Cross-Level Group Measurement Invariance When Groups Are at Different Levels of Multilevel Data

Educational and Psychological
Measurement

2018, Vol. 78(6) 973–997

© The Author(s) 2017

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164417739062

journals.sagepub.com/home/epm

Eun Sook Kim¹, Yan Wang¹, and Sarah M. Kiefer¹

Abstract

Studies comparing groups that are at different levels of multilevel data (namely, cross-level groups) using the same measure are not unusual such as student and teacher agreement in education or congruence between patient and physician perceptions in health research. Although establishing measurement invariance (MI) between these groups is important, testing MI is methodologically challenging because the groups compared for MI are at different levels with one group nested within the other group. We propose a multilevel confirmatory factor analysis (CFA) model that allows MI testing between cross-level groups at the between level and demonstrated testing MI between students and teachers using the promoting social interaction scale. Along with the demonstration, some methodological issues in implementing the proposed model (e.g., cluster invariance and reliability) and evaluating the model fit of multilevel CFA (e.g., Δ CFI and level-specific fit indices) and alternative approaches to the proposed model are discussed.

Keywords

measurement invariance, multilevel CFA, cross-level group measurement invariance, perception congruence

¹University of South Florida, Tampa, FL, USA

Corresponding Author:

Eun Sook Kim, Department of Educational and Psychological Studies, University of South Florida, 4202 East Fowler Avenue, EDU 105, Tampa, FL 33620, USA.

Email: ekim3@usf.edu

Multilevel data are prevalent in social sciences. Multilevel data consist of multiple units of analysis, for example, students and teachers as Level-1 and Level-2 units of analysis when students are nested within teachers. It is also not unusual that researchers compare these units *across* levels in multilevel data (Level-1 units vs. Level-2 units such as students vs. teachers; patients vs. caregivers, team members vs. leaders) where one group is at Level 1 and the other is at Level 2 (e.g., 200 students and their 30 teachers). For example, in educational research, congruence or disagreement between teachers' and students' perceptions has been studied for decades in many different aspects such as teaching effectiveness (e.g., Beaudrie, 2015), classroom learning environment (Chambers, 2015), and the use of power in the classroom (McCroskey & Richmond, 1982) because understanding the gap between teachers' and students' perceptions is important to improve educational and instructional practices and outcomes in the classrooms and schools (Beaudrie, 2015). Similar examples of congruence studies can be found across disciplines (e.g., Krupat et al., 2000; Lukowski, 2004; Watson, 2009).

In these studies that investigate agreement between two groups that are at different levels of multilevel data such as students and teachers with quantitative measures, the majority employed types of mean comparison (e.g., *t* tests), correlation, or interrater reliability to test the congruence between them using either items or scale composite scores (e.g., Beaudrie, 2015; Chambers, 2015). In the student teacher agreement studies, for example, student ratings or scores are averaged within classroom and compared to teacher scores. The mean differences between students and teachers are, then, tested for statistical significance to assess whether students and teachers agree on the tested measures. One fundamental assumption of these practices is measurement invariance between students and teachers, that is, students and teachers interpret and respond to the items of a measure in the same way. Even though it is widely recognized among applied researchers and practicing scientists that measurement invariance is a prerequisite to a meaningful comparison of scores between groups (Raykov, Marcoulides, & Li, 2012), measurement invariance between students and teachers, or between two groups that are at different levels of multilevel data, is often overlooked when agreement (i.e., a mean difference in this example) is assessed.

Before we move on to the issues of measurement invariance in agreement studies, it should be noted that the current study is based on research settings where agreement *between* groups is of focal interest. In the organizational climate research, there has been a long history of research on agreement *within* group (Schneider, Ehrhart, & Macey, 2011) which will be discussed later because agreement within group (e.g., agreement among students about their classroom learning environment) is essential to agreement between groups at different levels of multilevel data (e.g., agreement between students and teachers about the classroom learning environment). Another note is that agreement between groups can be measured in different ways. For demonstration purposes, we consider a mean difference between two groups as a measure of agreement.

When researchers are interested in testing measurement invariance (MI) between groups across levels of multilevel data, they may encounter several methodological challenges in implementing the MI test in this circumstance. Because one group (students) is nested within the other group (teachers), multilevel modeling that takes into account the dependency of students in the same classroom is essential. Previous simulation studies (French & Finch, 2010; Kim, Kwok, & Yoon, 2012) showed that failing to model data dependency in testing MI with multilevel data resulted in high Type I error (falsely rejecting invariance when a measure is invariant). In addition, the groups that are compared for MI are at different levels of multilevel data (students at Level 1 and teachers at Level 2). We term this type of group formation as cross-level groups hereafter. Because two groups in comparison are at different levels and the higher level group members are clusters to which the lower level group members belong, multiple group confirmatory factor analysis (CFA), which is a typical approach to MI testing for two groups (e.g., males and females), is not feasible.

The purpose of this article is twofold. First, we advocate testing measurement invariance in studies that investigate agreement between groups such as student teacher agreement studies. Second, we address the aforementioned methodological challenges in modeling measurement invariance when two groups are at different levels of multilevel data (i.e., cross-level groups), and introduce a multilevel CFA model with MI tested at Level 2. To this end, we demonstrate the proposed model by testing measurement invariance between student and teacher ratings on promoting social interaction in the classroom. Although this demonstration is based on educational data in classroom contexts, the proposed model may be applied to a wide variety of settings in different disciplines where units in different levels of multilevel data are compared for MI.

Multilevel Confirmatory Factor Analysis With Measurement Invariance Tested at Level 2

Measurement Invariance

Measurement invariance is defined as the equal probability of endorsing an item given latent ability regardless of group membership (Millsap, 2011). In other words, an item or a test is noninvariant or biased if individuals with the same latent ability have different expected item or test scores depending on their group membership. In CFA, measurement invariance is commonly tested with a multiple group analysis by comparing measurement models across groups. A measurement model is formulated for group g ($g = 1, 2, \dots, G$) as

$$Y_g = \tau_g + \Lambda_g \eta_g + \varepsilon_g \quad (1)$$

where Y , τ , Λ , η , and ε represent observed random variables, intercepts, factor loadings, factor scores, and residuals, respectively. Specifically, the measurement model shown in Equation (1) is constructed for G groups and compared for equality across

G groups. Depending on which element of the measurement model is compared, different levels of measurement invariance can be tested: (a) configural invariance if the factor structures are equal, (b) metric invariance if factor loadings are the same additionally, (c) scalar invariance if intercepts are equivalent additionally, and (d) strict invariance if residual variances are identical additionally across G groups. When metric invariance holds, the relations of factors are meaningfully comparable across groups. When scalar invariance holds, the means are additionally comparable across groups.

Building a Multilevel Model With Student Data

Multilevel Confirmatory Factor Analysis. When data are multilevel (i.e., two levels for exposition such as students nested within teachers), each score Y_{ij} for an individual i nested within cluster j consists of the cluster mean and its deviation from the cluster mean, which are represented by Y_{Bj} and Y_{Wij} , respectively (Zyphur, Kaplan, & Christian, 2008): $Y_{ij} = Y_{Bj} + Y_{Wij}$. Thus, the variance comes from two sources (within teacher and between teachers), and the total variance is decomposed into two components (Σ_T): pooled variance across students within teacher (i.e., within variance, Σ_W) and variance between teachers (i.e., between variance, Σ_B). Accordingly, a within measurement model can be constructed with the within variance and a between measurement model with the between variance. The within and between measurement models are written as

$$Y_{Wij} = \Lambda_W \eta_{Wij} + \varepsilon_{Wij}, \quad (2)$$

$$Y_{Bj} = \tau_B + \Lambda_B \eta_{Bj} + \varepsilon_{Bj}, \quad (3)$$

respectively, where W and B denote within (Level 1) and between (Level 2), respectively. Because the within data are basically deviation scores from the mean of each teacher, the intercepts are constrained at zero and not shown in Equation (2). Similar to the decomposition of observed scores, the latent factor scores are decomposed into within and between components (Muthén, 1994): $\eta_{ij} = \eta_{Wij} + \eta_{Bj}$. Given that the expectation of η_{Wij} is zero and the expectation of η_{Bj} is κ ,

$$E(Y_{Wij}) = 0; E(Y_{Bj}) = \tau_B + \Lambda_B \kappa.$$

Errors at the within and between levels are assumed to be normally distributed with mean zero and variance Θ_W and Θ_B , respectively: $\varepsilon_{Wij} \sim N(0, \Theta_W)$ and $\varepsilon_{Bj} \sim N(0, \Theta_B)$. It is also generally assumed that errors are not correlated with each other and with factors within the same level and also across levels.

Conceptualization of Latent Factors in Multilevel Confirmatory Factor Analysis. When researchers model latent factors at both within and between levels (Level 1 and Level 2, interchangeably), the meaning of those factors needs to be conceptualized at each level. There are different types of multilevel constructs depending on which

level construct is interpretable and of focal interest: for example, within-cluster construct, shared cluster construct, and configural cluster construct (Bliese, 2000; Kim, Dedrick, Cao, & Ferron, 2016; Stapleton, Yang, & Hancock, 2016). The multilevel constructs based on student ratings of classroom contexts (e.g., promoting social interaction) are considered as shared or reflective cluster constructs. For shared cluster constructs, the reference of the items is at Level 2, for example, classrooms or teachers, although students are the respondents (e.g., My teacher often allows us to discuss our work with classmates). That is, Level-2 factors are generally more meaningful as collective perceptions of students on teacher's practices than Level-1 factors that represent individual deviations from the common perception of teacher practices and classroom environment. For the same reason, the between measurement model of students is of interest and compared with the teachers' measurement model. For readers interested in different types of multilevel construct conceptualization and model specification, refer to Stapleton et al. (2016) and Bliese (2000). There is also a history of debate about the conceptualization of organizational climate as individual experiences and/or organizational attributes. An overview of the conceptualization of climate and its level of analysis theoretically and statistically is found in Schneider et al. (2011) as well as Yammarino and Dansereau (2011).

Agreement Within Groups. One underlying assumption of shared cluster constructs is agreement among individuals at Level 1 about the construct measured at Level 2. That is, it is assumed that individuals perceive the Level 2 construct (e.g., classroom environment) similarly. Although there are several measures that assess the degree of agreement within groups, we use intraclass correlation (ICC) in this article: ICC1 and ICC2. The first type of ICC is estimated using the formula, $ICC1 = \sigma_B^2 / (\sigma_B^2 + \sigma_W^2)$ where σ_B^2 and σ_W^2 denote between (Level-2) variance and within (Level-1) variance, respectively. The ICC1, that is, the proportion of between variance over the total variance, represents how much individuals within cluster correlate each other and is considered as an index of data dependency. The second type of ICC (ICC2) is often recommended for the shared cluster construct as a measure of agreement among individuals within cluster and estimated as $ICC2 = \sigma_B^2 / (\sigma_B^2 + (\sigma_W^2/n))$ where n represents the average cluster size (Bliese, 2000; Shrout & Fleiss, 1979; Stapleton et al., 2016). Of note is that the ICC2 formula is identical to that of reliability of cluster means (Raudenbush & Bryk, 2002). Unlike ICC1, ICC2 takes account of sampling error in estimating cluster means by dividing the within variance by cluster size. Holding within and between variances constant, increasing the average cluster size leads to higher ICC2, that is, more reliable cluster means. For reliable cluster means, ICC2 .70 or higher is suggested as good, between .50 and .70 as moderate, and below .50 as poor according to Klein et al. (2000).

Cluster Invariance. When a multilevel CFA model is constructed with student ratings, it is assumed that the student rating scale is invariant across classrooms (so called, cluster invariance) in order that one representative measurement model of students at

the classroom level can be compared to the teachers' measurement model. Cluster invariance is referred to as the invariance of a measure across Level-2 units of analysis in multilevel data (e.g., invariance across classrooms). Cluster invariance can be tested using multilevel CFA, which is formulated as Equations (2) and (3) (Jak, Oort, & Dolan, 2013). In cluster invariance testing, configural invariance means the equal factor structure across levels (i.e., factor structure of within measurement model = factor structure of between measurement model). Then, the equality of factor loadings across levels ($\Lambda_W = \Lambda_B$) indicates metric invariance across clusters. Finally, when the between-level residual variances equal zero ($\Theta_B = 0$) given cross-level factor loading invariance, it indicates scalar invariance across clusters. Cluster scalar invariance means that students interpret and respond to test items in the same way regardless of which classroom they belong to. Of note is that when factor loadings are unequal across levels but constrained equal for metric invariance testing, it increases the between-level residual variances. Because nonzero between-level residual variances could happen due to noninvariant factor loadings as well as noninvariant intercepts, metric and scalar invariance may not be always discernible in cluster invariance testing (Jak, Oort, & Dolan, 2014; Kim, Cao, Wang, & Nguyen, 2017). Thus, Kim and colleagues recommended to test scalar invariance directly from configural invariance. When cluster scalar invariance is violated, the sources of noninvariance can be further investigated by testing invariance of each parameter one by one (or examining modification indices).

Testing Measurement Invariance Between Students and Teachers

Multilevel Confirmatory Factor Analysis With Measurement Invariance Tested at Level 2. Once the multilevel CFA model based on student data is ready with cluster invariance, the teachers' measurement model is added to the Level-2 part of students' multilevel model. As illustrated in Figure 1, teachers' perceptions of their own practices in the classroom is another Level-2 factor on which the variables of teacher ratings load. As pointed out earlier, students and teachers cannot be considered as two independent groups because students are nested within teachers. Instead, they are represented by two separate but correlated factors at Level 2. Then, the between measurement model of students is compared with that of teachers for equivalence. The configural invariance refers to the identical structures of student and teacher factors at Level 2. If the loadings of the student and teacher factors are equal for all items ($\Lambda_{B_student} = \Lambda_{teacher}$), metric invariance holds. If the student and teacher intercepts are equal in addition to equal factor loadings, scalar invariance holds between students and teachers. This between-level model specification and MI testing procedure are similar to those of longitudinal MI testing (measurement invariance over time) in which variables at each time point measure a corresponding time factor (Time 1 factor, Time 2 factor, and so on) and factor loadings and intercepts of multiple time factors are compared for invariance (e.g., $\Lambda_{time1} = \Lambda_{time2} = \dots = \Lambda_{timeT}$ for metric invariance over T time points).

Figure 1. Model constructed for testing measurement invariance (MI) between students and teachers. FW = factor at Level 1 for student data; FB = factor at Level 2 for student data; FT = factor at Level 2 for teacher data. MI between students and teachers is tested at Level 2. Cluster invariance is imposed on the student measurement model.

In the model specification of student teacher measurement invariance testing, error covariance is worthy of note. For student and teacher measures, the same items are used with the only difference in the referent (“my teacher” for students; “I” for teachers). Also, a teacher rating and his or her students’ average rating are paired. In this case, it is common to allow the residual of a student item to covary with that of the corresponding teacher item. However, when cluster invariance holds across classrooms for the student measurement model, the residual variances at the between level are zero for student items, and thus, the covariance of errors between student and teacher items is not modeled (i.e., zero).

Alternative Models for Testing Measurement Invariance Between Students and Teachers. Although we promote multilevel CFA with MI tested at Level 2, researchers may consider single-level alternatives to multilevel CFA by either disaggregating or aggregating multilevel data. First, Level-2 data (e.g. teacher data) can be disaggregated to Level 1 (e.g., student level) and only the between-level part of the proposed model in Figure 1 is constructed as a student-level CFA model ignoring data dependency of students nested within teachers. The disadvantages of single-level approaches with multilevel data are widely discussed in the multilevel modeling

literature, and thus we discourage using this approach by all means. Second, the disaggregated single-level CFA model with adjusted standard errors that take account of the dependency of multilevel data can be considered alternatively. There are two major issues with this approach that applied researchers should be cognizant of. Creating a single-level model with multilevel data requires cross-level measurement equivalence (i.e., identical measurement models across levels) because a single set of factor loadings are estimated given the identical factor structure across levels (Zyphur et al., 2008). When the within and between factor structures are not identical, it is reported that this approach with adjusted standard errors (also called a design-based approach to multilevel data) yielded biased parameter estimates (Wu & Kwok, 2012) and inflated false detection of noninvariance in MI testing (Kim, Yoon, Wen, Luo, & Kwok, 2015). More problematically, when the construct tested for MI is a shared cluster construct with the focus on the between-level measurement model, constructing a student-level measurement model and testing measurement invariance between students and teachers at the student level is not conceptually appropriate even when this model is statistically feasible with cross-level equivalence.

Aggregating the student data to the teacher level could reduce model complexity if classroom means of student scores are reliable because the focal interest is at the between level in this case. Similar to the disaggregated single-level approach, the cross-level measurement equivalence is assumed in this approach. As pointed out earlier, observed cluster means (e.g., classroom average of student ratings) are not always reliable especially when cluster size is small, and previous studies evidenced that this approach yielded biased estimates of regression coefficients in regression analysis (Croon & van Veldhoven, 2007; Lüdtke et al., 2008). Croon and van Veldhoven suggested to use adjusted group means to correct bias. On the other hand, the multilevel CFA approach, which is demonstrated in this study using Mplus, is known to take into account the reliability of group means at the between level and produce unbiased estimates (Lüdtke et al., 2008).

Finally, a multilevel model with a saturated within model is another alternative to the proposed multilevel model. The within model is completely relaxed by allowing covariance among all within-level variables without imposing any specific model, which yields perfect model fit at the within level. This model is reasonable because the focal interest of the research is at the between level and MI is tested across groups at the between level. This saturated within model approach can be advantageous because the fit of the within-level measurement model is not of concern and all fit indices used for model evaluation and comparison are relevant to the level where MI is tested (i.e., between level). However, it should be kept in mind that the within level model can be theoretically important and informative although not directly used for MI testing between students and teachers. Thus, researchers “miss the opportunity to discover similarities and differences across levels of analysis in the functioning of their observed variables, differences that could have interesting theoretical importance” (Zyphur et al., 2008, p. 127).

In the following section, the proposed model was demonstrated with student and teacher ratings of promoting social interaction in the classroom. In this demonstration, we first tested measurement invariance of the promoting social interaction scale between students and teachers with the proposed multilevel CFA model and subsequently evaluated the agreement between students and teachers by comparing their factor means. For demonstration purposes, we also tested MI between students and teachers using alternative models and compared the results with those of the proposed model: single-level model ignoring data dependency, single-level approach with adjusted standard errors, aggregated data approach, and multilevel CFA with a saturated within model.

Method

Participants

Three hundred and thirty-six middle school students were nested within 31 teachers. The average cluster size, that is, average number of participating students per teacher is 10.84 ($SD = 3.72$) with the minimum 2 and maximum 19. The student sample had 52% males (48% females) and several ethnicities were represented (54% White, 21% Latino, 12% Other/Multiracial, 6% Asian, and 5% African American). The teacher sample had 23% males (77% females) and was predominantly White (74% White, 13% Latino, 7% African American, 3% Asian American, and 3% did not report their ethnicity in the survey).

Procedure

Students and teachers were recruited from social studies classrooms in three middle schools in the southeastern United States using a convenience sampling method. The research team requested one subject area to be recruited from; principals at participating schools selected social studies class due to practicality issues with state testing. Data used in this study were collected in fall 2009 during the first year of middle school (sixth grade) as part of a larger longitudinal study that investigated student motivation. The three middle schools recruited served a large, ethnically diverse, urban community. Active parental consent and participant assent were obtained prior to data collection; the average consent return rate for students was 57%. Participants were representative of demographics at each school and overall district demographics. Surveys were administered to participants during school hours. The research team returned one additional day to administer make-ups for participants who were absent. Of note is that measurement invariance among the three schools was not explicitly tested because of a small number of teachers per school (e.g., five teachers in one school) and nonconvergence, but indirectly evaluated in testing measurement invariance between students and teachers.¹

Measures

Variables in the current study included demographics and the Teacher Promotion of Social Interaction scale from the Classroom Social Environment developed by Ryan and Patrick (2001). The Promoting Social Interaction scale was self-report, used a 5-point Likert scale (1 = *not at all true*; 5 = *very true*), and was positively worded (i.e., higher scores indicated higher degrees of a given attribute).

Promoting Social Interaction Scale: Student Version. The student version of the Teacher Promotion of Social Interaction scale assessed the extent to which students perceived teachers as encouraging students to interact with one another during academic activities (Ryan & Patrick, 2001). This subscale comprised of four items, including “My teacher often allows students to discuss their work with classmates”, “My teacher encourages us to share ideas with one another in class”, “My teacher lets us ask other students when we need help with our work”, and “My teacher encourages us to get to know all the other students in class.” This scale has been previously administered to early adolescents and has been found to be valid and reliable (Patrick, Ryan, & Kaplan, 2007; Ryan & Patrick, 2001).

Promoting Social Interaction Scale: Teacher Version. The teacher version of the Teacher Promotion of Social Interaction scale assessed the extent to which teachers perceived themselves as encouraging students to interact with one another during academic activities (Ryan & Patrick, 2001). This measure was based on the student-version of the Teacher Promotion of Social Interaction subscale (Ryan & Patrick, 2001). The measure used similar items and Likert-type scale as the student-version, but was minimally reworded to reflect teachers’ own perceptions using the phrase (“I” instead of “My teacher”). Similar to the student-version, this measure contained four items, including “I often allow students to discuss their work with classmates”, “I encourage students to share ideas with one another in class”, “I let students ask other students when they need help with their work”, and “I encourage students to get to know all the other students in the class.” This is the first time this scale has been administered to teachers; psychometric information is not available from prior research studies.

Data Analytic Plan

First, descriptive statistics including mean, standard deviation, skewness, kurtosis, and correlation of items were examined for student and teacher data, separately, using SPSS version 22. For the student data, we also reported two types of intraclass correlation (ICC1 and ICC2). Regarding reliability, we estimated Cronbach’s alpha (α) and composite reliability omega (ω) using confirmatory factor analysis. Because Cronbach’s alpha is based on the stringent assumption of essentially tau-equivalent model and estimated by weighting all items equally (Raykov, 1997), we also examined omega that takes into account heterogeneous relations of items to the factor (i.e.,

unequal factor loadings) in its estimation. For the teacher measure, the Cronbach's alpha and composite reliability omega were estimated using single-level CFA. For the student data we constructed multilevel CFA. Because multilevel constructs (within and between) were modeled, we reported level-specific reliability estimates, that is, alpha and omega for within and between constructs separately from Level-1 and Level-2 measurement models, respectively (Geldhof, Preacher, & Zyphur, 2014).

For measurement invariance testing between students and teachers in terms of promoting social interaction, we first checked student and teacher measurement models separately including cluster invariance testing for the student model. Then, we applied the proposed model shown in Figure 1 to test configural invariance. We imposed cluster invariance for the student Level-1 and Level-2 measurement models with equal factor loadings across levels and zero Level-2 residual variances. For the identification of the covariance structure, the factor loading of one item (i.e., referent item) per factor was fixed at one. For the identification of the mean structure, the intercepts of the referent item were constrained equal between teachers and students, and the mean of the student Level-2 factor was fixed at zero (see the Mplus code in the appendix).

Metric invariance was tested by imposing the equality constraints on all factor loadings between students and teachers. After the establishment of metric invariance, the intercepts of all items were constrained equal between students and teachers for scalar invariance. Finally, factor means were compared between students and teachers to assess agreement between them. Because the mean of the student Level-2 factor was constrained at zero, the estimated mean of the teacher factor could be interpreted as the factor mean difference between students and teachers when scalar invariance was satisfied. We used Mplus version 7.1 (Muthén & Muthén, 2012) to conduct measurement invariance testing. The program default robust maximum likelihood (MLR) was adopted for model estimation which adjusts standard errors and chi-square statistics for nonnormal and nonindependent data.

In determining the level of invariance, we evaluated overall model fit and conducted likelihood ratio tests (LRTs) between two competing models (i.e., configural vs. metric invariance models; metric vs. scalar invariance models). Although Satorra-Bentler scaled likelihood ratio test (Satorra & Bentler, 2001) is considered optimal with the MLR estimation, we used the regular LRT because the Satorra-Bentler scaled LRT frequently produces negative values and the good performance of the regular LRT with MLR was also reported (Jak et al., 2014; Kim et al., 2017). The overall model fit was considered reasonable if chi-square $p \geq .05$, comparative fit index (CFI) $\geq .95$, and root mean square error of approximation (RMSEA) $\leq .06$ (Hu & Bentler, 1999). We also considered the changes in CFI and RMSEA. When $\Delta\text{CFI} < .01$ (Cheung & Rensvold, 2002) and $\Delta\text{RMSEA} < .015$ (Chen, 2007), the model with additional constraints (invariance model) was supported. These model fit criteria were suggested in the literature based on single-level CFA models, but for the demonstration purpose, we used these long-established conventional criteria for multilevel CFA. Considering that some scholars (e.g., Hsu, Kwok, Lin, & Acosta,

2015; Ryu & West, 2009) criticized the overall fit evaluation of multilevel models, we also assessed level-specific fit indices and discussed their usability for the proposed model evaluation.

The four alternative models were constructed with two correlated factors of students and teachers.² The first two alternative models were disaggregated single-level models without and with standard error adjustment, respectively ($n = 336$). We used TYPE = COMPLEX for adjusted standard errors in Mplus. For the aggregated approach, student scores of each item were averaged by teacher and combined with the teacher data ($n = 31$). Finally, for the saturated within model, the between-level model was identical to that of the proposed model, but at the within level student perception of teacher promoting social interaction was not modeled and all items were simply allowed to be correlated. For all four models, MI was tested with the same aforementioned procedure with the same model evaluation and comparison methods. For the single-level approach ignoring data dependency and the aggregated approach, maximum likelihood was used for model estimation. The MLR was used for the other two models.

Results

Descriptive Statistics

Table 1 presents descriptive statistics and correlations between items by students and teachers data separately. Overall, student ratings were on average lower than teacher ratings across all items. Items 2 ("Share ideas") and 4 ("Get to know others") had higher responses than other items across students and teachers, suggesting that both students and teachers perceived these two aspects of social interaction were promoted more than the others. Greater variability in student responses was observed across all items. Items were approximately normally distributed except for the slightly negatively skewed distribution of teacher responses for Items 2 and 4. Item responses were more highly correlated for students than teachers. For the student data, the item ICC1s ranged between .05 and .10 indicating small variability across classrooms relative to within-classroom variability. The estimated ICC2s were higher with the values from .36 to .55, but lower except one item than the minimum cutoff .50 for a moderately reliable shared cluster construct showing considerable variability among students about their teacher's promoting social interaction.

For the teacher responses, there was no missing case ($n = 31$). For the student responses, missing rates were very small across items. The missing cases were treated using full information maximum likelihood estimation (MLR) that included all 336 students in data analyses.

Reliability

Based on student data, Cronbach's alphas for the scale were .72 and .94, at the within and between levels, respectively, and composite reliability omegas .73 and .97,

Table 1. Descriptive Statistics and Correlations Between Items by Students and Teachers.

Item Descriptions ^a	N	M	SD	Skewness	Kurtosis	1	2	3	ICC1	ICC2
Students										
1. Discuss with classmates	328	2.99	1.16	.04	-.42	—	—	—	.05	.36
2. Share ideas	321	3.37	1.23	-.33	-.76	.45	—	—	.08	.47
3. Ask others for help	321	2.93	1.25	.11	-.84	.47	.55	—	.05	.35
4. Get to know others	318	3.70	1.31	-.61	-.72	.29	.45	.34	.10	.55
Teachers										
1. Discuss with classmates	31	4.00	.86	-.34	-.75	—	—	—	—	—
2. Share ideas	31	4.71	.53	-1.67	2.14	.15	—	—	—	—
3. Ask others for help	31	3.94	.85	-.22	-.84	.23	.11	—	—	—
4. Get to know others	31	4.58	.62	-1.22	.56	.31	.13	.33	—	—

^aItem descriptions are paraphrased. Items were measured on a 5-point Likert-type scale (1 = not at all true, 5 = very true). ICC1 = $\sigma_B^2 / (\sigma_B^2 + \sigma_W^2)$ and ICC2 = $\sigma_B^2 / (\sigma_B^2 + (\sigma_W^2/n))$ where σ_B^2 and σ_W^2 denote between and within variances, respectively and n is the average cluster size.

respectively. Cronbach's alpha for the scale using teacher data was .51 and composite reliability omega .53.

Measurement Invariance Testing

Student and teacher measurement models were constructed separately and the model fit was checked prior to measurement invariance testing. The multilevel CFA model with cluster invariance fitted student data well, $\chi^2(11) = 25.40$, $p = .008$, CFI = .967, RMSEA = .063, standardized root mean square residual-within (SRMR-W) = .030, and standardized root mean square residual-between (SRMR-B) = .167. The single-level measurement model for teachers had an excellent model fit, $\chi^2(2) = 1.06$, $p = .59$, CFI = 1.000, RMSEA = .000, SRMR = .011.

To test measurement invariance between students and teachers, the multilevel CFA model shown in Figure 1 was constructed with factor loadings and intercepts freely estimated for both student and teacher factors at the between level except the first item for model identification. This configural invariance model showed good fit, $\chi^2(28) = 33.24$, $p = .23$, CFI = .985, RMSEA = .024, SRMR-W = .030, SRMR-B = .132. Next, the metric invariance model showed good fit indicated by the nonsignificant p -value of χ^2 and model fit indices (see the top panel of Table 2). Additionally, metric invariance held over the configural invariance model, based on the likelihood ratio test, $\Delta\chi^2(5) = 3.45$, $p > .05$, and $\Delta\text{CFI} = .001 < .01$, $\Delta\text{RMSEA} = -.001 < .015$. When the metric and scalar invariance models were compared, results supported the scalar invariance model, based on the nonsignificant p -value of χ^2 , model fit indices, $\Delta\chi^2(3) = 1.97$, $p > .05$, and $\Delta\text{CFI} = -.003 < .01$, $\Delta\text{RMSEA} = -.003 < .015$. Therefore, we concluded that the scalar invariance was established for the promoting social interaction scale across students and teachers.

With the establishment of scalar invariance, factor mean comparisons were conducted. Results showed that the mean of the promoting social interaction factor for teachers (factor variance = 0.060) was 0.907 higher than that for students (factor variance = 0.046). Converting factor scores from the raw scale to the standardized scale (factor variances equaled 1.00), the standardized factor mean difference between teachers and students was 4.23. That is, the factor mean for teachers was 4.23 *SD* higher than that for students. In other words, teachers perceived themselves as actively promoting social interaction in class, while from students' perspective, social interaction was promoted to a lesser degree. The correlation between student and teacher factors was .43, but this was not statistically significant ($p = .15$).

The bottom panels of Table 2 showed the results of MI testing with four alternative models. When the single-level model was used to test MI ignoring data dependency, only configural invariance was satisfied with excellent fit. The model fit from configural to metric and from metric to scalar invariance deteriorated notably. The single-level model with adjusted standard errors showed mixed results. The likelihood ratio test did not support metric invariance. However, both ΔCFI and ΔRMSEA supported scalar invariance, and the fit of the scalar invariance model was excellent.

Table 2. Measurement Invariance Testing of the Promoting Social Interaction Scale across Students and Teachers.

Model	χ^2	df	p	CFI	RMSEA	SRMR-W	SRMR-B	$\Delta\chi^2$	Δdf	p ($\Delta\chi^2$)	ΔCFI	$\Delta RMSEA$	Mean ^a (SE)
Multilevel confirmatory factor analysis													
Configural	33.24	28	.23	.985	.024	.030	.132	—	—	—	—	—	—
Metric	36.69	31	.22	.984	.023	.030	.162	3.45	3	.631	.001	-.001	—
Scalar	38.66	34	.27	.987	.020	.031	.164	1.97	3	.579	-.003	-.003	.907 (.103)
Single-level analysis without standard error adjustment ^b													
Configural	18.20	19	.51	1.000	.000	.027	—	—	—	—	—	—	—
Metric	57.11	22	.00	.909	.069	.099	.38.91	38.91	3	.000	.091	.069	—
Scalar	70.43	25	.00	.883	.074	.099	13.32	13.32	3	.004	.026	.005	—
Single-level analysis with standard error adjustment ^b													
Configural	7.59	19	.99	1.000	.000	.027	—	—	—	—	—	—	—
Metric	17.68	22	.73	1.000	.000	.099	10.09	10.09	3	.018	.000	.000	—
Scalar	22.08	25	.63	1.000	.000	.099	4.40	4.40	3	.221	.000	.000	.997 (.137)
Aggregated analysis ^b													
Configural	15.60	19	.68	1.000	.000	.072	—	—	—	—	—	—	—
Metric	20.50	22	.55	1.000	.000	.148	4.90	4.90	3	.179	.000	.000	—
Scalar	22.47	25	.61	1.000	.000	.163	1.97	1.97	3	.579	.000	.000	.881 (.126)
Multilevel confirmatory factor analysis with saturated within-level model													
Configural	20.40	23	.62	1.000	.000	.005	.137	—	—	—	—	—	—
Metric	23.95	26	.58	1.000	.000	.005	.162	3.55	3	.314	.000	.000	—
Scalar	30.36	29	.40	.996	.012	.007	.165	6.41	3	.093	.004	.012	.969 (.169)

Note. CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR-W and SRMR-B = standardized root mean square residual at the within and between levels, respectively; $\Delta\chi^2$, Δdf , ΔCFI and $\Delta RMSEA$ refer to the changes in χ^2 , df, CFI, and RMSEA, respectively.

^aThe mean of the teacher factor, which indicates the factor mean difference between students and teachers. ^bSRMR was reported because model only involved one level of analysis.

The aggregated data analysis and the ML CFA with saturated within model supported scalar invariance between students and teachers, which is consistent to the conclusion of the proposed model. In all three approaches in which scalar invariance was fully or partly supported, the teacher factor mean was statistically significantly higher than the student factor mean indicating there was a considerable difference between students and teachers in their perception of promoting social interaction in the classroom. The smallest factor mean difference (0.881) was observed with the aggregated approach; the largest (0.997) with the single-level analysis with adjusted standard errors. The proposed ML CFA yielded the smallest standard error.

Discussion

This study demonstrated model specification and procedures of measurement invariance testing when the comparing groups are at different levels of multilevel data, in other words, one group is nested within the other group. In the following sections, we briefly discuss congruence or disagreement of student teacher perceptions about promoting social interaction given measurement invariance of the scale between students and teachers. Finally, remaining methodological issues in implementing the proposed model and future research directions are discussed.

Overall, the findings indicated that students and teachers perceived the promotion of social interaction in the classroom differently. This adds to the literature on congruence or disagreement between student and teacher perceptions of the classroom (Beaudrie, 2015, Chambers, 2015; McCroskey & Richmond, 1982) and may have implications for improving educational outcomes and instructional practices (Beaudrie, 2015). The findings extend prior studies using multilevel models by proposing a multilevel CFA model that allows MI testing between cross-level groups and demonstrating testing MI between students and teachers. One possible explanation for differing student and teacher perceptions is that reports of promoting social interaction may be more subjective than other aspects of the classroom such as instruction (Desimone, Smith, & Frisvold, 2010; Mitchell, Bradshaw, & Leaf, 2010). Given that data for the current study were collected during the first year of middle school, students and teachers may differ in their perceptions of the classroom environment due to systematic changes in classroom and school environments associated with the transition into middle school (Eccles et al., 1993). Students' perceptions of the classroom environment may be in flux following the transition into middle school as a result of experiencing changes in school structure, motivational and instructional techniques, grouping practices, and quality of teacher-student and student-student relationships (Eccles et al., 1993; Eccles & Roeser, 2011). Additionally, prior research suggests that student perceptions cannot be reliably aggregated at the classroom level (Lam, Ruzek, Schenke, Conley, & Karabenick, 2015; Miller & Murdock, 2007; Schenke, Ruzek, Lam, Karabenick, & Eccles, 2017; Schweig, 2014). This aligns with theory that student perceptions reflect individual interpretations of their experiences in the same environment (Ames, 1992; Maehr & Midgley, 1991) and

research indicating differential treatment of teachers toward students within the same classroom (Brattesani, Weinstein, & Marshall, 1984; Eccles & Blumenfeld, 1985; Kuklinski & Weinstein, 2000). Future research is needed to further examine the heterogeneity of student and teacher perceptions of the classroom and individual differences in the influence classroom contextual factors (Schenke et al., 2017). This may inform future educational interventions and align with recommendations for researchers to examine individual- and contextual-level moderating variables that may limit intervention results (Rosenzweig & Wigfield, 2016) and to consider individual's beliefs and understandings of themselves and their environments (Wilson & Buttrick, 2016).

Although this study showed scalar invariance of the promoting social interaction measure between students and teachers, several remaining issues are worthy of note. We observed low reliability estimates of teacher scores which raised a concern about the use of observed teacher scores in examining teacher student congruence. In this case, the advantage of multilevel CFA is more prominent over conventional approaches (e.g., an observed mean comparison) because CFA allows researchers to compare factor means after taking measurement error into account.

The item ICC2s of the student data were low with the values between .36 and .55. The shared cluster construct requires ICC2 greater than .50 for the meaningful interpretation of its between-level construct (Klein et al., 2000). The low ICC2 values of student variables indicate that students within classroom rated their teacher widely differently. Thus, although scalar invariance holds between students and teachers on average, it should be noted that students also disagree with each other about their teachers to some degree. The low ICC2s might not be surprising given small cluster size (10 students on average). That is, in the extreme case, a teacher was rated by two students, which obviously raises a concern of the reliability of the average student rating. Because the reliability of cluster means partly depends on the cluster size, when researchers conduct studies with the shared cluster construct, cluster size should be sufficiently large to have reliable cluster means for the between-level construct.

To establish measurement invariance between students and teachers using multilevel CFA, we mainly depended on the fit criteria developed for single-level CFA. Recently, Kim et al. (2017) investigated the conventional model comparison criteria, $\Delta CFI < .01$ and $\Delta RMSEA < .015$ in measurement invariance testing with multilevel CFA and showed reasonable performance of $\Delta CFI < .01$. Although they examined these cutoffs for specifically many group comparisons, it appears that the conventional cutoff is generally robust for multilevel CFA. However, researchers should be mindful of some criticism on the general use of CFI in measurement invariance testing. Lai and Yoon (2015) noted that CFI values depend on how a baseline model is specified and the conventional baseline model is not appropriate to MI testing. The conventional baseline model freely estimates the means and variances of observed variables between groups and, thus, the scalar invariance model that constrains the intercepts of observed variables is not nested within the baseline model. They proposed a modified baseline model and showed that CFI with the modified baseline

model was more sensitive to the violation of scalar invariance than the conventional CFI. Although ΔCFI performed reasonably with large noninvariance, its performance was not as good as the proposed CFI. Note that their investigation was conducted with single-level multiple group CFA. Given the popularity of ΔCFI in MI testing, further investigation is called for about the adequacy of ΔCFI in a variety of MI research settings.

Hsu et al. (2015) conducted a simulation study to assess the sensitivity of RMSEA, CFI, Tucker–Lewis index (TLI), SRMR-W, and SRMR-B to model misspecification in multilevel CFA and reported that SRMR-B was the only fit index sensitive to the model misspecification at the between level. Given their findings, evaluating between level-specific fit indices including SRMR-B is important for multilevel CFA with focal interest in the between-level model as demonstrated in this study. However, Hsu, Lin, Kwok, Acosta, and Willson (2016) showed that the performance of between level-specific indices depended on ICC. When ICC was low (about .09), they did not recommend SRMR-B and RMSEA-B for model evaluation. Instead, because Hsu et al.'s (2016) study evidenced reasonable performance of CFI-B and TLI-B ($\geq .95$), we evaluated these fit indices for the between-level models and observed that both between level-specific fit indices deteriorated notably when scalar invariance was imposed (CFI-B .928 from 1.00; TLI-B .930 from 1.117), which raises a concern of scalar noninvariance between students and teachers. To date, the criteria to evaluate and compare measurement invariance models in multilevel CFA have not been well established. Methodological efforts to develop reliable criteria based on between level-specific fit indices are needed. Especially, these between-level fit criteria are essential to evaluate multilevel CFA for cluster constructs such as shared cluster constructs (e.g., leadership, classroom climate).

In this study we demonstrated measurement invariance testing between students and teachers given classroom invariance of the student measure. However, the systematic review of 72 multilevel CFA studies (Kim et al., 2016) revealed that only a small number of studies tested cross-level invariance which is required for metric invariance across clusters and only two of those studies confirmed cross-level invariance. Given their report, cluster noninvariance or cluster bias (difference in student response patterns across classrooms) may not be unusual, which questions the feasibility of testing student teacher invariance. As suggested in general MI testing, researchers may establish partial cluster invariance and proceed for testing measurement invariance between students and teachers. However, the meaning of Level-2 constructs under cluster noninvariance (including different factor structures across levels such as one factor at Level 1 and two factors at Level 2) and the treatment of such constructs in subsequent analyses need further discussions and investigation from both substantive and methodological perspectives.

Measurement invariance between students and teachers was also investigated with four alternative models in this study. The single-level approach ignoring the hierarchical data structure reached a different level of MI (configural invariance) from what the proposed model yielded (scalar invariance). Previous studies (e.g., French

& Finch, 2010; Kim et al., 2012) demonstrated that single-level MI tests with multi-level data without taking into account data dependency resulted in high Type I error (i.e., high false detection of noninvariance). For both conceptual and statistical reasons, this approach is not recommended. The other three approaches reached the same conclusion of scalar invariance consistent to the conclusion of the proposed model, which is not surprising given cross-level measurement equivalence of promoting social interaction. The cross-level equivalence of measurement models is a fundamental underlying assumption of a single-level approach, either disaggregating to Level 1 or aggregating to Level 2 (Zyphur et al., 2008). When this assumption is satisfied, the MI tests at Level 1 with disaggregated data may perform reasonably in the detection of noninvariance if standard errors are properly adjusted as demonstrated in Kim et al. (2012) with within-level groups (e.g., gender). However, in this study the comparing groups are at different levels and the teacher construct of promoting social interaction is uninterpretable at the student level. Thus, even though this disaggregating approach may yield statistically reasonable solutions of MI under certain circumstances, this approach is not conceptually appropriate for the construct conceptualized at Level 2.

Because the construct of interest is at the between level, aggregating student scores and testing MI at the teacher level as a single level model or testing MI at the between level with a saturated within model seem appealing. The aggregated single-level approach has been adopted in the organizational climate and contextual studies (Hofmann, 2002; Lüdtke et al., 2008), but its disadvantages are also well documented in the literature. Croon and van Veldhoven (2007) pointed out that the aggregating single-level approach would yield unbiased parameter estimates if there is no within-group variability (or asymptotically unbiased as group size increases). Lüdtke et al. (2008) compared the aggregated approach with the multilevel approach (called multilevel latent covariate model) in estimating contextual effects, and showed the outperformance of the multilevel approach. In the current study, we observed the within-group agreement was not high with only one item meeting the minimal cutoff of ICC2 for reliable classroom means. In this situation, the proposed multilevel CFA approach is more compelling because this approach takes into account the unreliability of group means at the between level (Lüdtke et al., 2008) although we did not observe any difference between two approaches in MI testing results with both endorsing scalar invariance.

The multilevel CFA with a saturated within model also yielded identical results to those of the proposed model in MI testing between students and teachers. This approach could be very attractive to applied researchers because the specification of the within model is not needed and the model fit is evaluated only at the between level. However, the factor structure of student perception of promoting social interaction is not investigated at the within level, the tenability of the within model is not tested, and subsequently cluster invariance including cross-level measurement equivalence cannot be assessed and is not explicit as in the proposed model. Solely

for the between-level MI testing purposes, this approach is a potential alternative to the proposed model with advantages.

For the factor mean comparison between students and teachers, all models that endorsed scalar invariance showed a statistically significant mean difference. However, the estimated parameters and the associated standard errors vary to some degree across models as shown in Table 2. Because the estimated effect size was very large, some discrepancies in the estimated factor mean difference and its standard error across different approaches did not lead to different conclusions about student teacher agreement. However, the aggregated approach yielded the smallest factor mean difference possibly because of the unreliability of group means when student ratings were averaged as Lüdtke et al. (2008) consistently observed the underestimation of contextual effects with the aggregated approach in their simulation study. We also observed apparently large heterogeneity of factor variance between students and teachers for the disaggregated single-level approaches with notably small teacher variability, which again raises a concern of the legitimacy of the disaggregated analysis with teacher data. Finally, a future Monte Carlo simulation study is called for to compare the proposed method with potential alternatives under various research conditions in terms of MI testing and the accuracy and efficiency of the parameter estimates to fully discover the benefits of one model over another.

Conclusion

In studies investigating congruence between two groups, we advocate MI testing to ensure that the two groups interpret and respond to test items in the same way. Because the groups compared for MI are often at different levels of multilevel data with one group nested within the other group, we proposed and demonstrated a multilevel CFA model that allows MI testing between cross-level groups in multilevel data. This study is expected to promote MI testing between groups in congruence studies in different fields (e.g., agreement of student teacher perceptions in education). The proposed model can also be applied to any measurement invariance testing research with cross-level groups that are at different levels of multilevel data. Along with the demonstration, we discussed some methodological issues in implementing the proposed model and evaluating the model fit of multilevel CFA, which call for future research.

Appendix

Data: file is data_sturev.txt;

Variable: names are TeacherID StudentID Y1-Y4 T1-T4;

missing are all(999);

usevariables are Y1-Y4 T1-T4;

between = T1-T4;

cluster is TeacherID;

Analysis: type = twolevel;

Model:

%within%

FW by Y1

Y2(1)

Y3(2)

Y4(3);

%between%

FB by Y1

Y2(1)

Y3(2)

Y4(3); ! cross-level factor loading invariance for metric invariance of student data;

Y1-Y4@0; ! residual variances constrained to be zero for scalar invariance of student data;

FT by T1-T4; ! factor loading was fixed at one for the referent item to identify the covariance structure;

FT with FB; ! the correlation between teachers' and students' perceptions of promoting social interaction was estimated;

!Y1-Y4 pwth T1-T4; ! residual covariances were not estimated, because residual variances of student items were constrained to be zero;

[Y1](i1);

[Y2-Y4];

[T1](i1); ! intercepts of the referent item were constrained equal between teachers and students;

[T2-T4];

[FB@0 FT]; ! the mean of the student level-2 factor was fixed at zero;

Output: sampstat stdyx residual modindices(all) cinterval;

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. For the student data, if cluster invariance holds, that is, measurement invariance holds across classrooms regardless of students' school membership, it could imply measurement invariance between schools. For the teacher measure, if the measurement model fits well the data with all three schools combined, we consider this as no evidence of noninvariance between schools.

2. Error correlations between student and teacher items with the same wording were not specified in any of these models because the estimated error correlations were trivial and not statistically significant.
3. Because cluster invariance testing with the student data indicated partial cluster invariance, we also conducted the subsequent measurement invariance testing with partial cluster invariance of the student measure (one item factor loading relaxed from cross-level invariance). However, because the final model with full cluster invariance fit as well as that with partial cluster invariance and we reached the same conclusion either with full or partial cluster invariance, we reported the measurement invariance testing between students and teachers with full cluster invariance of the student measure for demonstration purposes. Because cluster invariance testing is not a major focus of this demonstration, we do not include all details, but the results are available on request.

References

- Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*, 261-271. doi:10.1037/0022-0663.84.3.261
- Beaudrie, S. M. (2015). Instructional effectiveness in the SHL classroom: Comparing teacher and student perceptions. *Journal of Hispanic Higher Education, 14*, 274-297. doi:10.1177/1538192715575372
- Brattesani, K. A., Weinstein, R. S., & Marshall, H. H. (1984). Student perceptions of differential teacher treatment as moderators of teacher expectation effects. *Journal of Educational Psychology, 76*, 236-247. doi:10.1037/0022-0663.76.2.236
- Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. 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, CA: Jossey-Bass.
- Chambers, C. B. (2015). *Examining and analyzing international high schools' teacher and student perceptions of the classroom learning environments*. (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses A&I. (3739505)
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*, 464-504. doi:10.1080/10705510701301834
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*, 233-255. doi:10.1207/S15328007SEM0902_5
- Croon, M. A., & van Veldhoven, M. J. P.M. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods, 12*, 45-57. doi:10.1037/1082-989X.12.1.45
- Desimone, L. M., Smith, T. M., & Frisvold, D. E. (2010). Survey measures of classroom instruction: Comparing student and teacher reports. *Educational Policy, 24*, 267-329. doi:10.1177/0895904808330173
- Eccles, J. S., & Blumenfeld, P. (1985). Classroom experiences and student gender: Are there differences and do they matter? In L. Wilinon & C. Marrett (Eds.), *Gender influences in classroom interaction* (pp. 79-114). Hillsdale, NJ: Lawrence Erlbaum.
- Eccles, J. S., Midgley, C., Wigfield, A., Buchanan, C. M., Reuman, D., Flanagan, C., & Mac Iver, D. (1993). Development during adolescence: The impact of stage-environment fit on

- adolescents' experiences in schools and families. *American Psychologist*, 48, 90-101. doi:10.1037/0003-066X.48.2.90
- Eccles, J. S., & Roeser, R. W. (2011). Schools as developmental contexts during adolescence. *Journal of Research on Adolescence*, 21, 225-241. doi:10.1111/j.1532-7795.2010.00725.x
- French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement*, 47, 299-317. doi:10.1111/j.1745-3984.2010.00115.x
- Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, 19, 79-91. doi:10.1037/a0032138
- Hofmann, D. A. (2002). Issues in multilevel research: Theory development, measurement and analysis. In Rogelberg, S. (Ed.), *Handbook of research methods in industrial and organizational psychology*. Oxford, England: Blackwell.
- Hsu, H.-Y., Kwok, O., Lin, J.-H., & Acosta, S. (2015). Detecting misspecified multilevel structural equation models with common fit indices: A Monte Carlo study. *Multivariate Behavioral Research*, 50, 197-215. doi:10.1080/00273171.2014.977429
- Hsu, H.-Y., Lin, J.-H., Kwok, O., Acosta, S., & Willson, V. (2016). The impact of intraclass correlation on the efficacy of level-specific fit indices in multilevel structural equation modeling: A Monte Carlo study. *Educational and Psychological Measurement*, 77, 5-31. doi:10.1177/0013164416642823
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 1-55. doi:10.1080/10705519909540118
- Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, 20, 265-282. doi:10.1080/10705511.2013.769392
- Jak, S., Oort, F. J., & Dolan, C. V. (2014). Measurement bias in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, 21, 31-39. doi:10.1080/10705511.2014.856694
- Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, 24, 524-544. doi:10.1080/10705511.2017.1304822
- Kim, E. S., Dedrick, R. F., Cao, C., & Ferron, J. M. (2016). Multilevel factor analysis: Reporting guidelines and a review of reporting practices. *Multivariate Behavioral Research*, 51, 881-898. doi:10.1080/00273171.2016.1228042
- Kim, E. S., Kwok, O., & Yoon, M. (2012). Testing factorial invariance in multilevel data: A Monte Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal*, 19, 250-267. doi:10.1080/10705511.2012.659623
- Kim, E. S., Yoon, M., Wen, Y., Luo, W., & Kwok, O. (2015). Within-level group factorial invariance with multilevel data: Multilevel factor mixture and multilevel MIMIC models. *Structural Equation Modeling: A Multidisciplinary Journal*, 22, 603-616. doi:10.1080/10705511.2014.938217
- Klein, K. J., Bliese, P. D., Kozlowski, S. W. J., Dansereau, F., Gavin, M. B., Griffin, M. A., & . . . Bligh, M. C. (2000). Multilevel analytical techniques: Commonalities, differences, and continuing questions. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 512-553). San Francisco, CA: Jossey-Bass.

- Krupat, E., Rosenkranz, S. L., Yeager, C. M., Barnard, K., Putnam, S. M., & Inui, T. S. (2000). The practice orientations of physicians and patients: The effect of doctor-patient congruence on satisfaction. *Patient Education and Counseling*, 39, 49-59. doi: 10.1016/S0738-3991(99)00090-7
- Kuklinski, M., & Weinstein, R. (2000). Classroom and grade level differences in the stability of teacher expectations and perceived differential teacher treatment. *Learning Environments Research*, 3, 1-34. doi:10.1023/A:1009904718353
- Lai, M. H. C., & Yoon, M. (2015). A modified comparative fit index for factorial invariance studies. *Structural Equation Modeling: A Multidisciplinary Journal*, 22, 236-248. doi:10.1080/10705511.2014.935928
- Lam, A. C., Ruzek, E. A., Schenke, K., Conley, A. M., & Karabenick, S. A. (2015). Student perceptions of classroom achievement goal structure: Is it appropriate to aggregate? *Journal of Educational Psychology*, 107, 1102-1115. doi:10.1037/edu0000028
- Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, 13, 203-229. doi:10.1037/a0012869
- Lukowski, A. V. (2004). *The relationship between employer-employee congruence of values, job satisfaction, and life satisfaction* (Doctoral dissertation). Retrieved from ProQuest Dissertation & Theses: A&I. (3139614)
- Maehr, M. L., & Midgley, C. (1991). Enhancing student motivation: A school-wide approach. *Educational Psychologist*, 26, 399-427. doi:10.1080/00461520.1991.9653140
- McCroskey, J. C., & Richmond, V. P. (1982, May). *Power in the classroom I: Teacher and student perceptions*. Paper presented at the Annual Meeting of the Eastern Communication Association, Hartford, CT.
- Miller, A. D., & Murdock, T. B. (2007). Modeling latent true scores to determine the utility of aggregate student perceptions as classroom indicators in HLM: The case of classroom goal structures. *Contemporary Educational Psychology*, 32, 83-104. doi:10.1016/j.cedpsych.2006.10.006
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.
- Mitchell, M. M., Bradshaw, C. P., & Leaf, P. L. (2010). Student and teacher perceptions of school climate: A multilevel exploration of patterns of discrepancy. *Journal of School Health*, 80, 271-279. doi:10.1111/j.1746-1561.2010.00501.x
- Muthén, B. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, 22, 376-398. doi:10.1177/0049124194022003006
- Muthén, B. O., & Muthén, L. K. (2012). *Mplus 7.1* [Computer software]. Los Angeles, CA: Muthén & Muthén.
- Patrick, H., Ryan, A. M., & Kaplan, A. (2007). Early adolescents' perceptions of the classroom social environment, motivational beliefs, and engagement. *Journal of Educational Psychology*, 99, 83-98. doi:10.1037/0022-0663.99.1.83
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau equivalence with fixed congeneric components. *Multivariate Behavioral Research*, 32, 329-353. doi:10.1207/s15327906mbr3204_2

- Raykov, T., Marcoulides, G. A., & Li, C. (2012). Measurement invariance for latent constructs in multiple populations: A critical view and refocus. *Educational and Psychological Measurement*, 72, 954-974. doi:10.1177/0013164412441607
- Rosenzweig, E., & Wigfield, A. (2016). STEM motivation interventions for adolescents: A promising start, but further to go. *Educational Psychologist*, 51, 146-163. doi:10.1080/00461520.2016.1154792
- Ryan, A. M., & Patrick, H. (2001). The classroom social environment and changes in adolescents' motivation and engagement during middle school. *American Educational Research Journal*, 38, 437-460. doi:10.3102/00028312038002437
- Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 16, 583-601. doi:10.1080/10705510903203466
- Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66, 507-514. doi:10.1007/BF02296192
- Schenke, K., Ruzek, E., Lam, A. C., Karabenick, S. A., & Eccles, J. S. (2017). Heterogeneity of student perceptions of the classroom climate: A latent profile approach. *Learning Environments Research*, 20, 289-306. doi:10.1007/s10984-017-9235-z
- Schneider, B., Ehrhart, M. G., & Macey, W. A. (2011). Organizational climate research: Achievements and the road ahead. In N. M. Ashkanasy, C. P. M. Wilderom, & M. E. Peterson (Eds.), *Handbook of organizational culture and climate* (2nd ed., pp. 29-49). Thousand Oaks, CA: Sage.
- Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. *Educational Evaluation and Policy Analysis*, 36, 259-280. doi:10.3102/0162373713509880
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420-428. doi:10.1037/0033-2909.86.2.420
- Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, 41, 481-520. doi:10.3102/1076998616646200
- Watson, E. (2009). *Inclusion: An examination of attitude congruency within elementary schools* (Doctoral dissertation). Retrieved from ProQuest Dissertation & Theses: A&I. (3366834)
- Wilson, T. D., & Buttrick, N. R. (2016). New directions in social psychological interventions to improve academic achievement. *Journal of Educational Psychology*, 108, 392-396. doi:10.1037/edu0000111
- Wu, J., & Kwok, O. (2012). Using SEM to analyze complex survey data: A comparison between design-based single-level and model-based multi-level approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, 19, 16-35. doi:10.1080/10705511.2012.634703
- Yammarino, E. J., & Dansereau, F. (2011). Multilevel issues in organizational culture and climate research. In N. M. Ashkanasy, C. P. M. Wilderom, & M. E. Peterson (Eds.), *Handbook of organizational culture and climate* (2nd ed., pp. 50-75). Thousand Oaks, CA: Sage.
- Zyphur, M. J., Kaplan, S. A., and 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, 127-140. doi:10.1037/1089-2699.12.2.127