



The Impact of Partial Measurement Invariance on Testing Moderation for Single and Multi-Level Data

Yu-Yu Hsiao^{1*} and Mark H. C. Lai²

¹ Center on Alcoholism, Substance Abuse, and Addictions, University of New Mexico, Albuquerque, NM, United States,

² School of Education, University of Cincinnati, Cincinnati, OH, United States

Moderation effect is a commonly used concept in the field of social and behavioral science. Several studies regarding the implication of moderation effects have been done; however, little is known about how partial measurement invariance influences the properties of tests for moderation effects when categorical moderators were used. Additionally, whether the impact is the same across single and multilevel data is still unknown. Hence, the purpose of the present study is twofold: (a) To investigate the performance of the moderation test in single-level studies when measurement invariance does not hold; (b) To examine whether unique features of multilevel data, such as intraclass correlation (ICC) and number of clusters, influence the effect of measurement non-invariance on the performance of tests for moderation. Simulation results indicated that falsely assuming measurement invariance lead to biased estimates, inflated Type I error rates, and more gain or more loss in power (depends on simulation conditions) for the test of moderation effects. Such patterns were more salient as sample size and the number of non-invariant items increase for both single- and multi-level data. With multilevel data, the cluster size seemed to have a larger impact than the number of clusters when falsely assuming measurement invariance in the moderation estimation. ICC was trivially related to the moderation estimates. Overall, when testing moderation effects with categorical moderators, employing a model that accounts for the measurement (non)invariance structure of the predictor and/or the outcome is recommended.

OPEN ACCESS

Edited by:

Ehri Ryu,
Boston College, United States

Reviewed by:

Jason C. Immekus,
University of Louisville, United States
Dexin Shi,
University of South Carolina,
United States

*Correspondence:

Yu-Yu Hsiao
yuyhsiao.office@gmail.com

Specialty section:

This article was submitted to
Quantitative Psychology and
Measurement,
a section of the journal
Frontiers in Psychology

Received: 31 July 2017

Accepted: 27 April 2018

Published: 15 May 2018

Citation:

Hsiao Y-Y and Lai MHC (2018) The
Impact of Partial Measurement
Invariance on Testing Moderation for
Single and Multi-Level Data.
Front. Psychol. 9:740.
doi: 10.3389/fpsyg.2018.00740

Keywords: measurement equivalence, measurement invariance, moderation, interaction effects, structural equation modeling, hierarchical linear modeling, multilevel modeling

Many theories in education and psychology rely on moderators, which in Baron and Kenny's (1986) words, "[affect] the direction and/or strength of the relation between an independent or predictor variable and a dependent or outcome variable" (p. 1,174). For many years, social and behavioral researchers are interested in understanding whether a specific moderation effect occurs as well as what factors may influence the extent of the moderation effect. Numerous methodological studies regarding different aspects of moderation effects have been done in contexts such as multiple regression (Aiken and West, 1991), multiple-group structural equation modeling (multiple-group SEM; Jaccard and Wan, 1996), latent variable models with observed composites (Bohrnstedt and Marwell, 1978; Busemeyer and Jones, 1983; Hsiao et al., 2018), within-subject designs (Judd et al., 1996, 2001), cross-level interactions (Kreft et al., 1995), and Bayesian estimations (Lüdtke et al., 2013).

Much of the methodological research regarding moderation effects focused on continuous variables, and less research has been done for categorical moderators. As an example of the latter, researchers may be interested in how the effect of social support on happiness differs by gender. Gender as a categorical variable is treated as the moderator, and social support and happiness are the predictor and outcome variables, respectively. In testing such a moderation with conventional methods such as multiple regression and multiple-group SEM, researchers implicitly assume that the predictor and the outcome variables are measurement invariant across the categorical moderators; that is, the measurement characteristics for social support and happiness are the same by different gender categories. However, such an assumption is seldom investigated before testing moderation effects. Additionally, little is known about how measurement non-invariance influences the estimation of the moderation effects. Hence, it is worth investigating whether measurement invariance for both the predictor and the outcome variables with respect to the moderator categories is a necessary prerequisite before conducting a moderation effect testing.

Measurement invariance (MI) is an important issue in a variety of social and behavioral research settings, especially when the data are collected from multiple populations (Millsap and Kwok, 2004). Full MI holds when individuals with identical ability but from different groups have the same propensity to get a particular score on that specific ability scale (Yoon and Millsap, 2007). Under the multiple-group confirmatory factor analysis framework, a simplified but commonly used version of MI analyses can be conducted by testing four models with hierarchical orders across groups: equal model structures (configural invariance), equal factor loadings (metric invariance), equal intercepts (scalar invariance), and equal unique factor variances (strict invariance; Vandenberg and Lance, 2000; Millsap and Kwok, 2004; Chen et al., 2005; Brown, 2015). Among the four types of MI, metric invariance has been suggested as one basic requirement for doing prediction (Vandenberg and Lance, 2000), which is closely related to moderation effect as moderation effect is about the difference in path coefficients across groups. Hence, in this paper we focus on the impact of metric non-invariance on the estimation of moderation effects. We also focus on testing moderation effects with the multiple-group approach, which is generally being used for examining measurement invariance.

PREVIOUS RESEARCH ON THE EFFECT OF METRIC NON-INVARIANCE ON PREDICTION

Millsap (1995, 1997, 1998, 2007) delineated several theorems and corollaries for the relationship between MI and prediction bias. Donahue (2006) conducted a simulation study to examine the change of the prediction accuracy when the measure of the exogenous (predictor) variable was non-invariant in some part of the factor loadings, or with the presence of partial metric invariance, across groups. Her study found that, if one correctly assumes a partial invariance model on the latent predictors' structures, the path coefficient estimates on the

outcome variables are unbiased even with a larger degree of metric non-invariance (i.e., more non-invariant items) on the latent predictors. However, the study only included the effects on tests of simple regression coefficient in each group, but not moderation, which can be defined as the difference in path coefficients across groups. Additionally, the study did not show the consequences of failing to correctly model the non-invariance structure.

Guenole and Brown (2014) used Monte Carlo studies to investigate the impact of ignoring measurement invariance (including metric invariance) on testing linear and nonlinear effects (including moderation effects). They adopted relative bias of the estimated path coefficients and 95% coverage rate of the estimated confidence intervals from both the reference group and focal group. They found biased estimates of the path coefficients from the two groups when two or more (out of six) ignored non-invariant loadings occurred. The same results were observed when the non-invariance occurred for predictors and outcomes¹.

In the present research, we address two gaps from the work of Donahue (2006) and Guenole and Brown (2014). First, we would show the degree to which estimations and tests of moderation are affected when researchers incorrectly assume that (metric) MI holds. Second, we are interested in whether the location of measurement non-invariance, particularly in the predictor or in the outcome variable, makes a difference. Furthermore, we extend their work by investigating the Type I error rate of misidentifying null moderation effect and the statistical power of detecting nonzero moderation effects in the presence of non-invariance.

Additionally, Donahue (2006) and Guenole and Brown (2014) focused on single level data structure, in which all the observations were assumed to be independent. However, educational and psychological data often have nesting structures (e.g., students nested within classrooms; Kim et al., 2012). For example, a researcher is interested in how the association between students' motivation and their academic achievement differs in public and private schools. Since students are nested within schools, the school variable is a moderator defined in the between level and motivation is a predictor defined in the within level. Therefore, the scenario represents a "cross-level" moderation effects. In this situation, the measurement characteristics of motivation and academic achievement are assumed invariant across school types (i.e., public vs. private). It is still unclear that how multilevel measurement metric (non)invariance across groups in the between level influences the cross-level moderation effects. Therefore, we also show how unique features of multilevel data affect the MI-moderation relationship².

¹One prerequisite to interpret moderation effects in the presence of non-invariance is that the constructs being measured are still conceptually comparable across groups (i.e., configural invariance). If the predictor and/or outcome represent different measurement structure across groups, the computed moderation effect may not be meaningful.

²Throughout this article, we investigated the multilevel measurement (non)invariance across an explicitly defined grouping variable, not the non-invariance for the between-versus within-level. Additionally, we assumed that the outcome variable is defined at the within-level but not at the between-level. Finally, the grouping variable was defined at the between-level but not the within-level in

STUDY 1

In Study 1, we aim to show the effect of measurement non-invariance on the power and Type I error rate when testing a moderator with two categories. Both the predictor and the outcome have a measurement structure and the moderation effects are tested with multiple-group approach, as shown in **Figure 1**. Specifically,

$$\begin{aligned} \mathbf{X}_g &= \lambda_{Xg}F_{Xg} + \delta_g, \\ \mathbf{Y}_g &= \lambda_{Yg}F_{Yg} + \epsilon_g, \\ F_{Yg} &= \gamma_g F_{Xg} + \zeta_g, \end{aligned}$$

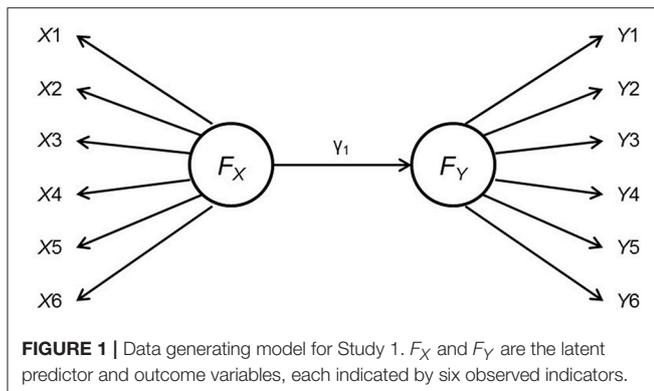
where $g = 1, 2$ was the group index number, $\mathbf{X} = [X_1, X_2, \dots]'$ and $\mathbf{Y} = [Y_1, Y_2, \dots]'$ were observed indicators as shown in **Figure 1**, λ_X and λ_Y were two vectors of factor loadings of the indicators on the latent variables, δ and ϵ were vectors of the effects of unique factor on \mathbf{X} and \mathbf{Y} , γ_g is the path coefficient between F_X and F_Y for group g , and ζ was the latent disturbance term for F_Y . In addition, both the impacts of having metric non-invariance on the outcome and on the predictor were investigated. The simulation study was described below.

Monte Carlo Simulation

The study had a 3 (p_{ni} , number of non-metric-invariant indicators) \times 4 ($\gamma = \{\gamma_1, \gamma_2\}'$, vector of population regression coefficients of the two groups) \times 2 (location of non-invariance) \times 2 (N , sample size of each group) design. In each condition there were two groups, and the sample sizes were assumed equal across groups. Both the predictor F_X and the outcome F_Y were latent variables with six indicators.

Number of Non-metric-Invariant Indicators, p_{ni}

Across the simulation conditions, p_{ni} will either be 0, 2, or 4. For all indicators in Group 1, the factor loadings were set to 0.7, while some of those in Group 2 were set to 0.3 to represent moderate degree of metric non-invariance. This was similar to



the multilevel case. All the mentioned conditions not investigated in the present study may have different implications that are also worthy of future studies.

the conditions in some previous studies (Kaplan and George, 1995; Donahue, 2006).

Regression Coefficients, γ

There were four levels of γ , two of which with equal regression coefficients ($\{0.1, 0.1\}$ and $\{0.5, 0.5\}$) and two with them different ($\{0.5, 0.33\}$ and $\{0.33, 0.5\}$). In the equal γ conditions the grouping variable did not moderate the effects of F_X on F_Y , and Type I error rates were investigated. We were also interested in whether the effect of F_X being large (i.e., 0.5) and small (i.e., 0.1) influences Type I error rates. In the conditions with different γ the effects of F_X on F_Y were different for Group 1 and for Group 2, so there were moderation effects between groups and F_X on F_Y and powers of detecting the true moderation effects were investigated. The numbers were chosen based on the benchmark of small ($\gamma = 0.1$), medium ($\gamma = 0.33$), and large effects ($\gamma = 0.5$; Cohen, 1988).

Location of Non-invariance

The metric non-invariance occurred either on only F_X or only F_Y . Note that this design factor were not applicable to conditions with $p_{ni} = 0$.

Sample Size, N

There were two levels of sample size: 200 and 500, in consistent with some previous studies (e.g., Yoon and Millsap, 2007).

Mplus 7.0 (Muthén and Muthén, 2012) was used to generate 500 data sets for each condition. All variables were assumed multivariate normally distributed. The two factor variances in Group 1 were 1.0 and those in Group 2 were 1.3. For Group 1, the unique factor variances of all indicators were set to 0.51 in the population, so that the invariant indicators had a variance of 1.0. The unique factor variances for Group 2 were set to $0.51 \times 1.3 = 0.663$ so that the proportion of explained variances for the invariant indicators was constant across groups. Because scalar invariance was not the focus of the present study and might not be required for correctly modeling moderation effects, all intercepts and factor means in the population were set to zero.

The data sets generated were then analyzed in Mplus. The analytic model was identified by fixing the factor loadings of the first indicators for F_X and for F_Y to the population value (i.e., 0.7), while allowing the latent factor variances of F_X and of F_Y to be freely estimated. Hence, both F_X and F_Y were scaled to the same unit as the population model and across replications, so that the γ values from the two groups were comparable. To identify the mean structure, the latent factor mean of F_X and the latent intercept of F_Y were fixed to zero for both groups, while the intercepts and the unique factor variances were allowed to be freely estimated without cross-group equality constraints, as scalar and strict invariance conditions were not assumed.

For conditions with $p_{ni} = 0$, the data sets were analyzed by fitting only the model with metric invariance. For other conditions with $p_{ni} > 0$, both the (misspecified) model with metric invariance and the (correct) model with partial metric invariance were fitted. Then for each data set, we obtained the point estimate of $\Delta\hat{\gamma} = \gamma_1 - \gamma_2$ (using the MODEL

CONSTRAINT command in Mplus) and the Wald test statistic (using the MODEL TEST command in Mplus) for the null hypothesis $\gamma_1 = \gamma_2$. Note that we also obtained the results for the likelihood ratio test, which is usually more accurate for finite samples, but we only presented the results for the Wald test as the two tests were nevertheless asymptotically equivalent and produced similar empirical powers and Type I error rates across simulation conditions.

The dependent variables of investigation for the simulations were the percentage of replications where the test statistics were statistically significant at 0.05 level and the standardized bias of $\Delta\hat{\gamma}$. If in the population, $\gamma_1 = \gamma_2$, then the percentage of replications with statistically significant Wald test statistic was the empirical Type I error rate (α^*). Taking into account the sampling variability in 500 replications, an α^* between 3.4% and 7.3% is within the 95% confidence interval when the true Type I error rate is 5%. Empirical Type I error rates over the range of [3.4%, 7.3%] are defined as biased. We expected to see biased Type I error rates and the standardized biases to be large when metric invariance is incorrectly assumed.

If in the population $\gamma_1 \neq \gamma_2$, the percentage where the test statistics were statistically significant at 0.05 level was the empirical power. Given that power is a function of effect size and sample size, the empirical power rates yielded from fitting the model with metric invariance in $p_{ni} = 0$ condition were treated as the baseline; those yielded with $p_{ni} > 0$ from incorrectly assuming measurement invariance and correctly assuming partial invariance models were then compared to the baseline. We expected to see power estimates from models incorrectly assuming measurement invariance were more different from the baseline than the correctly assuming partial invariance models.

Denote $\hat{\gamma}_1^{(i)}$ and $\hat{\gamma}_2^{(i)}$ as the estimated values of γ_1 and γ_2 for the i th replication, and $\bar{\gamma}_1$ and $\bar{\gamma}_2$ as the corresponding means

across replications. The standardized bias (Collins et al., 2001) was computed as

$$\text{standardized bias} = \frac{(\bar{\gamma}_1 - \bar{\gamma}_2) - (\gamma_1 - \gamma_2)}{SD(\hat{\gamma}_1 - \hat{\gamma}_2)},$$

where

$$SD(\hat{\gamma}_1 - \hat{\gamma}_2) = \sqrt{\frac{\sum_{i=1}^R [(\hat{\gamma}_1^{(i)} - \hat{\gamma}_2^{(i)}) - (\bar{\gamma}_1 - \bar{\gamma}_2)]^2}{R}},$$

and $i = 1, 2, \dots, R$ was the index of replications where $R = 500$. The standardized bias was the ratio of the average raw bias over the standard error of the sample estimator of the parameter, and a standardized bias with absolute value < 0.40 was regarded as acceptable (Collins et al., 2001).

Result

The simulation results for the condition with null moderation effects were displayed in **Table 1**. When the measurement invariance assumption held on both the predictor and the outcome in population model (i.e., $p_{ni} = 0$), using the analytic model assuming measurement invariance across groups yielded unbiased moderation effect estimates and unbiased α^* .

When the non-invariance occurred on F_X , as partial metric invariance was the correctly specified model, with a partial invariance model α^* was close to the 0.05 nominal significance level and the moderation effect was estimated with absolute values of standardized bias < 0.02 (< 0.40 as acceptable). On the other hand, α^* was inflated when metric invariance were falsely assumed. The difference between α^* from the nominal level increased as one or more of p_{ni} , N , and the values of γ increased. For example, when $N = 200$, $p_{ni} = 2$, and $\gamma = \{0.1, 0.1\}$, $\alpha^* = 4.2\%$; when $N = 500$, $p_{ni} = 4$, and $\gamma = \{0.1, 0.1\}$,

TABLE 1 | Empirical type I error rate (in percentage) and standardized bias for study 1.

N	γ	p_{ni}	Non-invariance on F_X				Non-invariance on F_Y			
			Type I error (%)		Std. Bias ($\Delta\hat{\gamma}$)		Type I error (%)		Std. Bias ($\Delta\hat{\gamma}$)	
			MI	pMI	MI	pMI	MI	pMI	MI	pMI
200	{0.1, 0.1}	0	4.2	–	0.00	–	–	–	–	–
		2	4.2	4.6	–0.12	0.00	4.2	4.0	0.11	0.01
		4	6.0	4.6	–0.34	–0.01	5.4	4.2	0.33	0.02
	{0.5, 0.5}	0	4.6	–	0.01	–	–	–	–	–
		2	8.0	4.4	–0.64	0.00	7.8	4.2	0.62	0.02
		4	35.2	4.0	–1.60	–0.01	38.8	4.8	1.74	0.02
500	{0.1, 0.1}	0	4.6	–	0.01	–	–	–	–	–
		2	5.4	5.2	–0.18	0.01	5.0	5.2	0.19	0.01
		4	7.6	5.0	–0.51	0.01	7.0	5.0	0.52	0.00
	{0.5, 0.5}	0	4.0	–	–0.01	–	–	–	–	–
		2	14.4	3.6	–1.06	–0.01	13.8	4.8	0.97	–0.01
		4	77.8	3.2	–2.79	–0.01	77.8	5.2	2.74	–0.01

p_{ni} , number of non-metric-invariant indicators; γ , population regression coefficient of F_Y on F_X ; MI, analytic model assumed metric invariance; pMI, analytic model correctly assumed partial metric invariance; Std. Bias, standardized bias = $\Delta\hat{\gamma}/SD(\Delta\hat{\gamma})$, where $SD(\Delta\hat{\gamma})$ is the standard deviation of the differences in the estimated γ s across all replications.

$\alpha^* = 7.6\%$; and when $N = 500$, $p_{ni} = 4$, and $\gamma = \{0.5, 0.5\}$, $\alpha^* = 77.8\%$. An analysis of variance (ANOVA) including N , γ , and p_{ni} showed that p_{ni} produced the largest impact on α^* ($\eta^2 = 0.34$), followed by γ ($\eta^2 = 0.21$) and N ($\eta^2 = 0.04$). The bias of the estimated values of $\Delta\gamma$ followed a similar pattern. For instance, With $N = 500$, $p_{ni} = 4$, and $\gamma = \{0.5, 0.5\}$, the standardized bias of the null moderation effects was -2.79 , which was a substantial bias.

The pattern of α^* and the absolute values of the standardized bias when non-invariance occurred in F_Y was very similar to those when non-invariance occurred in F_X . However, the sign of the standardized bias was reversed, which means that when non-invariance occurred in the outcome's structure, the moderation effects were overestimated. Considering both the locations of the non-invariance, we found that using models that incorrectly assumed measurement invariance would result in substantially biased moderation effect estimate and inflated Type I error rate.

Table 2 showed the results of both the powers and standardized biases with nonzero moderation effects. When the non-invariance occurred on F_X , the corrected partial metric invariance models performed well as they showed no bias on the moderation effect estimates with standardized biases from -0.03 to 0.01 . On the contrary, the metric invariance model yielded biased estimates of the moderation effects and the influence was more salient as both N and p_{ni} increased. For example, when $\gamma = \{0.5, 0.33\}$, the standardized bias was -0.43 with $N = 200$ and $p_{ni} = 2$; the standardized bias increased to -1.84 with $N = 500$ and $p_{ni} = 4$. An ANOVA showed that p_{ni} produced the largest impact on the biased moderation estimates ($\eta^2 = 0.79$), followed by N ($\eta^2 = 0.09$) and γ ($\eta^2 = 0.01$).

In terms of the powers for detecting the moderation effects, the corrected partial invariance model yielded powers around 30% and 60% for N equals 200 and 500, respectively. Such power estimates were close to population model with the measurement

invariance assumption held (33% for $N = 200$ and 70% for $N = 500$). On the other hand, if metric invariance was falsely assumed, there was a substantial decrease in powers for the conditions where non-invariance occurred. For example with $\gamma = \{0.5, 0.33\}$, $N = 500$, $p_{ni} = 2$, and non-invariance on F_X , the empirical power was half as would be obtained when metric invariance held in the population (33.8% vs. 70.2%); with $\gamma = \{0.33, 0.5\}$, $N = 200$, $p_{ni} = 4$, and non-invariance on F_Y , the empirical power was only 1/8 as the power would be obtained when metric invariance held in the population (4.2% vs. 33.0%).

Note that power loss was detected as both the N and p_{ni} increased when the non-invariance occurred on F_X and $\gamma = \{0.5, 0.33\}$; simulation conditions related to non-invariance occurred on F_Y and $\gamma = \{0.33, 0.5\}$ would lead to inflated power estimates as both the N and p_{ni} increased. The main reason for different patterns on the power estimates were that when the factor loadings of Group 1 (0.7) was larger than those of Group 2 (0.3) in the presence of non-invariance on F_X , the estimated moderation effect was negatively biased, whereas when non-invariance occurred in F_Y , the estimated moderation effect was positively biased. Additionally, the true moderation effect was -0.17 when $\gamma = \{0.33, 0.5\}$; therefore, the negative biases caused by falsely assuming measurement invariance would result in more negative moderation effects estimates and inflated power.

STUDY 2

In Study 2, we aim to extend the scope of the MI-moderation relation to multilevel data. We focused on how the measurement (non-)invariance across groups at the between level influences the test of cross-level moderation effect, which was one of the prevailing issues among social and behavioral research.

TABLE 2 | Empirical power (in percentage) and standardized bias for study 1.

N	γ	p_{ni}	Non-invariance on F_X				Non-invariance on F_Y			
			Power (%)		Std. Bias ($\Delta\hat{\gamma}$)		Power (%)		Std. Bias ($\Delta\hat{\gamma}$)	
			MI	pMI	MI	pMI	MI	pMI	MI	pMI
200	{0.5, 0.33}	0	32.2	–	–0.01	–	–	–	–	–
		2	16.0	30.8	–0.43	–0.02	53.0	31.2	0.48	0.00
		4	4.4	31.2	–1.12	–0.03	81.8	29.4	1.29	0.02
	{0.33, 0.5}	0	33.0	–	0.02	–	33.0	–	–	–
		2	50.0	30.8	–0.61	0.01	15.8	30.0	0.52	0.03
		4	77.0	26.8	–1.57	–0.01	4.2	27.4	1.52	0.04
500	{0.5, 0.33}	0	70.2	–	–0.01	–	–	–	–	–
		2	33.8	67.2	–0.69	–0.02	89.0	67.4	0.76	–0.01
		4	5.0	62.4	–1.84	–0.02	99.2	62.4	2.01	–0.02
	{0.33, 0.5}	0	67.6	–	0.00	–	–	–	–	–
		2	90.6	67.2	–1.01	–0.01	36.0	66.2	0.79	0.00
		4	99.4	59.4	–2.71	–0.01	5.0	62.8	2.38	0.00

p_{ni} , number of non-metric-invariant indicators; γ , population regression coefficient of F_Y on F_X ; MI, analytic model assumed metric invariance; pMI, analytic model correctly assumed partial metric invariance; Std. Bias, standardized bias = $\Delta\hat{\gamma}/SD(\Delta\hat{\gamma})$, where $SD(\Delta\hat{\gamma})$ is the standard deviation of the differences in the estimated γ s across all replications.

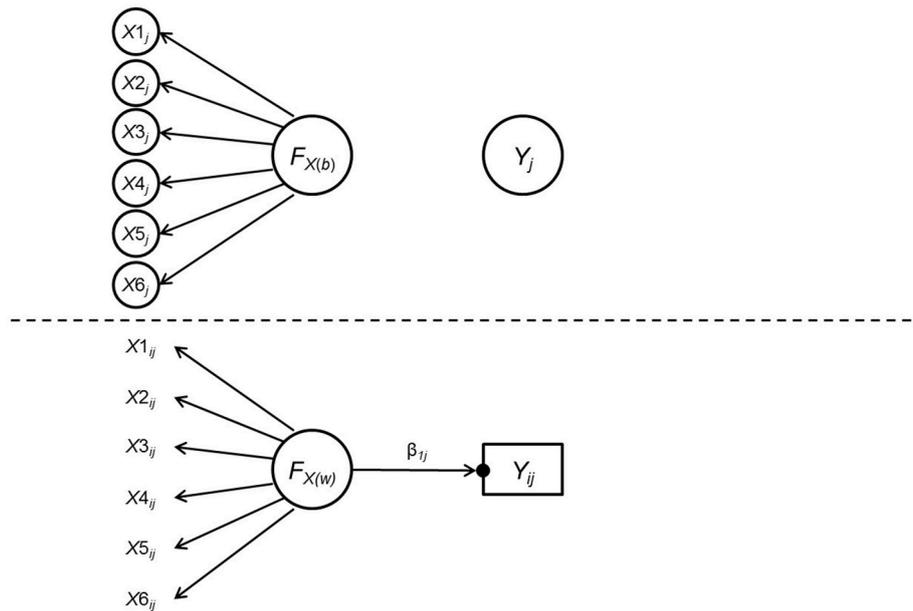


FIGURE 2 | Data generating model for Study 2. $F_{X^{(w)}}$ and $F_{X^{(b)}}$ are the latent predictor variable at the within-level and the between-level, respectively. Y_{ij} and Y_j are the within-level and the between-level components of the outcome variable Y . β_{ij} = within-level regression coefficient of Y on $F_{X^{(w)}}$, whose magnitude varies across clusters as indicated by the black dot. Conditioning on the grouping variable, measurement invariance was assumed across clusters such that the within-level and the between-level factors loadings were identical, and that there were no residual variances for the six indicators at the between-level.

Specifically, we used the data generating model shown in **Figure 2**, which was one of the simplest models including multilevel measurement (non-)invariance and a within-level predictor, to depict the cross-level moderation effect. As can be seen in **Figure 2**, the latent predictor was measured by six indicators and the cross-level interaction effect was denoted by the difference between the within-level path coefficient from the predictor to the outcome across groups. It was assumed that the predictor did not have an effect on the outcome in the between level.

Because multilevel data are usually of larger sample size, we expect the impact of multilevel non-invariance on the Type I error rate and power to be bigger. In addition, we are interested in whether the impact varies across multilevel specific design factors such as the intraclass correlation (ICC), number of clusters, and cluster size. Because in Study 1, we found that different locations of non-invariance mainly resulted in changes in signs of the biases of the moderation effects, in Study 2 we only focused on measurement non-invariance on the predictor side. Likewise, we only consider the positive moderation effects condition in Study 2 given that negative moderation effect led to similar results in biases in Study 1. A second Monte Carlo simulation study was conducted, as described below.

Monte Carlo Simulation

The study had a $2 (p_{ni}) \times 2 (\gamma) \times 2$ (ICC, intraclass correlation) $\times 2 (m, \text{ number of clusters}) \times 2 (c, \text{ cluster size})$ design. In each condition there were two groups (Group 1 and Group 2), and sample sizes (both within and between) were assumed

equal across groups. The latent predictor, F_X , had the same six-indicator measurement structure in both the within and the between level as in Study 1; the observed outcome, Y_{ij} , of the i th observation in the j th cluster, contained no measurement error and was assumed measurement invariant. The model can be expressed as

$$\begin{aligned} \text{Within level: } Y_{ijg} &= Y_{(b)jg} + \gamma_{10g} F_{X^{(w)ijg}} + r_{ijg}, \\ \mathbf{X}_{ijg} &= \mathbf{X}_{(b)jg} + \boldsymbol{\lambda}_{(w)g} F_{X^{(w)ijg}} + \boldsymbol{\delta}_{(w)ijg}; \\ \text{Between level: } Y_{(b)jg} &= \gamma_{00g} + \zeta_{jg}, \\ \mathbf{X}_{(b)jg} &= \alpha_g + \boldsymbol{\lambda}_{(b)g} F_{X^{(b)jg}} + \boldsymbol{\delta}_{(b)jg}, \end{aligned}$$

where $\mathbf{X} = [X_1, X_2, \dots]'$ was a vector containing the observed values of the indicators, and their group means comprised $\mathbf{X}_{(b)}$. The vectors $\boldsymbol{\lambda}_{(w)}$ and $\boldsymbol{\lambda}_{(b)}$ contained the within-level and between-level factor loadings, respectively. In this study we assumed that $\boldsymbol{\lambda}_{(w)g} = \boldsymbol{\lambda}_{(b)g} = \boldsymbol{\lambda}_g$. In the within level, $F_{X^{(w)}}$ (the within-level exogenous factor) had an effect of magnitude γ_{10g} on Y , where g is the group index. In the between level, $F_{X^{(b)}}$ (the between level exogenous factor) had no effect on Y . Note that there were no between-level random effects on γ_{10g} and on the factor loadings. We also assumed measurement invariance across clusters, implying that $\boldsymbol{\delta}_{(b)jg} = \mathbf{0}$ and homogeneous $\boldsymbol{\delta}_{(w)ijg}$ across clusters (Jak et al., 2013), in addition to $\boldsymbol{\lambda}_{(w)g} = \boldsymbol{\lambda}_{(b)g}$. The design factors were described below.

Number of Non-metric-Invariant Indicators, p_{ni}

p_{ni} was either 0 or 2 out of the six indicators of F_X . Whereas two of the factor loadings were always set to 0.7 in Group 1, for

conditions with $p_{ni} = 2$, those loadings were set to 0.3 in Group 2. The factor loadings for other four indicators were 0.7, 0.3, 0.5, and 0.6, for both the within level and the between level.

Regression Coefficients, γ

There were two levels of γ : {0.3, 0.2} (moderation present) and {0.3, 0.3} (moderation absent).

Intraclass Correlation, ICC

Based on previous simulations (Kim et al., 2012), in this study there were two levels of ICC: 0.10 and 0.35, representing small and large within-cluster correlations for the latent variable F_X and for the outcome Y .

Cluster Size, c

Based on previous literature (Clarke, 2008; Kim et al., 2012), there were two levels of cluster size: 5 and 20, representing small and medium number of observations within a cluster. For simplicity we generated data with all clusters having the same size in both groups.

Number of Clusters, m

Hox and Maas (2001) suggested the number of groups larger than 100 as the minimum requirement for yielding accurate multilevel regression estimates. Later on, Maas and Hox (2005) found groups number equal to 30 could also yield accurate multilevel regression estimates. McNeish (2017) did a literature review on 70 multilevel studies and found 90% of them fail to meet Hox and Maas's criterion of 100 clusters, and that the median number of clusters was 44. In the present study, we specified the number of clusters in each group either 30 or 100, representing the small and large number of clusters.

Mplus 7.0 was used to generate and analyze (with ESTIMATOR=MLR) 500 data sets for each condition. All exogenous variables and random effects were assumed multivariate-normally distributed. For both groups the variances of $F_{X(w)}$ and Y both equaled to 1.0, and that of $F_{X(b)}$ and ζ_{jg} were functions of the ICC. The variance of $\delta_{(w)ijg}$ was set to 0.51 so that the level-1 unique factor variances were similar in values to those in Study 1 (i.e., 0.51 in Study 1 when the latent factor variance is one). The covariance and mean structure were identified similarly as in Study 1 by fixing the factor loadings of the first indicators for $F_{X(b)}$ to the population value and the latent mean of $F_{X(b)}$ to zero for both groups. Additionally, within the same group the factor loadings were constrained to be equal in the between and the within levels so that metric invariance was assumed across clusters (Jak et al., 2013). Because scalar invariance was not the focus of the present study and may not be required for correctly modeling moderation effects, all intercepts and factor means in the population were set to zero.

The dependent variables of investigation were the standardized biases and the rejection rates of the Wald test statistics for the difference in γ_{10} , which reflected either the empirical Type I error rate (α^*) or the empirical power, and were obtained in the same manner as in Study 1. Also as in Study 1, for conditions with $p_{ni} = 2$, both the metric invariance model and the partial metric invariance model were fitted. We expected that model falsely assuming measurement invariance would lead

to biased moderation estimation, inflated Type I error rate (when $p_{ni} = 0$), and power more different from the baseline (when $p_{ni} = 2$).

Result

Results for Study 2 were shown in **Table 3**. In the conditions absent of moderation effects ($\gamma = \{0.3, 0.3\}$), fitting data with a metric invariance model when the true population model followed the measurement invariance assumption ($p_{ni} = 0$) led to unbiased moderation effect estimates and unbiased α^* , regardless of the level of ICC, m , and c . The same pattern was observed while employing the corrected partial metric invariance model to fit data from a measurement non-invariance population, as such practice also led to unbiased moderation estimates and α^* close to the 5% nominal significance level across different ICC, m , and c simulation conditions.

When non-invariance occurred ($p_{ni} = 2$), fitting data with a metric invariance model yielded substantially underestimated moderation effect and inflated α^* . Such trend became more salient as m and c increased. For example with ICC = 0.10, $p_{ni} = 2$, $m = 30$, and $c = 5$, the standardized bias was -0.55 with α^* of 9%; with ICC = 0.10, $p_{ni} = 2$, $m = 100$, and $c = 20$, the standardized bias increased to -2.13 with α^* of 55.2%. An ANOVA analysis on the standardized bias with p_{ni} , m , c , and ICC showed that p_{ni} had the largest impact on estimation biases ($\eta^2 = 0.70$), followed by c ($\eta^2 = 0.08$), m ($\eta^2 = 0.06$), and ICC (η^2 close to 0). ICC showed no impact of falsely assuming metric invariance on yielding biased moderation estimates and inflated α^* . For example with ICC = 0.10, $p_{ni} = 2$, $m = 100$, and $c = 5$, the standardized bias was -1.07 with α^* of 16.4%; increasing the ICC to 0.35 while keeping the other design factors to be the same led to similar results with standardized bias = -1.04 and $\alpha^* = 15.8\%$.

In the conditions with nonzero moderation effect (γ : {0.3, 0.2}), again, employing the correctly specified partial invariance model resulted in unbiased moderation effect estimates. On the other hand, falsely assuming metric invariance led to substantially underestimated moderation effects across simulation conditions with standardized biases from -0.39 to -1.61 . Consistent with the null moderation condition, an ANOVA analysis on the standardized bias indicated p_{ni} had the largest impact on estimation biases ($\eta^2 = 0.70$), followed by c ($\eta^2 = 0.10$), m ($\eta^2 = 0.06$), and ICC (η^2 close to 0). There was also a substantial loss in power when fitting a metric invariance model to data draw from a population with non-invariance. For example, the power of the simulation condition of ICC = 0.10, $p_{ni} = 0$, $m = 100$, and $c = 20$ was 80% but it dropped to 16.8% when ICC = 0.10, $p_{ni} = 2$, $m = 100$, and $c = 20$. Again, ICC only had a trivial effect on the deflation of power.

DISCUSSION

In the literature, the impact of measurement invariance on testing moderation effects has not been fully examined. The ratio of the non-invariant items have been found to be an important factor on the estimation accuracy of the path coefficients by the moderating groups (e.g., Guenole and Brown, 2014). In much of

TABLE 3 | Empirical type I error rate, power, and standardized bias for study 2.

ICC	p_{ni}	m	c	$\gamma = \{0.3, 0.3\}$				$\gamma = \{0.3, 0.2\}$			
				Type I Error (%)		Std. Bias ($\Delta\hat{\gamma}$)		Power (%)		Std. Bias ($\Delta\hat{\gamma}$)	
				MI	pMI	MI	pMI	MI	pMI	MI	pMI
0.10	0	30	5	6.60	–	0.01	–	15.00	–	0.01	–
			20	7.20	–	–0.02	–	34.80	–	–0.02	–
		100	5	5.80	–	0.02	–	31.20	–	0.02	–
			20	6.00	–	0.04	–	81.00	–	0.03	–
	2	30	5	9.00	5.80	–0.55	0.01	7.00	14.20	–0.41	0.01
			20	25.60	7.00	–1.17	–0.02	10.00	34.40	–0.91	0.02
		100	5	16.40	6.40	–1.07	0.01	4.80	28.60	–0.81	0.02
			20	55.20	5.80	–2.13	0.03	16.80	79.20	–1.60	0.03
0.35	0	30	5	6.20	–	0.02	–	15.20	–	0.02	–
			20	7.40	–	–0.02	–	35.20	–	–0.03	–
		100	5	5.40	–	0.01	–	30.80	–	0.02	–
			20	5.40	–	0.03	–	80.60	–	–0.03	–
	2	30	5	7.60	5.20	–0.53	0.02	6.80	14.60	–0.39	0.02
			20	25.20	7.20	–1.16	–0.02	10.20	32.80	–0.91	–0.02
		100	5	15.80	5.80	–1.04	0.01	4.60	25.80	–0.79	0.02
			20	54.60	5.40	–2.13	0.03	18.20	78.20	–1.61	0.03

γ , population regression coefficient of F_Y on F_X ; ICC, intraclass correlation; p_{ni} , number of non-metric-invariant indicators; m , number of clusters; c , number of observations in a cluster; MI, analytic model assumed metric invariance; pMI, analytic model assumed partial metric invariance; Std. Bias, standardized bias = $\Delta\hat{\gamma}/SD(\Delta\hat{\gamma})$, where $SD(\Delta\hat{\gamma})$ is the standard deviation of the differences in the estimated γ s across all replications.

previous work, the focus was limited to single level data structure, without considerations of nested data structure. Additionally, the direct statistical test of the moderation effect was largely ignored in previous research. The current study investigated the impact of partial measurement invariance, with a focus on the metric invariance, on the estimation and testing of moderation effects on both single and multilevel structures, in terms of standardized bias, power and Type I error rate.

The results suggest that incorrectly assuming metric invariance holds while estimating moderation effects would lead to biased estimates. The impact is more salient as the number of non-invariant items increases, which is consistent with Guenole and Brown (2014)'s and Shi et al. (2017)'s findings with direct effects. On the other hand, fitting models correctly assuming partial metric invariance yielded accurate estimates regardless of samples size, main effects, number of non-invariant items, and the location of the non-invariance occurred.

In testing null moderation effects (i.e., γ s are equal between two groups), the high Type I error rate yielded from falsely assuming metric invariance is not only related to the non-invariant item ratio but the magnitude of the main effects. These results suggest that evaluation of measurement invariance is of more importance when the main effect of the predictor is larger. On the other hand, the Type I error rates were on or below 5% with models correctly assuming partial metric invariance. Thus, before examining moderation effects, the metric invariance assumption should not be presumed without conducting any invariance test, even for cases in which the moderation tests turn out to be non-significant.

The location of the non-invariance (predictor vs. outcome) is associated with the direction of the biases of the moderation effects. In our simulations, all of the non-invariance conditions were specified such that factor loadings of Group 1 were equal or larger than those of Group 2. As evident from the simulation results, under such settings ignoring predictor non-invariance leads to underestimation of the moderation effects, whereas ignoring outcome non-invariance results in overestimated moderation effects. Our findings are consistent with Chen (2008) and Guenole and Brown (2014), in which they found that non-invariance on the predictor with lower factor loadings in group 2 would lead to underestimated path coefficient in group 1 (γ_1) and overestimated path coefficient in group 2 (γ_2). Hence, the moderation effect ($\gamma_1 - \gamma_2$) would likely be underestimated. On the other hand, non-invariance on the outcome changes the association to the opposite direction and results in the overestimation of the moderation effects.

Compared with models correctly assuming partial metric invariance, models falsely assuming metric invariance yielded moderation test with statistical power varying substantially. Taking into account the signs of the moderation effects, when the moderation effects are positive, ignoring non-invariance on the predictors leads to power loss, but ignoring non-invariance on the outcomes leads to increased power (at the cost of highly inflated Type I error rate). Likewise, an opposite association between the location of non-invariance and power is observed when the moderation effects are negative. Therefore, the increase in power in half of our simulation conditions in **Table 2** is actually a byproduct of sacrificing the estimation accuracy of

the moderation effects (i.e., overestimation). Ignoring non-invariance and resulting in power gain or loss depends on (a) the location of the non-invariance, (b) the signs of the moderation effects. Overall, it is not recommended to fit a model assuming metric invariance when the assumption is actually violated, even though it may increase the power of the moderation test.

There is also prospective evidence that falsely assuming multilevel metric invariance across groups has a negative impact on the estimation of the cross-level moderation effects, which leads to either substantially inflated Type I error rate or inflated/deflated statistical power of the moderation test. Both increases in m and in c , or in other words an increase in the total sample size, resulted in bigger problems in the estimation accuracy as well as α^* and power. Thus, for multilevel data, even with only one-third of the indicators being non-metric-invariant, tests of moderation can become hugely misleading.

Across simulation conditions, the number of non-invariant items played a huge role in influencing the performance of the moderation estimates. Researchers use multilevel data with a number of clusters (e.g., number of classrooms) larger than 100 or cluster size larger than 20 (e.g., 20 students in each classroom) should be particularly cautious about the negative impact of non-invariant items. Additionally, the cluster size (c) seemed to have a larger impact than the number of clusters (m) when falsely assuming measurement invariance in the moderation estimation. Intraclass correlation (ICC) was trivially related to the moderation estimates, probably because the path of interest was defined in the within-level. On the contrary, previous research has shown that ICC is highly related to between-level analysis (Kim et al., 2012). Thus, one potential explanation for the discrepancy is that, given that the cross-level moderation coefficients were mainly defined in the within-level, the level of data dependency has less influence on the moderation effect estimates.

Findings from Study 1 and Study 2 highlight the importance of testing metric invariance before conducting a moderation test with both single and multilevel data structure. If the metric invariance assumption is violated, a partial metric invariance model in which the non-invariant factor loadings between groups are correctly reflected should be employed. Researchers should also be aware that the MI-moderation relationship is highly affected by the ratio of non-invariant items in the scale

and the overall sample size. Overall, while testing moderation effects in a multiple-group analysis setting, we recommend the test of measurement invariance for both the predictors and outcomes by the moderator groups. If the measurement invariance assumption holds, then employing models with such an assumption implied is appropriate. On the other hand, if the measurement invariance assumption is violated, then the use of a corrected partial invariance model would yield more accurate estimates and unbiased Type I error rates or power.

Some limitations and future study directions should be addressed. First, the research scenario only focused on metric invariance (i.e., invariance of the factor loadings). In practice, non-invariance may exist in the intercepts, factor loadings, unique factor variances, or some combinations of them. Previous simulation studies on latent growth modeling have shown that ignoring intercept non-invariance only leads to biased factor mean (or intercept) estimates (Kim and Willson, 2014). Research on multiple-group analysis also showed that ignoring intercept non-invariance has less impact on the prediction bias of the path coefficient in each group (Guenole and Brown, 2014). Therefore, we suspect that the impact of intercept non-invariance on the moderation effect estimates should be much smaller than that of factor loading non-invariance, but more conclusive evidence needs to be obtained from future methodological inquiries.

Second, for Study 2 we only tested cross-level moderation in the present study, but moderation effects may also occur at the between level, in which factors such as ICC may play a more important role in affecting the moderation estimates. Lastly, in the simulation, the indicators were assumed to be continuous and normal distributed when conditioned on the latent factors. It is important to see how measurement non-invariance with skewed and categorical indicators influence the estimation of the moderation effects. Therefore, future study can investigate the impact of falsely assuming measurement invariance under more complicated research settings.

AUTHOR CONTRIBUTIONS

Y-YH led the implementation and manuscript writing of the study. ML designed and conducted the simulation. Both authors (Y-YH and ML) contributed to the design, analysis, interpretation of data, writing, and revising of the manuscript.

REFERENCES

- Aiken, L. S., and West, S. G. (1991). *Multiple Regression: Testing and Interpreting Interactions*. Thousand Oaks, CA: Sage Publications, Inc.
- Baron, R. M., and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* 51, 1173–1182.
- Bohrnstedt, G. W., and Marwell, G. (1978). The reliability of products of two random variables. *Sociol. Methodol.* 9:254.
- Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research, 2nd Edn.* New York, NY: Guilford.
- Busemeyer, J. R., and Jones, L. E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychol. Bull.* 93, 549–562.
- Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *J. Pers. Soc. Psychol.* 95, 1005–1018. doi: 10.1037/a0013193
- Chen, F. F., Sousa, K. H., and West, S. G. (2005). Teacher's corner: testing measurement invariance of second-order factor models. *Struct. Equat. Model.* 12, 471–492. doi: 10.1207/s15328007sem1203_7
- Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *J. Epidemiol. Community Health* 62, 752–758. doi: 10.1136/jech.2007.060798.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn.* Hillsdale, NJ: Erlbaum.
- Collins, L. M., Schafer, J. L., and Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol. Methods* 6, 330–351. doi: 10.1037/1082-989X.6.4.330

- Donahue, B. H. (2006). *The Effect of Partial Measurement Invariance on Prediction*. Ph.D. thesis, University of Georgia, Athens.
- Guenole, N., and Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Front. Psychol.* 5:980. doi: 10.3389/fpsyg.2014.00980
- Hox, J. J., and Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Struct. Equat. Model.* 8, 157–174. doi: 10.1207/S15328007SEM0802_1
- Hsiao, Y.-Y., Kwok, O.-M., and Lai, M. H. C. (2018). Evaluation of two methods for modeling measurement errors when testing interaction effects with observed composite scores. *Educ. Psychol. Meas.* 78, 181–202. doi: 10.1177/0013164416679877
- Jaccard, J., and Wan, C. K. (1996). *LISREL Approaches to Interaction Effects in Multiple Regression*. Thousand Oaks, CA: Sage.
- Jak, S., Oort, F. J., and Dolan, C. V. (2013). A test for cluster bias: detecting violations of measurement invariance across clusters in multilevel data. *Struct. Equat. Model.* 20, 265–282. doi: 10.1080/10705511.2013.769392
- Judd, C. M., Kenny, D. A., and McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychol. Methods* 6, 115–134. doi: 10.1037/1082-989X.6.2.115
- Judd, C. M., McClelland, G. H., and Smith, E. R. (1996). Testing treatment by covariate interactions when treatment varies within subjects. *Psychol. Methods* 1, 366–378.
- Kaplan, D., and George, R. (1995). A study of the power associated with testing factor mean differences under violations of factorial invariance. *Struct. Equat. Model.* 2, 101–118.
- Kim, E. S., Kwok, O.-M., and Yoon, M. (2012). Testing factorial invariance in multilevel data: a Monte Carlo study. *Struct. Equat. Model.* 19, 250–267. doi: 10.1080/10705511.2012.659623
- Kim, E. S., and Willson, V. L. (2014). Measurement invariance across groups in latent growth modeling. *Struct. Equat. Model.* 21, 408–424. doi: 10.1080/10705511.2014.915374
- Kreft, I. G., de Leeuw, J., and Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivar. Behav. Res.* 30, 1–21.
- Lüdtke, O., Robitzsch, A., Kenny, D. A., and Trautwein, U. (2013). A general and flexible approach to estimating the social relations model using Bayesian methods. *Psychol. Methods* 18, 101–119. doi: 10.1037/a0029252
- Maas, C. J. M., and Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology* 1, 86–92. doi: 10.1027/1614-1881.1.3.86
- McNeish, D. (2017). Multilevel mediation with small samples: a cautionary note on the multilevel structural equation modeling framework. *Struct. Equat. Model.* 24, 609–625. doi: 10.1080/10705511.2017.1280797
- Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivar. Behav. Res.* 30, 577–605.
- Millsap, R. E. (1997). Invariance in measurement and prediction: their relationship in the single-factor case. *Psychol. Methods* 2, 248–260.
- Millsap, R. E. (1998). Group differences in regression intercepts: implications for factorial invariance. *Multivar. Behav. Res.* 33, 403–424.
- Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika* 72, 461–473. doi: 10.1007/S11336-007-9039-7
- Millsap, R. E., and Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychol. Methods* 9, 93–115. doi: 10.1037/1082-989X.9.1.93
- Muthén, L. K., and Muthén, B. O. (1998–2012). *Mplus User's Guide, 7th Edn.* Los Angeles, CA: Muthén & Muthén.
- Shi, D., Song, H., and Lewis, M. D. (2017). The impact of partial factorial invariance on cross-group comparisons. *Assessment* 1:1073191117711020. doi: 10.1177/1073191117711020
- Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–70. doi: 10.1177/109442810031002
- Yoon, M., and Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: a Monte Carlo study. *Struct. Equat. Model.* 14, 435–463. doi: 10.1080/10705510701301677

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Hsiao and Lai. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.