HHS Public Access

Author manuscript

*Dev Rev*. Author manuscript; available in PMC 2017 September 01.

Published in final edited form as:

*Dev Rev*. 2016 September ; 41: 71–90. doi:10.1016/j.dr.2016.06.004.

# Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research

**Diane L. Putnick** and **Marc H. Bornstein**

*Eunice Kennedy Shriver* National Institute of Child Health and Human Development

## Abstract

Measurement invariance assesses the psychometric equivalence of a construct across groups or across time. Measurement noninvariance suggests that a construct has a different structure or meaning to different groups or on different measurement occasions in the same group, and so the construct cannot be meaningfully tested or construed across groups or across time. Hence, prior to testing mean differences across groups or measurement occasions (e.g., boys and girls, pretest and posttest), or differential relations of the construct across groups, it is essential to assess the invariance of the construct. Conventions and reporting on measurement invariance are still in flux, and researchers are often left with limited understanding and inconsistent advice. Measurement invariance is tested and established in different steps. This report surveys the state of measurement invariance testing and reporting, and details the results of a literature review of studies that tested invariance. Most tests of measurement invariance include configural, metric, and scalar steps; a residual invariance step is reported for fewer tests. Alternative fit indices (AFIs) are reported as model fit criteria for the vast majority of tests; $\chi^2$ is reported as the single index in a minority of invariance tests. Reporting AFIs is associated with higher levels of achieved invariance. Partial invariance is reported for about one-third of tests. In general, sample size, number of groups compared, and model size are unrelated to the level of invariance achieved. Implications for the future of measurement invariance testing, reporting, and best practices are discussed.

### Keywords

Comparative psychology; measurement invariance; structural equation modeling; confirmatory factor analysis; multiple-group analysis

Measurement invariance assesses the (psychometric) equivalence of a construct across groups or measurement occasions and demonstrates that a construct has the same meaning to those groups or across repeated measurements. Measurement invariance takes many forms and is key to psychological and developmental research because it is a prerequisite to comparing group means. Measurement invariance applies to group comparisons, to mean

Address Correspondence to: Diane L. Putnick, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, 6705 Rockledge Drive, Suite 8030, Bethesda, MD 20892, 301-496-6291, putnickd@mail.nih.gov.

comparisons across measurement occasions, and to differential relations between constructs by group (e.g., interactions by group), all of which are staples in psychological and developmental science. It applies equally to research topics in clinical, cognitive, social, experimental, cross-cultural, and developmental psychology. For example, before testing mean differences in a construct, researchers should test for invariance across child genders (Hong, Malik, & Lee, 2003), mothers and fathers (Wang et al., 2006), ethnic groups (Glanville & Wildhagen, 2007), and cultural groups (Senese, Bornstein, Haynes, Rossi, & Venuti, 2012). Because the interpretation of a construct can change over time, developmental researchers should test for invariance across measurement occasions (e.g., invariance across time; Little, 2013; Widaman, Ferrer, & Conger, 2010) and even for pretests and posttests before and after interventions (Nolte, Elsworth, Sinclair, & Osborne, 2009).

A study may measure the same cognition or behavior (form) across groups or times, but that cognition or behavior can have a different meaning (function) for the different groups or at different times (Bornstein, 1995). Like the meaning derived from sound-word associations in language, meaning is essentially conventionalized, and so different groups can apply different meanings to the same cognition or behavior. Appropriate and proper comparison of a construct between groups or across times, therefore, depends first on ensuring equivalence of meaning of the construct. The untoward consequences of measurement noninvariance can be readily illustrated in the study of depression in men and women. Suppose frequency of crying, weight gain, and feelings of hopelessness are indicative of the severity of depression in women, but only feelings of hopelessness are indicative of the severity of depression in men. If the three indicators are combined into a scale to compare depression in women and men, mean differences on the scale may mislead because crying and weight gain have little relation to depression in men. In this example, men may score lower than women on the depression scale because they cry less and gain less weight. However, crying and weight gain are not associated with depression in men in the first place. Another example comes from experimental designs. It is common to assess the effectiveness of an intervention, protocol, or trial by comparing pretest and posttest scores, or treatment and control groups. However, the intervention, protocol, or trial could change the way participants interpret the constructs under study. In an investigation of the effects of a training program to improve reading fluency, the intervention group may have learned to rely on one component of fluency (e.g., prosody – intonation patterns and phrasing), rather than fluency as a whole (including accuracy and automaticity) which could affect the measurement of the fluency construct (e.g., prosody could have a stronger loading on fluency in the treatment group or on the posttest). If the measure used at pretest and posttest (or in treatment and control groups) is noninvariant, then it is no longer clear that the change in reading fluency overall can be attributed to the intervention. Rather, it could be that the intervention only improves one aspect of fluency. Hence, noninvariance of a construct across groups or measurements can lead to erroneous conclusions about the effectiveness of a trial. As is clear, measurement invariance is a central and pervasive feature of psychological and developmental science.

In this report, we provide a brief history and non-technical description of measurement invariance, review current practices for testing and reporting measurement invariance, and discuss best practices and future directions for measurement invariance testing in psychological research.

## History of Measurement Invariance

The conceptual importance of testing measurement invariance entered the literature more than 50 years ago (e.g., Meredith, 1964; Struening & Cohen, 1963), but statistical techniques for testing invariance have become more accessible to, and expected from, the research community only relatively recently. Indeed, measurement invariance is fast becoming *de rigueur* in psychological and developmental research. Around the turn of the 21st century, methodologists increasingly directed attention to the significance of measurement invariance, especially within a structural equation modeling framework (Cheung & Rensvold, 1999, 2002; Little, 2000; Rensvold & Cheung, 1998; Steenkamp & Baumgartner, 1998; Vandenberg, 2002). In landmark papers, Widaman and Reiss (1997) and Vandenberg and Lance (2000) synthesized the measurement invariance literature, delineated the ladder-like approach to measurement invariance testing, and provided researchers with step-by-step guides to conducting invariance tests (interested readers should turn back to these articles for a more thorough accounting of the history and mathematics supporting measurement invariance; see also Millsap, 2011). As Vandenberg (2002) noted shortly thereafter, “there is adoption fervor” (p. 140) with respect to measurement invariance. This fervor continued to crest in the succeeding decade, but it has not been accompanied by consistent or adequate advice, explication, best practices, or understanding.

## Testing Measurement Invariance

Measurement invariance can be tested in an item-response theory (IRT) framework or a structural equation modeling (SEM) framework, and some contemporary researchers are working to integrate the two approaches (e.g., Raju, Laffitte, & Byrne, 2002; Reise, Widaman, & Pugh, 1993; Stark, Chernyshenko, & Drasgow, 2006; Widaman & Grimm, 2014). Here, we focus exclusively on the SEM framework using confirmatory factor analysis (CFA) because SEM is more commonly used than IRT. Readers interested in the IRT approach are referred to Tay, Meade, and Cao (2015) for an overview and tutorial, and Meade and Lautenschlager (2004) for a comparison of the SEM-CFA and IRT approaches.

In a CFA, items that make up a construct (e.g., questionnaire items that form a scale) load on a latent or unobserved factor representing the construct. Widaman and Reiss (1997) described four main steps for testing measurement invariance: configural, weak factorial (also known as metric), strong factorial (also known as scalar), and strict (also known as residual or invariant uniqueness). Vandenberg and Lance (2000) soon after outlined 8 steps for testing measurement invariance, with the first 5 steps consisting of the main tests of measurement invariance. The last 3 steps reflected structural invariance of the derived latent factors themselves (e.g., equivalence of factor variances, covariances, and means) and will not be considered in this report. Here, we review the four measurement invariance steps that Widaman and Reiss (1997) identified and which coincide with steps 2-5 in Vandenberg and Lance (2000). We exclude Vandenberg and Lance's (2000) first step, invariant covariance matrices, because rejection of this test is “uninformative with respect to the particular source of measurement inequivalence” (Vandenberg & Lance, 2000, p. 36), because contemporary guidelines now omit this first step (Milfont & Fischer, 2010; van de Schoot, Lugtig, & Hox, 2012), and because this test is rarely performed in practice. Hence, the four measurement

invariance steps considered are: (1) *configural*, equivalence of model form; (2) *metric* (weak factorial), equivalence of factor loadings; (3) *scalar* (strong factorial), equivalence of item intercepts or thresholds; and (4) *residual* (strict or invariant uniqueness), equivalence of items' residuals or unique variances.

To concretely illustrate each step we invoke an example comparing parental warmth and control in the United States and China. Figure 1 displays the multiple-group CFA model that will serve as the example. In Figure 1A, warmth and control are the latent variable constructs that we ultimately want to compare across the two cultures. Parental warmth and control are measured by a 10-item questionnaire with 5 continuously distributed items that load on a latent factor that represents warmth (love, praise, etc.) and 5 continuously distributed items that load on a latent factor that represents control (monitor, punish, etc.). The item loadings (weights) on each latent factor are indicated by parameters *l1-l10*. The item intercepts (means) are indicated by parameters *i1-i10*. Item residual variances (item-specific variance + error variance) are indicated by parameters *r1-r10*. In the multiple-group framework, this model is applied to parents in China and the United States separately, and, following an initial test, various levels of model constraints are applied (i.e., parameters are set to be equal) across the two cultures. In general through this example, we are focusing on estimation of measurement invariance with continuously distributed items, but we also note some variations in the procedure for items that are measured in a different scale (e.g., ordinal or dichotomous items).

## Configural invariance

The first, and least stringent, step in the measurement invariance ladder is configural invariance, or invariance of model form. This step is designed to test whether the constructs (in this case, latent factors of parental warmth and control) have the same pattern of free and fixed loadings (e.g., those that are estimated by the model and those that are fixed at 0) across groups (in this case the two cultures). Invariance at the configural level (how this is determined is discussed below) means that the basic organization of the constructs (i.e., 5 loadings on each latent factor) is supported in the two cultures (Figure 1A applies to both groups). Configural noninvariance (assume Figure 1A applies to one group and Figure 1B applies to the other) means that the pattern of loadings of items on the latent factors differs in the two cultures (e.g., in one culture only, at least one item loads on a different factor, cross-loads on both factors, etc).

Finding configural noninvariance leaves two options: (1) redefine the construct (e.g., omit some items and retest the model) or (2) assume that the construct is noninvariant and discontinue invariance and group difference testing. Redefining the construct in any step of invariance testing (e.g., by omitting items or making other post-hoc alterations to the model) is a “data-driven” strategy, rather than a “theory-driven” one, that is exploratory in nature, and results should be interpreted with this concern in mind. To validate exploratory model alterations, empirical replication in similar samples is particularly important.

## Metric invariance

If configural invariance is supported, the next step is to test for metric invariance, or equivalence of the item loadings on the factors. Metric invariance means that each item contributes to the latent construct to a similar degree across groups. Metric invariance is tested by constraining factor loadings (i.e., the loadings of the items on the constructs) to be equivalent in the two groups. In the example, the loadings of the 5 warmth questionnaire items (noted as *l1-l5* in Figure 1A) are set to be equivalent across U.S. and Chinese groups, and the loadings of the 5 control questionnaire items (noted as *l6-l10* in Figure 1A) are set to be equivalent across U.S. and Chinese groups (e.g., *l1* in China = *l1* in U.S.; *l2* in China = *l2* in U.S., etc.). The model with constrained factor loadings (Figure 1C) is then compared to the configural invariance model (Figure 1A) to determine fit. If the overall model fit is significantly worse in the metric invariance model compared to the configural invariance model (model fit is discussed below), it indicates that at least one loading is not equivalent across the groups, and metric invariance is not supported. For example, noninvariance of a loading related to kissing a child on the warmth factor would indicate that this item is more closely related to parental warmth in one culture than in the other (assume Figure 1C applies to one group and Figure 1D applies to the other). If the overall model fit is not significantly worse in the metric invariance model compared to the configural invariance model, it indicates that constraining the loadings across groups does not significantly affect the model fit, and metric invariance is supported.

Finding metric noninvariance leaves three options: (1) investigate the source of noninvariance by sequentially releasing (in a backward approach) or adding (in a forward approach; see e.g., Jung & Yoon, 2016) factor loading constraints and retesting the model until a partially invariant model is achieved (partial invariance is discussed below), (2) omit items with noninvariant loadings and retest the configural and metric invariance models, or (3) assume that the construct is noninvariant and discontinue invariance and group difference testing.

## Scalar invariance

If full or partial metric invariance is supported, the next step is to test for scalar invariance, or equivalence of item intercepts, for metric invariant items. Scalar invariance means that mean differences in the latent construct capture all mean differences in the shared variance of the items. Scalar invariance is tested by constraining the item intercepts to be equivalent in the two groups. The constraints applied in the metric invariance model are retained. In the example, assuming full metric invariance, the intercepts (means) of the 5 questionnaire items that load on parental warmth (noted as *i1-i5* in Figure 1A) are set to be equivalent across U.S. and Chinese groups, and the intercepts of the 5 questionnaire items that load on control (noted as *i6-i10* in Figure 1A) are set to be equivalent across U.S. and Chinese groups (e.g., *i1* in China = *i1* in U.S.; *i2* in China = *i2* in U.S., etc.). Any items that had unequal loadings in the metric invariance model (and were allowed to vary) should be allowed to vary in the scalar invariance model because it is meaningless to test for equal item intercepts if the metric of the items differs across groups. The model with constrained item intercepts (Figure 1E) is then compared to the metric invariance model (Figure 1C) to determine fit. If the overall model fit is significantly worse in the scalar invariance model compared to the metric

invariance model, it indicates that at least one item intercept differs across the two groups, and scalar invariance is not supported. For example, noninvariance of an item intercept for kissing a child would mean that parents in one culture kiss their children more, but that increased kissing is not related to increased levels of parental warmth in that culture (assume Figure 1E applies to one group and Figure 1F applies to the other). If the overall model fit is not significantly worse in the scalar invariance model compared to the metric invariance model, it indicates that constraining the item intercepts across groups does not significantly affect the model fit, and scalar invariance is supported. Scalar invariance is not always tested separately from metric invariance. For example, for models with items that are measured with two categories (binary) rather than on a continuous scale, metric and scalar invariance may be tested in a single step (Muthén & Asparouhov, 2002), and models intended to compare SEM-CFA to IRT methods may also report a combined metric-and-scalar invariance step because discrimination and location parameters are conventionally tested together in IRT (e.g., Stark et al., 2006).

Finding scalar noninvariance leaves three options: (1) investigate the source of noninvariance by sequentially releasing (in a backward approach) or adding (in a forward approach) item intercept constraints and retesting the model until a partially invariant model is achieved, (2) omit items with noninvariant intercepts and retest the configural, metric, and scalar invariance models, or (3) assume that the construct is noninvariant and discontinue invariance and group difference testing.

### Residual Invariance

If scalar invariance is supported, the final step for establishing measurement invariance is to test for residual invariance, or equivalence of item residuals of metric and scalar invariant items. Residual invariance means that the sum of specific variance (variance of the item that is not shared with the factor) and error variance (measurement error) is similar across groups. (It should be noted that there could be larger measurement error and less specific variance in one group than another, and residual invariance could still be supported if the totals of these two components were similar.) Although a required component for full factorial invariance (Meredith, 1993), testing for residual invariance is not a prerequisite for testing mean differences because the residuals are not part of the latent factor, so invariance of the item residuals is inconsequential to interpretation of latent mean differences (Vandenberg & Lance, 2000). On this account, many researchers omit this step. However, we include it here because residual invariance is still reported in many tests of measurement invariance. Residual invariance is tested by constraining the item residuals (noted as *r1*-*r10* in Figure 1A) to be equivalent in the two groups (e.g., *r1* in China = *r1* in U.S.; *r2* in China = *r2* in U.S., etc.). The constraints applied in the scalar invariance model are retained. Like in the scalar invariance model, any items with unequal loadings and/or intercepts should be allowed to vary across groups (i.e., not constrained) in the residual invariance model. The model with constrained item residuals (Figure 1G) is then compared to the scalar invariance model (Figure 1E) to determine fit. If the overall model fit is significantly worse in the residual invariance model compared to the scalar invariance model, it indicates that at least one item residual is different across the two groups, and residual invariance is not supported. If the overall model fit is not significantly worse in the residual invariance model compared

to the scalar invariance model, it indicates that constraining the residuals across groups does not significantly affect the model fit, and residual invariance is supported.

If residual noninvariance is found, researchers can (1) investigate the source of residual noninvariance by sequentially releasing (in a backward approach) or adding (in a forward approach) item residual constraints and retesting the model until a partially invariant model is achieved, or (2) accept the noninvariant residuals and proceed with tests of mean differences or differential relations in the latent factors across groups.

## Mean Differences in Latent Factors

Once the configural, metric, and scalar invariance steps have been passed, the researcher is free to compare group means on the latent factors (warmth and control in Figure 1E, for example). One common way to do this is to set the latent factor mean to 0 in one group and allow it to vary in the second group. The estimated mean parameter in the second group represents the difference in latent means across groups. For example, if the latent factor variance is set to 1.0 and the standardized mean of the parental control latent factor is estimated at 1.00, $p < .05$, in the United States, then control in the United States is one standard deviation higher than control in China. Regardless of the method used to compare means, an effect size like Cohen's (1988) *d* should be reported to allow comparisons across different studies.

## Parameterization and Model Identification

There are two main ways to parameterize (set the parameters of) the tests of metric and scalar invariance in a CFA model in a multiple-group test of measurement invariance. The first approach is to set the variance of the latent factor at 1 and the mean of the latent factor at 0 in both groups. However, if the factor variance and mean are not actually the same for both groups, the metric and scalar invariance tests may be incorrect because the factor loadings and intercepts for each group are on different scales. As there is no way to know whether the factor variance and mean are truly equivalent across groups, this is not an ideal way to parameterize the model.

The second and most commonly used approach to parameterize the metric and scalar invariance models is to fix the loading of one item on a factor (the referent or marker item) at 1 and the intercept of the same item to 0 in both groups to identify the model and set the scale of the latent factor (latent factors have no scale by default, so this method makes the factor follow the same scale as the referent item). However, the choice of referent item has implications for the interpretation of the model (Johnson, Meade, & DuVernet, 2009). If a noninvariant item is chosen as the referent item, all other items on the factor may appear metric and/or scalar noninvariant because the scales of the latent factor are different for the groups. Of course, the researcher does not know which items are invariant prior to conducting invariance tests. Consequently, Cheung and colleagues (Cheung & Lau, 2012; Cheung & Rensvold, 1999; Rensvold & Cheung, 2001) described two procedures, the factor-ratio test and the stepwise portioning procedure, for testing pairs of items to identify those that are noninvariant. These tests are lengthy and somewhat complex, and the evidence

is mixed regarding their ability to correctly identify invariant and noninvariant items (French & Finch, 2008; Jung & Yoon, 2016).

## Fit of Measurement Invariance Models

Measurement invariance is tested by evaluating how well the specified model (e.g., the model set up by the researcher) fits the observed data. Current practice emphasizes the importance of using multiple fit statistics to assess model fit (Kline, 2015). Configural invariance is tested by evaluating the overall fit of the model. Which fit statistics should be reported is a source of debate (discussed below), but most scholars recommend reporting the chi-square ($\chi^2$) and two to four alternative fit indices (AFIs), here defined as all fit statistics other than $\chi^2$: Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean-square Residual (SRMR), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), McDonald's (1989) Noncentrality Index (McNCI), etc. (Definitions and advantages and disadvantages of these various fit statistics are described in the Supplementary Information.) The fit of metric, scalar, and residual invariance models is typically evaluated by comparing the fit of two nested models that are identical except for a target set of restrictions in one. For example, the configural and metric invariance models are nested because they have the same model structure except that the metric invariance model imposes equality constraints on the factor loadings. Differences between the two models can therefore be attributed to the imposed constraints. Nested model comparisons involve computing the difference between fit statistics for the two models (e.g., $\Delta\chi^2$, $\Delta$CFI). Current conventions for evaluating model fit are described below.

## Measurement Invariance Conventions

There remain many methodological issues surrounding the execution of measurement invariance testing. Among them are (1) the number and order of tests required to establish measurement invariance, (2) the criteria used to evaluate model fit, (3) partial invariance, and (4) the sample and model characteristics that may moderate invariance. The literature on each of these topics is still evolving; here we summarize and evaluate contemporary conventions and recommendations for each.

### Number and Order of Tests

In their review of the literature from the 1980s and 1990s, Vandenberg and Lance (2000) noted that few studies tested all levels of invariance, and in particular testing scalar invariance was infrequent. However, the literature on measurement invariance converged considerably in the 2000s. To test measurement invariance, it is becoming more common to use a mean and covariance structure (MACS; Little, 1997; Ployhart & Oswald, 2004) framework, which includes item intercepts and factor means, rather than just item loadings and residual terms. An increase in MACS models should be accompanied by an increase in scalar invariance tests. Below, we document which measurement invariance steps are commonly tested, and the order in which they are tested, in a sampling of recent studies.

### Criteria Used to Evaluate Model Fit

Classically, measurement invariance was evaluated using a single criterion, significance of the change in $\chi^2$ for two nested models (Byrne et al., 1989; Marsh & Hocevar, 1985; Reise et al., 1993). However, some researchers have shifted from a focus on absolute fit in terms of $\chi^2$ to a focus on alternative fit indices because $\chi^2$ is overly sensitive to small, unimportant deviations from a "perfect" model in large samples (Chen, 2007; Cheung & Rensvold, 2002; French & Finch, 2006; Meade, Johnson, & Braddy, 2008). Cheung and Rensvold's (2002) criterion of a -.01 change in CFI for nested models is commonly used, but other researchers have suggested different criteria for CFI (Meade et al., 2008; Rutkowski & Svetina, 2014) or the use of other alternative fit indices (e.g., ΔRMSEA, ΔSRMR; Chen, 2007; Meade et al., 2008), and some question the use of alternative fit indices entirely because of their lack of precision (Barrett, 2007; Bentler, 2007; Fan & Sivo, 2009). For sample sizes with adequate power, equal group sizes, and mixed invariance (i.e., some loadings are higher and some lower in the first group), Chen (2007) also suggested a criterion of a -.01 change in CFI, paired with changes in RMSEA of .015 and SRMR of .030 (for metric invariance) or .015 (for scalar or residual invariance). Meade et al. (2008) suggested a more conservative cutoff of -.002 for the change in CFI as well as using a condition-specific cutoff (i.e., a cutoff value that depends on the number of items and factors in the model) for McDonald's (1989) noncentrality index (McNCI), but cautioned that neither criterion should be used for models with low statistical power. Rutkowski and Svetina (2014) investigated model fit in conditions comparing 10 or 20 groups. They concluded that changes in CFI of -.02 and RMSEA of .03 were most appropriate for tests of metric invariance with large group sizes, but the traditional criteria of -.01 for ΔCFI and .01 for ΔRMSEA were appropriate for scalar invariance tests. There is no consensus about the best fit indices or cutoff values for alternative fit indices under all conditions, leaving researchers to choose fit criteria. Below, we document the fit statistics used to evaluate measurement invariance in a sampling of recent studies.

### Partial Invariance

Because full measurement invariance in all four steps is often not supported, it is becoming common practice to accept some violations of measurement invariance (e.g., releasing constraints on one or more loadings or intercepts or both) and continue with tests of mean differences or relations among constructs using the partially invariant factor. However, standards for partial invariance vary. Byrne, Shavelson, and Muthén (1989) described testing partial measurement invariance, but placed no restrictions on the number of parameters released other than nebulously suggesting that it "makes substantive sense to do so" (p. 465). Steenkamp and Baumgartner (1998) suggested that ideally more than half of items on a factor should be invariant. Similarly, Vandenberg and Lance (2000) suggested that a factor can be considered partially invariant if the majority of items on the factor are invariant. However, no empirical studies are cited to support these guidelines. Other researchers explored the consequences of partial invariance for interpretation of mean differences. Chen (2008) demonstrated that, as the proportion of noninvariant items on a factor increased, so did the bias in mean estimates for subgroups (and therefore the estimated difference between subgroup means), but made no suggestion for an "acceptable" proportion of invariant items (see also Guenole & Brown, 2014, for bias of regression parameters under noninvariance).

In a Monte-Carlo simulation, Steinmetz (2013) demonstrated that metric noninvariance (unequal factor loadings) had a negligible effect on mean differences of a latent factor, but that scalar noninvariance (unequal intercepts) led to serious misinterpretation of true mean differences (see also Schmitt, Golubovich, & Leong, 2011). Given the burgeoning literature on partial invariance, below we document the level of invariance (no, partial, full) reported in each invariance step of a set of contemporary studies.

## Sample and Model Characteristics that May Moderate Measurement Invariance

There are several sample and model characteristics that may affect the level of measurement invariance achieved. Three characteristics that have demonstrated relations with the fit of measurement invariance models - sample size, the number of groups compared, and model size (e.g., the complexity of the model that is being tested) - are considered.

**Sample size—**The number of participants included in tests of measurement invariance is known to affect the power of the tests, and hence the test's sensitivity to detecting differences in absolute model fit. Because $\chi^2$ increases in power to reject the null hypothesis as the sample size increases, having a larger total sample may lead to over-rejection of measurement invariance tests if the change in $\chi^2$ is the only criterion used to evaluate fit. Change in alternative fit indices (AFIs) may be less sensitive to sample size (Cheung & Rensvold, 2002), but some evidence suggests that measures of absolute model fit (like the RMSEA) over-reject correct models in small samples ($N < 100$; Chen, Curran, Bollen, Kirby, & Paxton, 2008). As it becomes common practice to use AFIs as fit criteria, sample size (assuming adequate power) may be less important to the level of measurement invariance achieved because AFIs are less sensitive to sample size.

**Number of groups—**It is unclear whether the number of groups compared in tests of measurement invariance affects the ability to achieve full or partial invariance. For practical reasons, most studies of the power and sensitivity of measurement invariance tests compare two or three groups. We located only one study that investigated the performance of fit indices with a larger number of groups. Rutkowski and Svetina's (2014) simulation study of 10-20 groups suggested that, as the number of groups increased, the change in CFI decreased and the change in RMSEA increased. This result led the authors to recommend less stringent cutoff values for tests of metric invariance (but not for the other steps) with 10 or more groups.

**Model size—**There is some evidence that the performance of fit statistics varies by the size of the model (e.g., the number of observed variables and factors estimated; model degrees of freedom). The $\chi^2$ statistic is sensitive to model size, and Herzog, Boomsma, and Reinecke (2007) recommended using a Swain correction to the $\chi^2$ (which corrects for sample size and model size) when large models are tested. Fan and Sivo (2009) questioned the use of change in alternative fit indices (e.g., ΔCFI, ΔRMSEA) because the performance of these tests was highly related to model size (e.g., tests were more sensitive in small models and much less sensitive as model size grew) in their Monte Carlo simulation. They reasoned that applying a single AFI cut-off to models of various sizes was likely to lead to dubious conclusions about measurement invariance. Sensitivity to model size is also the reason Meade et al. (2008)

suggested using a condition-specific McNCI cutoff that depends on the number of items and factors in the model. Using simulated data, Kenny, Kaniskan, and McCoach (2015) also found that the RMSEA is overly sensitive in models with few (e.g., < 10) degrees of freedom.

Following, we determine how sample size, number of groups, and model size relate to achieved levels of measurement invariance. Using adjusted model fit criteria with large sample sizes, numbers of groups, and models may reduce the effects of these moderators on measurement invariance outcomes, but more targeted simulation research is needed to support differential cutoffs under various special conditions, such as small samples, large numbers of groups, and small and large models.

## Contemporary Uses of Measurement Invariance

Here, we document the levels of invariance (no, partial, full) achieved for each invariance step tested (configural, metric, scalar, residual) in a sampling of articles published over a 1-year period, and explore whether model fit criteria and sample size, number of groups being compared, and model size (as indexed by model degrees of freedom) relate to the invariance level (no, partial, full) achieved for each step. We expected that: (1) most researchers would, at a minimum, test configural, metric, and scalar invariance because those tests are required prior to group mean comparisons; (2) partial invariance tests would be relatively common; (3) use of the $\chi^2$ difference test as the only index of model fit would be associated with lower levels of measurement invariance; (4) use of alternative fit indices would be associated with higher levels of measurement invariance; and (5) larger sample size, number of groups compared, and model size would be associated with lower levels of measurement invariance.

## Method

To illustrate common contemporary practices in measurement invariance research, we surveyed general psychological articles indexed in APA's PsycNet database and published from May 2013 through April 2014. Search criteria included: Any Field: “measurement invariance” OR “metric invariance” OR “measurement equivalence” OR “metric equivalence” OR “multiple group model” OR “multisample model”.

A total of 157 articles was identified, but 31 were excluded because they were dissertations ($n$ = 7), theoretical, or included only simulated data ($n$ = 9), did not include a test of measurement invariance ($n$ = 7), used a non-SEM technique ($n$ = 7), or were not in English ($n$ = 1). Therefore, 126 articles which included 269 tests of at least one measurement invariance step were evaluated.

### Coding

**Measurement Invariance—**Four invariance steps were documented: configural, metric, scalar, and residual. Three levels of measurement invariance were coded for each step: *No* (0) -coded if the study was unable to achieve adequate model fit (with or without modifications). *Partial* (1) - coded if the study reported adequate model fit only after modifications (e.g., releasing parameters). *Full* (2) - coded if the study reported adequate

model fit with no model modifications. Because of varying guidelines in the literature for different sample sizes, procedures, and models, level of invariance achieved was judged based on the criteria stated in each study (e.g., nonsignificant $\Delta\chi^2$, ΔCFI < .01, absolute fit).

**Model Fit Criteria—**The model fit criteria used for each test of invariance were recorded. Fit criteria that were coded included change in chi-square ($\Delta\chi^2$); change in alternative fit indices (ΔAFI), including the Comparative Fit Index (ΔCFI), Root Mean Square Error of Approximation (ΔRMSEA), Standardized Root Mean Square Residual (ΔSRMR), Tucker-Lewis Index (ΔTLI), McDonald's Noncentrality Index (ΔMcNCI), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Expected Cross Validation Index (ECVI); and absolute model fit (i.e., no comparison of nested models).

**Model Characteristics—**Total sample size, number of groups compared, and model size (*df* of baseline model) were recorded for each test of measurement invariance.

## Results

Most tests compared 2 (75%) or 3 groups (12%), but up to 19 groups were compared. Median total sample size was 725 (range=152-43,093), and median *df* of the base model was 62 (range=0-3,138). The fit statistics used to evaluate measurement invariance were: only the $\Delta\chi^2$ for 16.7%, only alternative fit indices (ΔAFI; e.g., ΔCFI, ΔRMSEA) for 34.1%, both $\Delta\chi^2$ and ΔAFI for 45.9%, and only measures of absolute fit (no change in model fit) for 3.3%. Overall, the $\Delta\chi^2$ was reported for 62.6% of tests, the change in CFI was reported for 73.2% of tests, and the change in another alternative fit index (e.g., ΔRMSEA, ΔTLI, ΔSRMR) was reported for 56.1% of tests.

Table 1 displays the percentage of measurement invariance tests that established the three levels of measurement invariance in each of the four steps. All tests of measurement invariance surveyed included a test of configural invariance. The majority included tests of metric invariance (82%) and scalar invariance (86%), but only 41% included a test of residual invariance. Contrary to Vandenberg and Lance's (2000) observation that scalar invariance was rarely tested, more tests of measurement invariance included scalar than metric invariance steps. However, the order of tests was invariably from least to most restrictive. Full configural and metric invariance were established for most comparisons, but full scalar and residual invariance were established for 60% or fewer comparisons. Partial invariance was reported for 2-26% of individual invariance steps (configural, metric, scalar, residual), and overall 32% of measurement invariance tests reported partial invariance for one or more steps.

Relations of model fit criteria and sample and model characteristics with levels of invariance appear in Table 2. The criteria used to determine fit of nested models were associated with the level of invariance achieved. Using only $\Delta\chi^2$ as a measure of exact fit (and no other criteria) was associated with lower levels of scalar invariance. Using the ΔCFI (with or without other criteria) was associated with higher levels of metric, scalar, and residual invariance. Including another AFI was associated with slightly higher levels of scalar

invariance. Finally, including $\Delta\chi^2$ as well as ΔAFI was associated with slightly higher levels of metric invariance.

Larger sample size was associated with a higher level of residual invariance, but otherwise sample size, number of groups, and model degrees of freedom were unrelated to the level of invariance achieved. Implications of these findings (or lack thereof) are discussed below.

## Discussion

Measurement invariance is a key methodological feature of comparative and developmental psychological science that is gaining prominence. Assessing contemporary measurement invariance practices is critical to understanding what is expected of research reports, as well as identifying where more research is needed. We explored current practices for testing measurement invariance in a sampling of recent psychological studies. Issues included the number and order of measurement invariance tests, the model fit criteria reported, the frequency with which partial invariance was reported, and how sample and model characteristics moderated the level of measurement invariance achieved. Each of these practices is discussed in the context of our overall discussion of measurement invariance, followed by suggestions for best practices and future directions.

### Number and Order of Tests

As expected, most measurement invariance tests include configural, metric, and scalar invariance steps. The lower percentage for metric than scalar invariance likely arises because some analyses do not allow for or recommend separate tests of metric invariance (e.g., Stark et al., 2006). For example, separate metric invariance tests are often not computed when using the weighted least squares mean and variance adjusted estimator in M*plus* (WLSMV; for estimating models with categorical variables; Muthén & Muthén, 2010) and multiple indicators multiple causes models (MIMIC; an alternative method of testing measurement invariance; Willse & Goodman, 2008). As expected, because it is not necessary for latent mean comparisons, fewer than half of tests of measurement invariance included the residual invariance step. With the exception of occasionally skipping the metric invariance test, steps occurred in the order of least to most restrictive (configural, metric, scalar, residual).

### Model Fit

Our review of studies which reported measurement invariance showed that 80% used one or more ΔAFI as a criterion for model fit (either alone or in combination with the $\Delta\chi^2$). Including ΔAFIs (and ΔCFI in particular) as a criterion may prevent over-rejection of models that demonstrate practical fit in large samples. Our survey showed, consistent with expectations, that use of ΔCFI was associated with higher levels of metric, scalar, and residual invariance and use of only $\Delta\chi^2$ was associated with lower levels of scalar invariance. However, the criterion most often used, Cheung and Rensvold's (2002) ΔCFI < −.01, has been criticized as overly liberal (e.g., allowing meaningful differences in the measurement of the construct across groups to go unchecked). Meade et al. (2008) recommended a ΔCFI < .002 based on the results of their Monte Carlo simulation, but Little (2013) suggested that the simulation parameters used by Meade and colleagues were too

strict for real-world models and therefore Meade et al.'s proposed cut-off may be too conservative. More simulation studies are needed to determine the best fit criteria for determining measurement invariance as well as the practical effects of noninvariance in different steps on the variances, covariances, and mean differences of the latent factor across groups.

One reason different researchers suggest different model fit criteria is that the model parameters tested in each study differ. No one researcher can test all possible models (sample size, number of groups, number of factors, number of items, loading sizes, latent mean differences, degree of misspecification, etc.). Consequently, any given Monte Carlo simulation is testing limited conditions, and the results of a given simulation apply only to the conditions tested. Focusing on tests of a particular element (e.g., large numbers of groups), as some do (e.g., Rutkowski & Svetina, 2014), helps to pin-point the model fit statistics and cutoffs that perform best for particular model conditions.

## Partial Invariance

Despite ambiguity in the research literature about the effects of partial invariance and lingering lack of consensus about the best ways to test partial invariance, approximately one-third of our sample of measurement invariance studies reported partial invariance for one or more steps. Researchers seem to be adopting the practice of releasing constraints as a way of managing noninvariance across groups, but little is known about the statistical or conceptual implications of accepting partial invariance. Even the process for identifying noninvariant items is still debated. Yoon and Kim (2014) suggested that sequentially releasing constraints (backward method) based on the highest modification index has much smaller Type I (false positive) error rates than releasing all problematic constraints in a single pass. However, the sequential release method was shown to have serious limitations when the proportion of noninvariant items was high (Yoon & Milsap, 2007). Jung and Yoon (2016) compared backward, forward, and factor-ratio methods for identifying noninvariant parameters. Generally, the forward method (sequentially adding parameters) worked better than the backward method (sequentially releasing parameters), but both methods worked well when adjusted criteria were used (a 99% confidence interval for the forward method and a modification index of 6.635 for the backward method). The factor-ratio method had the highest error rates.

There is also very little research on the accuracy of mean-level tests for partially invariant models, and much more research is needed to identify the statistical and conceptual consequences of partial metric and scalar invariance. Steinmetz (2013) suggested that researchers may permit more violations of metric invariance than scalar invariance because the effects of metric noninvariance on the accuracy of mean-level analyses are minimal, whereas the effects of scalar noninvariance are large. For example, just one noninvariant intercept on a 4- or 6-item factor lead to a spurious significant mean difference in 13% of samples (well above the expected 5% error rate) and reduced the ability to detect a real mean difference by about half. Guenole and Brown (2014) explored the effects of ignoring nonivariance across groups (e.g., not modeling separate parameters across groups when they are warranted) on relations among constructs. Ignoring one noninvariant factor loading (out

of six; i.e., metric noninvariant) or intercept (i.e., scalar noninvariant) did not produce significant bias (> 10%) in parameter estimates, but ignoring one fully noninvariant item (i.e., metric and scalar noninvariant) produced significant bias. Therefore, they concluded that it is important to release noninvariant parameters to account for partial invariance of a construct. Chen (2008) similarly demonstrated bias > 10% in slopes and factor means when more than 25% of loadings were noninvariant and the invariance was uniform (e.g., always higher in one group than the other), but the degree of bias was worse when one group was small ($n = 60$) and as the percentage of noninvariant items increased.

To manage partial noninvariance, Chen (2008) suggested comparing the results of interest (e.g., mean differences across groups, regression coefficients between latent variables, etc.) using a partially invariant model (imposing constraints on invariant items only) to those using a fully invariant model (imposing constraints on invariant *and* noninvariant items). If the substantive conclusions using the two models are similar, the researcher could conclude that noninvariance had little impact on the results. If the models have different substantive conclusions, however, there is no clear path forward. The construct is still noninvariant, neither model is clearly superior or correct, and we still do not know what the results would be if the construct were truly invariant. Another option is to compare the partial invariance model to a reduced (noninvariant items removed) fully invariant model (Cheung & Rensvold, 1998) *if* the argument could be made (and empirically demonstrated – e.g., with a very high correlation between the reduced fully invariant and partially invariant factors) that the constructs with and without the removed items were interchangeable. In either case, if the results of the two models (full and partial invariance, or reduced full and partial invariance) were similar, the researcher would have the option to choose which model specification to report as the main analyses. If, however, the results differed, the researcher in the latter case (reduced full and partial invariance) would have more information about the source of noninvariance and how the noninvariant items affected the results, and would have the option of reporting the results of the reduced fully invariant model.

## Sample and Model Characteristics that Moderate Fit

We explored whether achieving different levels of measurement invariance was related to sample size, the number of groups being compared, and model size. Contrary to our expectations, the level of measurement invariance achieved was generally unrelated to these factors. The lack of relations between level of invariance and sample size, the number of groups being compared, and model size is particularly notable because, as detailed above and in Supplementary Table 1, many researchers have documented the susceptibility of the $\chi^2$ difference test and other ΔAFIs to sample size and model size. Hence, researchers may use model fit criteria that correct for these problems. For example, researchers may appropriately choose to use AFIs rather than $\chi^2$ for model comparisons in large samples, AFIs that adjust for model size when testing large models, or adjusted cut-off values for models with many groups.

The absence of significant relations of sample size with measurement invariance most likely results from the increased use of ΔAFIs instead of, or as a supplement to, $\Delta\chi^2$. Even in studies that used only $\Delta\chi^2$ as the criterion, sample size was not associated with invariance

levels achieved at each step. The lack of association between sample size and invariance in tests that used only $\Delta\chi^2$ may have come about because relaxing the requirement to have a nonsignificant $\Delta\chi^2$ (although controversial, see e.g. Barrett, 2007; Bentler, 2007) has allowed researchers to choose which criteria to report. Researchers may report the $\Delta\chi^2$ as their only criterion when it is nonsignificant and default to ΔAFIs when the $\Delta\chi^2$ is significant. Furthermore, with a median sample size of over 700, most tests of measurement invariance include large samples, precluding exclusive use of the $\Delta\chi^2$.

Most (89%) measurement invariance tests in our sample data set compared two or three groups. Comparing two vs. more than two groups was unrelated to levels of measurement invariance achieved. However, the small number of studies that compared larger numbers of groups impedes our interpretation of the relations with number of groups because the statistical difference between testing two and three groups is small. Perhaps tests of many groups rarely appear in the literature because fewer researchers study many groups, measurement invariance is not achieved (e.g., the file drawer problem; Rosenthal, 1979), or model complexity causes researchers to shy away from tests with many groups. More research on the mathematical and practical implications of testing few versus many groups is needed.

Surprisingly, model size (as indexed by the *df* of the baseline model) was unrelated to the level of measurement invariance in any step. Based on previous research (Fan & Sivo, 2009; Kenny et al., 2015; Meade et al., 2008), we anticipated that testing more complex models would result in more violations of measurement invariance. Researchers may be using adjusted fit criteria to account for large samples, many groups, and complex models.

## Reporting Requirements

In the contemporary literature, reporting of measurement invariance models is haphazard. About 20% of 269 tests of measurement invariance in our sample data set failed to report the model degrees of freedom, and many reports were vague about the criteria used to determine model fit. Many failed to report the sample sizes per group. We propose requiring that measurement invariance studies uniformly include the following minimal information: (1) sample size used in the models if different from the total *N* in the study, (2) how missing data were handled, (3) number of groups being compared and *n*s in each group, (4) specific model fit criteria for the configural model and nested model comparisons, and (5) a table detailing the models tested, *df*, fit statistics, which models were compared, model comparison statistics (including Δ*df* for nested models), and statistical decisions for each model comparison. Table 3 provides an example.

## Future Directions

Vandenberg and Lance's (2000) review of the applied literature across two decades included only 67 articles conducted on measurement invariance (3.35 articles per year). Schmitt and Kuljanin (2008) identified 75 articles about measurement invariance over a 8-year period from 2000-2007 (9.38 articles per year), and Johnson et al. (2009) reviewed a 3-year period from 2005-2007 and uncovered 153 articles on measurement invariance (51 articles per year). Our study identified 126 articles with over 250 tests of measurement invariance for a

period of just one year (2013-2014). To say that measurement invariance is being increasingly adopted by the psychological community is an understatement. Growth is exponential. With measurement invariance expectations and applications on the rise, there are still several major issues that require more research and theoretical attention: (1) conceptual and statistical sensitivity, (2) invariance across continuous variables, and (3) statistical techniques for analyzing noninvariant data.

**Conceptual and Statistical Sensitivity—**The most important future direction, to our minds, is grappling with the issue of sensitivity (Vandenberg, 2002), that is, the ability of measurement invariance tests to detect real, meaningful differences in constructs across groups or measurement occasions. Vandenberg (2002) questioned "what thresholds need to be reached for a true difference or shift [in conceptual frames of reference] to exist?" (p. 144). Conversely, Millsap (2005) asked "when are group differences in factor structure small enough to ignore?" (p. 157). Currently, we cannot quantify the impact of violations of measurement invariance, and a lot is at stake. If virtually all comparative analyses must first show measurement invariance, it is important to know what violations at different invariance steps mean for the construct and interpretations of results. Even small violations of invariance could preclude researchers from making meaningful group or temporal comparisons. Perhaps small violations are serious, but the question of effect size or the sensitivity of measurement invariance tests is still open (Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014).

To address this issue, Meuleman (2012) and Oberski (2014) suggested ways to compute the impact of a noninvariant parameter on the change in the estimated latent mean. Rather than relying on nested model comparisons or modification indices that are sensitive to sample size, Meuleman's (2012) and Oberski's (2014) procedures estimate the impact of noninvariance on mean differences in the construct – a more practical test. Similarly, Nye and Drasgow (2011) developed an effect size index, $d_{\mathrm{MACS}}$, to quantify the degree of measurement noninvariance of an item across groups, as well as estimates of the change in mean and variance of a factor as a result of noninvariance of an item. However, these indices assume a simple CFA model where all items are measured on the same scale and load on a single factor (e.g., no cross-loadings or secondary factors). Furthermore, at least one indicator must be fully invariant to produce accurate estimates. For models that lead to selection of individuals into a group, Milsap and Kwok (2004) developed a method for testing whether noninvariance has a significant impact on group membership. Returning to the example of measuring depression symptoms in women and men, noninvariance of depression items may have implications for the proportion of men and women who are identified as clinically depressed. Milsap and Kwok's procedure provides an estimate of the practical importance of noninvariance through the difference in proportions of women and men selected as depressed, for example, under fully invariant and partially invariant models. These innovative procedures may help to identify practical differences in a construct under different conditions, but research aimed at quantifying the impact of noninvariance in real-world models in still in its infancy.

Currently, the limited literature on sensitivity mostly focuses on the mathematical aspects of model fit criteria to evaluate between-group differences in model fit; namely, which fit

indices and criteria should be used to judge "practical" differences in model fit between groups (e.g., Chen, 2007; Cheung & Rensvold, 2002). Perhaps more important, however, is the conceptual question. What does it mean for the construct's validity if a model is found to be noninvariant or only partially invariant? How much deviation in the configural, metric, and scalar steps is too much for meaningful group or temporal comparisons to be made? Methodologists suggest that comparing means across groups or time from a noninvariant model is akin to comparing "apples to oranges" (or any manner of different objects), but what if we are really comparing red delicious apples to granny smith? Is that close enough? When is noninvariant is *too* noninvariant?

**Invariance across Continuous Variables—**In general, measurement invariance is discussed across unordered groups (e.g., men and women, measurement occasions). However, sometimes a researcher may be interested in invariance across an ordinal variable or a continuous variable that does not lend itself to easy or valid post-hoc grouping (e.g., socioeconomic status, child age). In cases like these, researchers have begun to study methods, such as moderated nonlinear factor analysis (Bauer & Hussong, 2009; Molenaar, Dolan, Wicherts, & van der Maas, 2010) or score-based tests of measurement invariance (Wang, Merkle, & Zeleis, 2014), that allow for estimation of invariance across a range of scores. More research is needed on the performance of these tests across different conditions, and methods are currently only implemented using maximum likelihood estimation, but these procedures could be applied to other estimation methods in future research (Wang et al., 2014).

**Statistical Techniques for Analyzing Noninvariant Data—**When a researcher finds that a construct is nonivariant or only partially invariant, it is unclear whether, or how, to proceed with analyses. Sometimes noninvariance is expected, especially across time. As children grow, they may acquire skills that reorganize their cognitions. For example, language ability moves from single-word production to a more complex expression of grammatical rules to reading and comprehending printed words across the first decade of the child's life. Some constructs shouldn't be invariant, but if they aren't then how should they be analyzed? Should the researcher simply abandon a noninvariant construct, or could an alternative analysis be used to manage noninvariance? Cheung and Rensvold (1998) describe a procedure for teasing apart the group difference attributable to item responses and unequal item-construct relations. This procedure can provide the researcher with more information about the source of noninvariance, but it is still unclear what to do with a noninvariant construct. Is it better to drop noninvariant items or to model noninvariance under a partially invariant model? What if 75% of the items are noninvaraint? What if dropping items changes the meaning of the intended construct? Could groups be analyzed separately and then aggregated as if across separate studies (e.g., Curran & Hussong, 2009)? More research is needed about statistical approaches that could be used to handle partially or fully noninvariant data.

## Practical Considerations for Demonstrating Measurement Invariance

Many researchers recognize the importance of demonstrating measurement invariance, but current limitations of the methods must also be acknowledged. A failure to demonstrate

invariance should not necessarily preclude all further analyses of group or developmental differences. Noninvariance can be informative and may lead researchers to important conclusions about how different groups interpret the same construct. Returning to the example of warmth and control in the United States and China, noninvariance of an item for kissing a child informs the form and function of warmth across groups. Perhaps warmth is not communicated physically, but verbally, or through provisions in the environment in one or the other culture. Perhaps physical affection is interpreted as warm in one culture but inappropriate or overly indulgent in the other. If the parental warmth construct is noninvariant across groups, the researcher could explore mean differences and intercorrelations of the individual items to inform development of a better measure of parental warmth in future research. Some constructs may simply be non-comparable across groups because they are experienced so differently (e.g., how new mothers vs. fathers experience the pain of childbirth). However, minor deviations from invariance could be stated as a limitation of the study, and group differences could be interpreted accordingly. The concern is that potentially important comparative research will never see the light of print if full invariance cannot be achieved. Without solid research on the real-life implications of noninvariance, we see rejecting all noninvariant models as premature. Instead, we encourage researchers to test invariance, report their results and interpret any deviations from invariance in the context of the construct, test group differences if it makes sense to do so, and report any limitations of the tests. We also encourage editors and reviewers to view measurement invariance tests as a dynamic and informative aspect of the functioning of a construct across groups, rather than as a gateway test, and to accept well-reasoned arguments about how and why small deviations from invariance may not be practically meaningful to the interpretation of group or developmental differences.

### Conclusions

The literature on measurement invariance is rapidly evolving, but more research is needed. If establishing measurement invariance is a prerequisite to all tests of developmental change, mean comparisons, or differential relations across groups, a large portion of comparative research hangs in the balance. Controlled experimental studies, Monte Carlo simulations, and other creative analytic techniques that elucidate the consequences of model noninvariance at different steps, are needed. However, like many nuanced statistical topics, clear and easy general guidelines may not be forthcoming. Through careful exploration and analysis of measurement invariance, researchers can learn more about their constructs and the groups they study, and make more informed decisions about whether their constructs are reasonably invariant and therefore comparable across groups and measurements.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This research was supported by the Intramural Research Program of the NIH, NICHD. Many thanks to Charlene Hendricks and Patrick S. Malone for feedback on early drafts.

## References


Barrett P. Structural equation modelling: Adjudging model fit. Personality and Individual Differences. 2007; 42:815–824. DOI: 10.1016/j.paid.2006.09.018

Bauer DJ, Hussong AM. Psychometric approaches for developing commensurate measures across independent studies: traditional and new models. Psychological Methods. 2009; 14:101–125. DOI: 10.1037/a0015583 [PubMed: 19485624]

Bentler PM. On tests and indices for evaluating structural models. Personality and Individual Differences. 2007; 42:825–829. DOI: 10.1016/j.paid.2006.09.024

Bornstein MH. Form and function: Implications for studies of culture and human development. Culture & Psychology. 1995; 1:123–137. DOI: 10.1177/1354067X9511009

Byrne BM, Shavelson RJ, Muthén B. Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychological Bulletin. 1989; 105:456–466. DOI: 10.1037/0033-2909.105.3.456

Chen F, Curran PJ, Bollen KA, Kirby J, Paxton P. An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. Sociological Methods & Research. 2008; 36:462–494. DOI: 10.1177/0049124108314720 [PubMed: 19756246]

Chen FF. Sensitivity of goodness of fit indexes to lack of measurement invariance. Structural Equation Modeling. 2007; 14:464–504. DOI: 10.1080/10705510701301834

Chen FF. What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. Journal of Personality and Social Psychology. 2008; 95:1005–1018. DOI: 10.1037/a0013193 [PubMed: 18954190]

Cheung GW, Lau RS. A direct comparison approach for testing measurement invariance. Organizational Research Methods. 2012; 15:167–198. DOI: 10.1177/1094428111421987

Cheung GW, Rensvold RB. Cross-cultural comparisons using non-invariant measurement items. Applied Behavioral Science Review. 1998; 6:93–110. DOI: 10.1016/S1068-8595(99)80006-3

Cheung GW, Rensvold RB. Testing factorial invariance across groups: A reconceptualization and proposed new method. Journal of Management. 1999; 25:1–27. DOI: 10.1177/014920639902500101

Cheung GW, Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling. 2002; 9:233–255. DOI: 10.1207/S15328007SEM0902_5

Cohen, J. Statistical power analysis for the behavioral sciences. 2nd. Hillsdale, NJ: Erlbaum; 1988.

Curran PJ, Hussong AM. Integrative data analysis: the simultaneous analysis of multiple data sets. Psychological Methods. 2009; 14:81–100. DOI: 10.1037/a0015914 [PubMed: 19485623]

Davidov E, Meuleman B, Cieciuch J, Schmidt P, Billiet J. Measurement equivalence in cross-national research. Annual Review of Sociology. 2014; 40:55–75. DOI: 10.1146/annurev-soc-071913-043137

Fan X, Sivo SA. Using ΔGoodness-of-Fit indexes in assessing measn structure invariance. Structural Equation Modeling. 2009; 16:54–69. DOI: 10.1080/10705510802561311

French BF, Finch WH. Confirmatory factor analytic procedures for the determination of measurement invariance. Structural Equation Modeling. 2006; 13:378–402. DOI: 10.1207/s15328007sem1303_3

French BF, Finch WH. Multigroup confirmatory factor analysis: Locating the invariant referent sets. Structural Equation Modeling. 2008; 15:96–113. DOI: 10.1080/10705510701758349

Glanville JL, Wildhagen T. The measurement of school engagement: Assessing dimensionality and measurement invariance across race and ethnicity. Educational and Psychological Measurement. 2007; 67:1019–1041. DOI: 10.1177/0013164406299126

Guenole N, Brown A. The consequences of ignoring measurement invariance for path coefficients in structural equation models. Frontiers in Psychology. 2014; 5:980.doi: 10.3389/fpsyg.2014.00980 [PubMed: 25278911]

Herzog W, Boomsma A, Reinecke S. The model-size effect on traditional and modified tests of covariance structures. Structural Equation Modeling. 2007; 14:361–390. DOI: 10.1080/10705510701301602

Hong S, Malik ML, Lee MK. Testing configural, metric, scalar, and latent mean invariance across genders in sociotropy and autonomy using a non-Western sample. Educational and Psychological Measurement. 2003; 63:636–654. DOI: 10.1177/0013164403251332

Johnson EC, Meade AW, DuVernet AM. The role of referent indicators in tests of measurement invariance. Structural Equation Modeling. 2009; 16:642–657. DOI: 10.1080/10705510903206014

Jung E, Yoon M. Comparisons of three empirical methods for partial factorial invariance: Forward, backward, and factor-ratio tests. Structural Equation Modeling: A Multidisciplinary Journal. 2016; doi: 10.1080/10705511.2015.1138092

Kenny DA, Kaniskan B, McCoach DB. The performance of RMSEA in models with small degrees of freedom. Sociological Methods and Research. 2015; 44:486–507. DOI: 10.1177/0049124114543236

Kline, RB. Principles and practice of structural equation modeling. 4th. New York: Guilford Press; 2015.

Little TD. Mean and covariance structure (MACS) analyses of cross-cultural data: Practical and theoretical issues. Multivariate Behavioral Research. 1997; 32:53–76. DOI: 10.1207/s15327906mbr3201_3 [PubMed: 26751106]

Little TD. On the comparability of constructs in cross-cultural research: A critique of Cheung and Rensvold. Journal of Cross-Cultural Psychology. 2000; 31:213–219. DOI: 10.1177/0022022100031002004

Little, TD. Longitudinal structural equation modeling. New York: Guilford Press; 2013.

Marsh HW, Hocevar D. Application of confirmatory factor analysis to the study of self-concept: First- and higher order factor models and their invariance across groups. Psychological Bulletin. 1985; 97:562–582. DOI: 10.1037/0033-2909.97.3.562

McDonald RP. An index of goodness-of-fit based on noncentrality. Journal of Classification. 1989; 6:97–103.

Meade AW, Johnson EC, Braddy PW. Power and sensitivity of alternative fit indices in tests of measurement invariance. Journal of Applied Psychology. 2008; 93:568–592. DOI: 10.1037/0021-9010.93.3.568 [PubMed: 18457487]

Meade AW, Lautenschlager GJ. A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. Organizational Research Methods. 2004; 7:361–388. DOI: 10.1177/1094428104268027

Meredith W. Notes on factorial invariance. Psychometrika. 1964; 29:177–185.

Meredith W. Measurement invariance, factor analysis and factorial invariance. Psychometrika. 1993; 58:525–543.

Meuleman, B. When are item intercept differences substantively relevant in measurement invariance testing?. In: Salzborn, S.; Davidov, E.; Reinecke, J., editors. Methods, theories, and empirical applications in the social sciences: Festschrift for Peter Schmidt. Wiesbaden, Germany: Springer; 2012. p. 97-104.

Milfont TL, Fischer R. Testing measurement invariance across groups: Applications in cross-cultural research. International Journal of Psychological Research. 2010; 3:111–121. Available at: http://mvint.usbmed.edu.co:8002/ojs/index.php/web/article/view/465.

Millsap, RE. Four unresolved problems in studies of factorial invariance. In: Maydeu-Olivares, A.; McArdle, JJ., editors. Contemporary psychometrics: A festschrift for Roderick P McDonald Multivariate applications book series. Mahwah, NJ: Erlbaum; 2005. p. 153-171.

Millsap, RE. Statistical approaches to measurement invariance. New York: Routledge; 2011.

Millsap RE, Kwok OM. Evaluating the impact of partial factorial invariance on selection in two populations. Psychological Methods. 2004; 9:93–115. DOI: 10.1037/1082-989X.9.1.93 [PubMed: 15053721]

Molenaar D, Dolan CV, Wicherts JM, van der Maas HL. Modeling differentiation of cognitive abilities within the higher-order factor model using moderated factor analysis. Intelligence. 2010; 38:611–624. DOI: 10.1016/j.intell.2010.09.002

Muthén, B.; Asparouhov, T. Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. 2002. Retrieved from http://www.statmodel.com/download/webnotes/CatMGLong.pdf

Muthén, LK.; Muthén, BO. Mplus User's Guide. Sixth. Los Angeles, CA: Muthén & Muthén; 2010.

Nolte S, Elsworth GR, Sinclair AJ, Osborne RH. Tests of measurement invariance failed to support the application of the “then-test”. Journal of Clinical Epidemiology. 2009; 62:1173–1180. DOI: 10.1016/j.jclinepi.2009.01.021 [PubMed: 19595570]

Nye CD, Drasgow F. Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. Journal of Applied Psychology. 2011; 96:966–980. DOI: 10.1037/a0022955 [PubMed: 21463015]

Oberski DL. Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. Political Analysis. 2014; 22:45–60. DOI: 10.1093/pan/mpt014

Ployhart RE, Oswald FL. Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. Organizational Research Methods. 2004; 7:27–65. DOI: 10.1177/1094428103259554

Raju NS, Laffitte LJ, Byrne BM. Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. Journal of Applied Psychology. 2002; 87:517–529. DOI: 10.1037/0021-9010.87.3.517 [PubMed: 12090609]

Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychological Bulletin. 1993; 114:552–566. DOI: 10.1037/0033-2909.114.3.552 [PubMed: 8272470]

Rensvold RB, Cheung GW. Testing measurement models for factorial invariance: A systematic approach. Educational and Psychological Measurement. 1998; 58:1017–1034. DOI: 10.1177/0013164498058006010

Rensvold, RB.; Cheung, GW. Testing for metric invariance using structural equation models: Solving the standardization problem. In: Schriesheim, CA.; Neider, LL., editors. Research in management: Vol 1 Equivalence of measurement. Greenwich, CT: Information Age; 2001. p. 21-50.

Rosenthal R. The file drawer problem and tolerance for null results. Psychological Bulletin. 1979; 86:638–641. DOI: 10.1037/0033-2909.86.3.638

Rutkowski L, Svetina D. Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. Educational and Psychological Measurement. 2014; 74:31–57. DOI: 10.1177/0013164413498257

Schmitt N, Golubovich J, Leong FT. Impact of measurement invariance on construct correlations, mean differences, and relations with external correlates: An illustrative example using Big Five and RIASEC measures. Assessment. 2010; 18:412–427. DOI: 10.1177/1073191110373223 [PubMed: 20622198]

Schmitt N, Kuljanin G. Measurement invariance: Review of practice and implications. Human Resource Management Review. 2008; 18:210–222. DOI: 10.1016/j.hrmr.2008.03.003

Senese PV, Bornstein MH, Haynes OM, Rossi G, Venuti PA. Cross-cultural comparison of mothers' beliefs about their parenting very young children. Infant Behavior and Development. 2012; 35:479–488. DOI: 10.1016/j.infbeh.2012.02.006 [PubMed: 22721746]

Stark S, Chernyshenko OS, Drasgow F. Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. Journal of Applied Psychology. 2006; 91:1292–1306. DOI: 10.1037/0021-9010.91.6.1292 [PubMed: 17100485]

Steenkamp JEM, Baumgartner H. Assessing measurement invariance in cross-national consumer research. Journal of Consumer Research. 1998; 25:78–90. Available from: http://www.jstor.org/stable/10.1086/209528.

Steinmetz H. Analyzing observed composite differences across multiple groups: Is partial measurement invariance enough? Methodology: European Journal of Research Methods for the Behavioral and Social Sciences. 2013; 9:1–12. DOI: 10.1027/1614-2241/a000049

Struening EL, Cohen J. Factorial invariance and other psychometric characteristics of five opinions about mental illness factors. Educational and Psychological Measurement. 1963; 23:289–298. DOI: 10.1177/001316446302300206

Tay L, Meade AW, Cao M. An overview and practical guide to IRT measurement equivalence analysis. Organizational Research Methods. 2015; 18:3–46. DOI: 10.1177/1094428114553062

van de Schoot R, Lugtig P, Hox J. A checklist for testing measurement invariance. European Journal of Developmental Psychology. 2012; 9:486–492. DOI: 10.1080/17405629.2012.686740

Vandenberg RJ. Toward a further understanding of and improvement in measurement invariance methods and procedures. Organizational Research Methods. 2002; 5:139–158. DOI: 10.1177/1094428102005002001

Vandenberg RJ, Lance CE. A review and systhesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational Research Methods. 2000; 2:4–69. DOI: 10.1177/109442810031002

Wang M, Summers JA, Little T, Turnbull A, Poston D, Mannan H. Perspectives of fathers and mothers of children in early intervention programmes in assessing family quality of life. Journal of Intellectual Disability Research. 2006; 50:977–988. DOI: 10.1111/j.1365-2788.2006.00932.x [PubMed: 17100958]

Wang T, Merkle EC, Zeileis A. Score-based tests of measurement invariance: use in practice. Frontiers in Psychology. 2014; 5doi: 10.3389/fpsyg.2014.00438

Widaman KF, Ferrer E, Conger RD. Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. Child Development Perspectives. 2010; 4:10–18. DOI: 10.1111/j.1750-8606.2009.00110.x [PubMed: 20369028]

Widaman, KF.; Grimm, KJ. Advanced psychometrics: Confirmatory factor analysis, item response theory, and the study of measurement invariance. In: Reis, HT.; Judd, CM., editors. Handbook of research methods in social and personality psychology. 2nd. New York: Cambridge University Press; 2014. p. 534-570.

Widaman, KF.; Reise, SP. Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In: Bryant, KJ.; Windle, ME.; West, SG., editors. The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research. Washington, DC: American Psychological Association; 1997. p. 281-324.

Willse JT, Goodman JT. Comparison of multiple-indicators, multiple-causes–and item response theory–based analyses of subgroup differences. Educational and Psychological Measurement. 2008; 68:587–602. DOI: 10.1177/0013164407312601

Yoon M, Kim ES. A comparison of sequential and nonsequential specification searches in testing factorial invariance. Behavior Research Methods. 2014; 46:1199–1206. DOI: 10.3758/s13428-013-0430-2 [PubMed: 24356995]

Yoon M, Millsap RE. Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. Structural Equation Modeling. 2007; 14:435–463. DOI: 10.1080/10705510701301677

## Highlights

- Measurement invariance assesses the psychometric equivalence of a construct across groups.
- Appropriate and proper comparison between groups depends first on ensuring equivalence of a construct across those groups though measurement invariance testing.
- Current practices for testing and reporting measurement invariance are reviewed in a sample of 126 articles with 269 tests of invariance.
- Implications for the practice of measurement invariance and areas of research need are discussed.

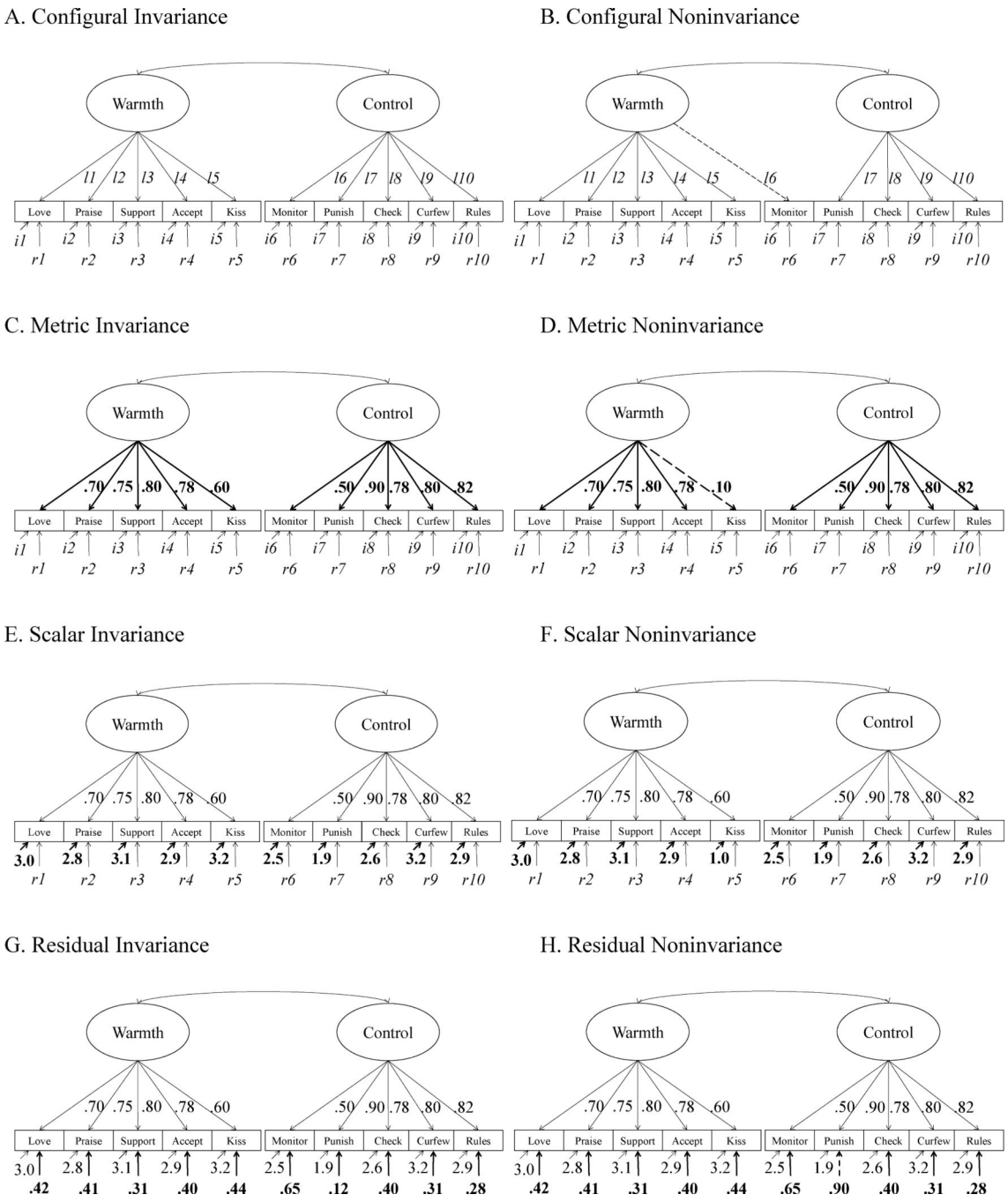


**Figure 1.**

A simulated confirmatory factor analysis model of parenting warmth and control.

*Note*. In measurement invariance tests, all models are fit to Chinese and United States groups and parameters are constrained to be equal across the groups. For the invariance models depicted in C, E, and G, the bolded parameters are the focal constraints, which are set to be equivalent in the two groups. For the noninvariance models depicted in B, D, F, and H, there is a path or constraint, represented by a dashed line, that applies only to one group

(compared to the base invariance model in A, C, E, and G, respectively, which applies to the other group).

## Table 1

**Percentages of measurement invariance tests in 126 research articles that established three levels of measurement invariance in each of the four steps**

| | Configural (*n* = 269; 100%) | Metric (Weak) (*n* = 220; 82%) | Scalar (Strong) (*n* = 232; 86%) | Residual (Strict) (*n* = 110; 41%) |
|---|---|---|---|---|
| No (0) | 3.3 | 6.4 | 15.5 | 33.6 |
| Partial (1) | 2.2 | 11.4 | 26.7 | 22.7 |
| Full (2) | 94.4 | 82.3 | 57.8 | 43.6 |

*Note*. A total of 269 measurement invariance tests was reported in 126 articles, but some articles did not include tests of all steps.

## Table 2

**Spearman correlations of model fit criteria and sample and model characteristics with level of invariance achieved in each step**

| | Configural (*df* = 215-267) | Metric (*df* = 181-218) | Scalar (*df* = 192-230) | Residual (*df* = 83-108) |
|---|---|---|---|---|
| *Fit criteria* | | | | |
| Only $\Delta\chi^2$ | .02 | -.03 | -.20** | -.08 |
| ΔCFI[a] | -.07 | .25*** | .14* | .24* |
| Other AFI[a] | .03 | .11 | .17* | .08 |
| $\Delta\chi^2$ and AFI | -.05 | .17* | .11 | .16 |
| *Sample and Model Characteristics* | | | | |
| Total sample size[b] | .10 | .07 | -.01 | .30*** |
| Group size (2 vs. >2) | -.04 | -.10 | -.10 | .01 |
| *df* of base model[b] | -.08 | -.02 | .01 | .11 |

*Note*. CFI = Comparative Fit Index. AFI = Alternative Fit Index - including Root Mean Square Error of Approximation (ΔRMSEA), Tucker-Lewis index (ΔTLI), McDonald's Noncentrality Index (ΔMcNCI), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Expected Cross Validation Index (ECVI).

*$p \leq .05$.

**$p \leq .01$.

***$p \leq .001$

[a]With or without other indices.

[b]Log transformed to approximate normality.

**Table 3**

**Sample table for reporting tests of measurement invariance**

| Model | $\chi^2$ (*df*) | CFI | RMSEA (90% CI) | SRMR | Model comp | $\Delta\chi^2$ ($\Delta df$) | ΔCFI | ΔRMSEA | ΔSRMR | Decision |
|---|---|---|---|---|---|---|---|---|---|---|
| M1: Configural Invariance | 1.20 (2) | .990 | .010 (.000-.090) | .03 | -- | -- | -- | -- | -- | -- |
| M2: Metric Invariance | 10.20$^{*}$(4) | .928 | .031 (.001-.110) | .08 | M1 | 9.00$^{*}$(2) | .062 | .021 | .05 | Reject |
| M2a: Partial Metric Invariance | 2.00 (3) | .988 | .020 (.000-.100) | .05 | M1 | .80 (1) | .002 | .010 | .02 | Accept |
| M3: Scalar Invariance | 15.00$^{**}$(6) | .946 | .042 (.026-.131) | .09 | M2a | 13.00$^{**}$(3) | .042 | .022 | .04 | Reject |
| M3a: Partial Scalar Invariance | 3.10 (5) | .979 | .030 (.000-.109) | .05 | M2a | 1.10 (2) | .009 | .010 | .00 | Accept |
| M4: Residual Invariance | 4.95 (9) | .975 | .032 (.002-.115) | .06 | M3a | 1.85 (4) | .001 | .002 | .01 | Accept |

Note. $N$ = 400; group 1 $n$ = 220; group 2 $n$ = 180.

$^{*}p \leq .05$.

$^{**}p \leq .01$.