Practical Assessment, Research, and Evaluation

Volume 12 *Volume 12, 2007* Article 3

2007

Decoding the Meaning of Factorial Invariance and Updating the Practice of Multi-group Confirmatory Factor Analysis: A Demonstration With TIMSS Data

Amery D. Wu

Zhen Li

Bruno D. Zumbo

Recommended Citation
Wu, Amery D.; Li, Zhen; and Zumbo, Bruno D. (2007) "Decoding the Meaning of Factorial Invariance and Updating the Practice of Multi-group Confirmatory Factor Analysis: A Demonstration With TIMSS Data," *Practical Assessment, Research, and Evaluation*: Vol. 12 , Article 3.
DOI: https://doi.org/10.7275/mhqa-cd89
Available at: https://scholarworks.umass.edu/pare/vol12/iss1/3

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 12, Number 3, February 2007 ISSN 1531-7714

Decoding the Meaning of Factorial Invariance and Updating the Practice of Multi-group Confirmatory Factor Analysis: A Demonstration With TIMSS Data

Amery D. Wu
Zhen Li
Bruno D. Zumbo
University of British Columbia.

Measurement invariance (MI) has been developed in a very technical language and manner that is generally not widely accessible to social and behavioral researchers and applied measurement specialists. Primarily relying on the widely known concepts in regression and linear statistical modeling, this paper decoded the concept of MI in the context of factor analysis. The paper began by describing what is MI (and lack of MI) and how the concept can be realized in the context of factor analysis. Next, we explained the need for modeling the mean and covariance structure (MACS), instead of the traditionally applied covariance structure, in detecting factorial invariance. Along the way, we addressed the related matter of statistically testing for MI using the Chi-squared likelihood ratio test and fit indices in multi-group MACS confirmatory factor analysis. Bringing to bear current developments by Cheung and Rensvold (2002) and others, we provided an update on the practice of using change in fit statistics to test for MI. Throughout the paper we concretized our discussion, without lack of generality to other constructs and research settings, with an example of 21 cross-country MI comparisons of the 1999 TIMSS mathematics scores.

The validity of cross-country (or cross-cultural) score comparisons is vital to many practices in applied psychological and educational research. The premise of validity in cross-country comparison is *construct comparability,* which necessitates that test[1] scores from different countries (or cultures) measure the same construct of interest on the same metric. Only then can score differences across countries be the true representation of the discrepancy in the performance/attribute, and the exercise of explaining variation by group membership be meaningful. In reality, however, difference in scores may be clouded with many confounding variables such as test adaptation (e.g., translation), curriculum differences, familiarity with item response formats, and many other socio-cultural factors. Unless evidence is demonstrated, construct comparability should never be naively assumed. Throughout this paper we will discuss cross-country comparisons but our recommendations apply to any groups of

[1] Throughout the paper, the terms "test" and "scale" are used interchangeably. If one is concerned with Psychological measures, for example, one may speak of scales, whereas in the Educational or certification settings one speaks of "tests".

respondents (e.g., groups based on gender, race, interventions, or the same respondents across time).

The paper is organized into three sections – reflecting the purposes of the paper. First, we decoded the meaning of measurement invariance (MI). Using the concept and framework of regression, we explained the meaning of *strict invariance* (Meredith, 1993), a necessary condition for construct comparability, and why a multi-group confirmatory factory analysis (MG-CFA) based on mean and covariance structure (MACS) is crucial for an investigation of strict MI. Second, we reviewed the controversy surrounding the use of the Chi-squared likelihood ratio test and fit indices as the decision rule for MI, which are traditionally used in MG-CFA, and provided updated criteria for making the statistical decision of MI. Third, we resolved the disagreement surrounding the necessity for testing for strict invariance by showing the impact that lack of strict invariance has on construct comparability. Along the way, we demonstrated the four complete steps for investigating strict MI using the TIMSS mathematics example. Throughout this paper, our notation remains as consistent as possible with those of Jöreskog & Sörbom (1999).

Before we move to our discussion on what MI actually means and its impact on construct comparability, it is crucial to point out the distinction between a latent variable and a construct. As Zumbo (2007) reminds us, although it is often confused even in the technical measurement literature, the construct is not the same as the true score or latent variable, which, in practical settings, is not the same as the observed item or task score. The essential difference being that a latent variable is a statistical and mathematical variable created by the data analyst and statistical modeler for which respondents (or examinees) could receive a predicted score based on their item responses. A construct, on the other hand, is an abstract or theoretical entity that has meaning because of its relation to other abstract variables, and a theory of the concept being studied. In short, one cannot get an empirically realized score on a construct, as they can on a latent variable. Test validity then involves an inference from the item responses to the construct via the latent variable; please see Zumbo (2007) for more details.

In settings such as cross-cultural comparison, an obvious and popular distortion of these concepts of construct and latent variable is nearly ubiquitous in the use of the term "construct comparability". In these studies what is, at best, often being demonstrated is the equivalence of latent variables. This remark is to inform readers that that even though we are following the literature and using the term "construct comparability", we want to acknowledge that construct comparability (as used in various domains of study to include, for example, cross-cultural differences, gender differences) is more than the equivalence of latent variables, or measurement invariance.

What Constitutes MI?

Mellenburgh (1989), Meredith (1993), and Meredith and Millsap (1992) provided a statistical definition of MI. Namely, an observed score is said to be measurement invariant if a person's probability of an observed score does not depend on his/her group membership, conditional on the true score. That is, respondents from different groups, but with the same true score, will have the same observed score.

More formally, given a person's true score, knowing a person's group membership does not alter the person's probability of getting a specific observed score. That is, the statistical definition of MI is:

> *Definition.* The observed random variable Y is said to be measurement invariant with respect to selection on G, if $F(y|\ \eta, g) = F(y|\ \eta)$ for all (y, η, g) in the sample space, where Y denotes an observed random variable with realization y; H denotes the latent variable (i.e., factor) with realization η that is measured by Y, or underlies Y; G denotes a random variable with realization g that functions as a selection of a subpopulation from the parent population by application of a selection function s(g), $0 <= s(g) <= 1$ (see Meredith, 1993, p. 528).

Therefore, MI holds if and only if the probability of an observed score, given the true score and the group membership, is equal to the probability of that given only the true score. To

this point, the definition of MI can apply to any observed variables at either the item or test level and hence is broad enough to provide the statistical basis for psychometric techniques such as differential item functioning (DIF) or item response theory methods, as well as factor analytic invariance.

This definition of MI, however, fits nicely into the framework of factor analysis wherein a factor score (i.e., the score on the latent variable) can be seen as the proxy for a person's true score, and the items are the observed random variables. Because a factor, in the context of factor analysis, can be construed as a type of latent variable, throughout this paper we will use the terms "latent variable" and "factor" interchangeably. The factor analysis framework allows one to empirically test for MI. To translate Meredith's (1993) notion of MI into factor analytic language, MI necessitates that the same latent variable is measured, and is measured on the same metric, so that cross-group factor scores are comparable. That is, factorial invariance requires that the measurement model linking the observed indicators to the unobserved factor(s) be identical across subgroups.

In research practice, cross-group factorial invariance is widely tested by multi-group confirmatory factor analysis (MG-CFA). It is important, at this point, to distinguish *covariance structure* (CS) modeling from *means and covariance structure* (MACS) modeling because MG-CFA can be applied to either CS or MACS data. The essential difference between MACS and CS is that MACS not only models covariances and variances but also the means of the observed variables – hence, in practice, resulting in intercepts being incorporated in the factor analytic model. Modeling factorial invariance based on MACS, instead of the more commonly used CS, is necessary for understanding and empirically testing Meredith's definition of MI. Throughout the remainder of this paper we use "MG-CFA" as a short-hand for "MG-CFA on MACS data".

To address the question "what constitutes MI?" a factor analysis model incorporating MACS is represented with a regression equation,

$$y_{ij} = \tau_j + \lambda_{j1}\eta_{1i} + \lambda_{j2}\eta_{2i} + \ldots \lambda_{jp}\eta_{pi} + r_{ij}, \qquad (1)$$

where y_{ij} denotes the i^{th} person's score (i= 1…N) on the j^{th} manifest variable (j= 1…J). Each response is assumed to be a linear combination of the intercept, τ_j, one or more factors, η_{pi} (p= 1…P), and a normally distributed random residual term, r_{ij} due to unpredictable fluctuation in the response process. The regression coefficients, λ_{jp} (i.e., slopes), are the loadings for item j on factor p, and the intercept, τ_j, is the y_{ij} score at which the factor(s) score is 0. The right-hand side of Equation (1) can be dissected into seven elements:

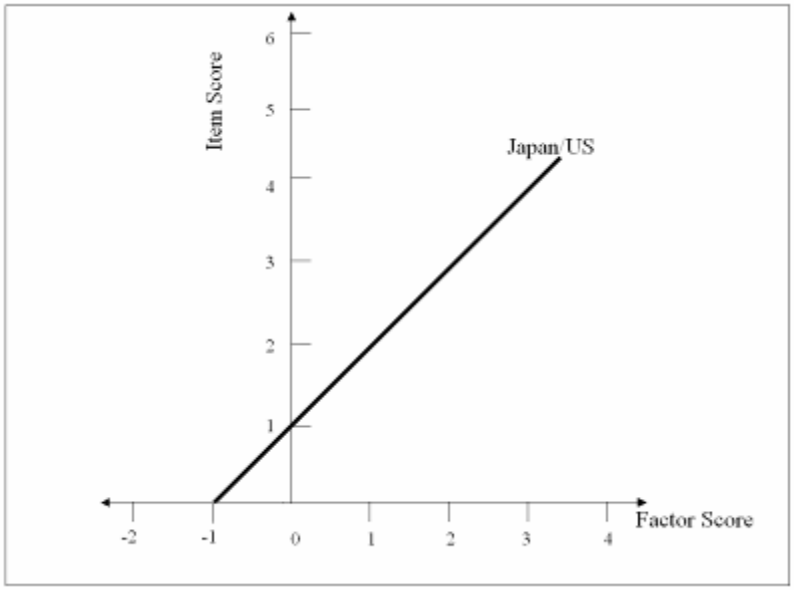
1. the model specification (number of factors and loading pattern),
2. the regression coefficient,
3. the regression intercept term,
4. the regression residual variance,
5. the means of the common factors,
6. the variances of the common factors, and
7. the covariances among the common factors.

The first four elements are related to the measurement model, which specifies how the observed indicators are related to the latent common factors. The last three elements are related to the structural model, which specifies the distribution of and the relationships among the latent variables. There is agreement in the research literature that cross-group equality in the last three structural elements is not a necessary condition for MI because equality in these elements is not involved in defining the relationship between the items and the factors (Little, 1999; Meredith & Millsap, 1992; Millsap, 1998; Widaman & Reise, 1997). In fact, explaining or predicting group differences in the mean of, the variance of, and the interrelationships among the true scores are often the fruit of much substantive research, and are widely analyzed statistically using a t-test, ANOVA, or ordinary least squares regression. Nonetheless, support for equality in the last three elements may suggest that the two groups may, in fact, belong to the same population regarding the construct of interest.

Unfortunately, the same agreement has not been reached regarding the necessity of the equality in the first four measurement elements, especially the regression residual variance (Cheung & Rensvold, 2002; Deshon, 2004; Lubke & Dolan, 2003; Little, 1999; Vandenberg & Lance, 2000). Equality in the first three elements ensures the

observed indicators have identical quantitative relationships with the latent variable(s) for each population of interest. Namely, the regression lines in (1) should be identical across groups for MI to hold. Figure 1 shows this ideal condition that MI holds. The two regression lines for, for example, U.S. and Japan, are identical in depicting the item-factor relationship. In order to achieve this ideal condition, it is necessary to examine whether the item-factor score scalings are equal across groups. The fourth element ensures that the MI established by the first three elements is not obfuscated by the non-random residuals so that the cross-group item-factor relationships remain identical when the effects of regression residuals are brought into the picture.

Figure 1: Item-Factor Regression Condition for MI



Four levels of nested hierarchy of factorial invariance have been formulated in the psychometrics literature, in correspondence to the increasing equality constraints on the four measurement elements in equation (1). The four levels of factorial tests are i) configural invariance, ii) weak invariance, iii) strong invariance, and iv) strict invariance (Meredith, 1993). Configural invariance requires that the same factor model specification holds across groups. In addition to configural invariance's equality constraints, weak invariance requires the cross-group equality in the loadings, strong invariance requires the cross-group equality in the loadings and intercepts, and strict invariance requires the cross-group equality in the loadings, intercepts, and residual variances. Meredith argued that strict invariance is a necessary condition for a fair and equitable comparison. However, in the 1990s to date, the governing belief reflected in research practice is that weak invariance, or strong invariance at best, would constitute sufficient evidence for MI (Little, 1997; Marsh, 1994; McArdle, 1998; Vanderberg & Lance, 2000). Not until recently, in support of Meredith's long-neglected argument, Deshon (2004) and Lubke and Dolan (2003) revisited the legitimacy of strict invariance for MI and affirmed the necessity for testing equality in the residual variances, in addition to loadings and intercepts.

However, this periodic advocacy for strict invariance has been largely neglected in applied measurement practice. A thorough review of empirical tests of MI in applied psychology by Vandenberg and Lance (2000) revealed that although 99% of the studies that they had reviewed investigated loading invariance, only 12% investigated intercept equality and 49% investigated residual variance equality. Our position is that cross-group equality in **all** four measurement-elements is a necessary condition for MI.

We believe that the inconsistency between the call for testing strict invariance and the day-to-day MI practices may be partially due to the lack of awareness of the (a) meaning of intercept and residual variance inequality and (b) impact of such inequality on MI. For this reason, the notion of factorial invariance, the necessity for strict invariance as well as the impact of lack of strict invariance will be described and demonstrated in more detail in the third section in the TIMSS example. Before we can proceed, however, to our TIMSS demonstration, we need to address the matter of the criterion for empirically testing MI using MG-CFA – i.e., what fit statistics and cut-offs should be used in applying MG-CFA?

Current Thinking About MG-CFA Decision Rules

Statistically, MG-CFA (Jöreskog & Sörbom, 1999) has become the most widely used method for investigating factorial invariance because of its reliance on formal hypothesis testing using the likelihood ratio test to support a decision about MI. (Byrne, Shavelson, & Muthén, 1989; Jöreskog & Sörbom, 1999; Zumbo, Sireci, & Hambleton, 2003). MG-CFA involves a sequence of hypothesis tests of nested models beginning with the least constrained model, often the configural model (Horn & McArdle, 1992), and then progressively placing equality constraints on the parameters across groups[2]. Hence, the subsequent test of MI is an augmentation of the parameter constraints of each proceeding hypothesis test. More demanding tests of MI will proceed only if the less demanding level of invariance is demonstrated.

Conventionally, the two-point decision about whether MG-CFA supports or rejects MI has solely relied upon the test of Chi-squared difference between two nested models ($\Delta\chi^2$), which itself also follows a Chi-squared distribution with degree of freedom equal to the difference between those of the two nested models. The decision rule for whether MI holds relies upon whether the added constraints make a significant improvement to the model fit. A non-significant improvement in fit is considered as evidence for MI (see comprehensive reviews in Cheung & Rensvold, 2002; Vandenberg & Lance, 2000) [3]. However, the practical usefulness of the $\Delta\chi^2$ test in MG-CFA has been questioned (Brannick, 1995; Kelloway, 1995). The concern with the $\Delta\chi^2$ test as a decision rule for MI can be understood first by focusing on single group CFA. In a single group CFA, the χ^2 fit statistic quantifies the magnitude of discrepancy between the sample and the fitted covariance matrices, and is calculated as $\chi^2 = (N-1)*(\text{Minimum Fitting Function})$.

The χ^2 is, therefore, clearly a function of the sample size, N. For this reason, the χ^2 test is susceptible to sample size in the sense that it rejects the null hypothesis with too much power if the sample size is large. In other words, the χ^2 test may reject trivial model-data differences and tends therefore to lose practical usefulness when used as the sole decision rule.

A variety of fit indexes were developed to accommodate the problems with sample size and model complexity, such as the Comparative Fit Index (CFI, Bentler, 1990) and RMSEA (Steiger, 1989). However, the sampling distributions of many of these fit indexes are unknown, hence, the formal hypothesis testing of the fit statistics cannot be conducted. For this reason, numerous cut-off criteria such as CFI $\geq$ 0.90 or 0.95 and RMSEA $\leq$ 0.06 or 0.08 were proposed to assist in determining model fit (see Fan & Sivo, 2005; Hu & Bentler, 1998; 1999; Marsh, Hau, & Wen, 2004; Schermelleh-Engel, Moonsbrugger, & Müller, 2003; Vandenberg & Lance, 2000). In an important sense, these fit statistics became the descriptive effect size indices for the χ^2 test of fit. However, the move toward descriptive fit indices for indication of model fit did not turn out to be a panacea, either. Despite the fit indices' efforts in adjusting model complexity (i.e., number of items and number of factors and the ratio of the two), most fit indices were still shown to be sensitive to model complexity

[2] Some authors have suggested a hypothesis testing strategy that involves beginning with the most constrained model and hence testing invariance involves relaxing equality constraints.

[3] Strictly speaking, a non-significant improvement in fit indicates that the null hypothesis is retained. Like any hypothesis testing, it does not prove that the parameters constrained to equality are, in fact, equal in the population. At best, a non-significant improvement in fit only provides a weak form of evidence that cross-group parameters are likely to be equal in the population.

(Cheung & Rensvold, 2002; Marsh et. al., 2004) in such a manner that a relatively less stringent cut-off is appropriate for a more complex model, and a relatively more stringent cut-off is appropriate for a simpler model. In other words, the idea that the same cut-off value applies to all models is inappropriate (Cheung & Rensvold, 2002; Marsh et. al., 2004).

Despite the warning that χ^2-related and the unequivocal cut-offs for fit indices were inappropriate for indicating model fit, they are still being used alone or in a combined manner as a decision rule for MI for multi-group nested models. Cheung and Rensvold (2002) reminded researchers that, like the χ^2 test, $\Delta\chi^2$ test is also susceptible to sample size and/or model complexity and has less value in making practical decisions about MI. Their argument can be illustrated by the formula for obtaining the $\Delta\chi^2$,

$\Delta\chi^2$ = (N-1)[Min Fitting Function $_{(aug)}$ - Min Fitting Function $_{(com)}$]

where "aug" denoted the augmented model such as configural invariance and "com" denoted the more compact model such as weak, strong or strict MI models.

One can see that $\Delta\chi^2$ is also a function of sample size, and should not be used solely for practical decisions on MI (Cheung & Rensvold, 2002; Brannick, 1995; Kelloway, 1995; Vandenberg & Lance, 2000). In a Monte Carlo study, Cheung and Rensvold also showed that most of the fit indices were susceptible to model complexity for the MG-CFA nested models and should not be trusted as the sole criterion in making decisions about MI.

For reasons stated above, early literature advocated the use of *change in fit indices* such as ΔTLI $\leq$ 0.05 (Tucker & Lewis, 1973) and ΔRho $\leq$ 0.022 (McGaw & Jöreskog, 1971) as descriptive indices for nested models such as those in MG-CFA. However, these early recommendations have not been widely applied. Recently, however, Cheung and Rensvold (2002) recommended the revival of change in fit indices. In a Monte Carlo study of 20 different fit indices, Cheung and Rensvold showed that, as expected, $\Delta\chi^2$ was sensitive to sample size and model complexity. They also showed that despite many fit indices' efforts in adjusting for model complexity, only RMSEA is not affected by model complexity and RMSEA$\leq$ 0.05 was recommended for indicating the configural model fit. They also examined the appropriate cut-offs for change in fit indices to determine MI for nested models and suggested that ΔCFI (Hu & Bentler, 1990) $\leq$ -0.01, ΔGamma Hat (Steiger, 1989) $\leq$ -0.001, or ΔMcdonald's (1989) Non-Centrality Index $\leq$ -.02 were the best indication of support of MI. Although more research like Cheung and Rensvold's is needed to validate their findings in other settings, their suggestions have been the most justifiable theoretically or empirically to date. Hence, we adopted their decision rules for our TIMSS construct comparability example.

Why Should We Concern Ourselves With Strict Invariance?

The following section discusses the meaning of strict invariance and the impact that lack of strict invariance has on construct comparability, with an eye toward making a case for why we should concern ourselves with testing strict invariance. The discussion is set in the context of an example with real data and the demonstration of the four complete steps for investigating whether strict factorial invariance holds. It should be noted that although our demonstration is with international comparative educational achievement data, our conclusions about invariance testing apply more broadly to other constructs and other grouping variables (e.g., over time, or across genders).

The example data were retrieved from the first booklets of TIMSS 1999 grade 8 mathematics tests. The factor model is based on the 1999 TIMSS test blueprint, where the construct, mathematics proficiency, is measured by five content domains (indicators), each consisting of a pre-specified observed set of items (i.e., item parcel or item bundle). The use of item parcelling was justified for three reasons. First, the parcelling of the items was based on the item-content domain specification developed by TIMSS 1999. Second, items in each domain were tested for unidimensionality, which is the empirical prerequisite for item parcelling (Bandalos, 2002). Finally, the focus of this MI study was on the construct level, which, according to the TIMSS test blueprint, could be better represented

theoretically by the content domains rather than the individual items. The readers should note that by using homogeneous parcels (i.e., clusters or bundles, testlets) the techniques are applicable to computer adaptive tests. The score range for the five domains are: 0-16 for Fraction and Number Sense, 0-9 for Measurement, 0-5 for Geometry, 0-5 for Algebra, and 0-5 for Data Representation. Because the data are continuous, MG-CFA analyses were conducted using maximum likelihood estimation in LISREL/SIMPLIS on Pearson correlation matrices. Other popular statistical software packages such as SPSS cannot perform such analysis because they do not allow specification of free and fixed loadings in the same manner as confirmatory factor analysis software like LISREL, Mplus, or EQS. In addition, neither SPSS nor SAS allow one to conduct simultaneous multi-group factor analyses.

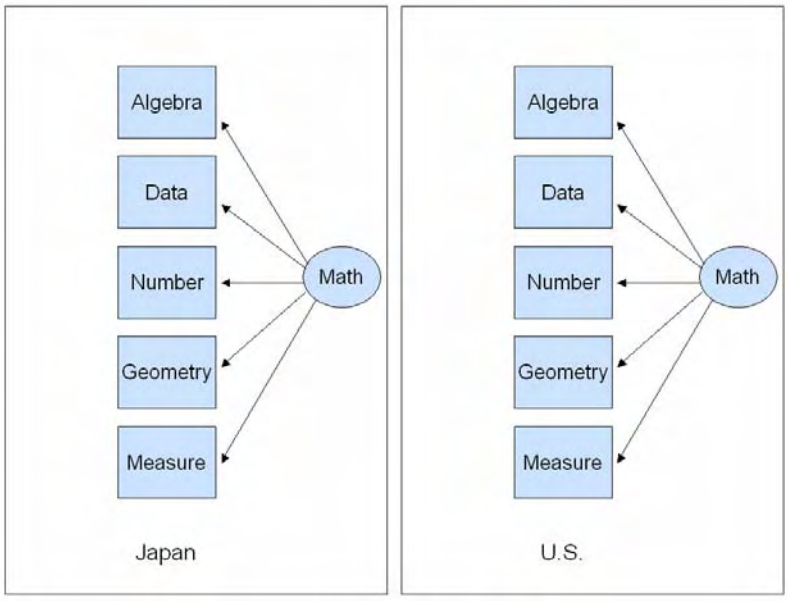
Data from seven countries, Australia (AUS), New Zealand (NZL), USA, Canada (CAN), Korea (KOR), Japan (JPN), and Taiwan (TWN) were examined. All the possible pairs of comparisons among the seven countries were investigated. By choosing these multiple countries, we intend to investigate the prevalence with which strict MI holds. Also, we intend to examine how sensitive the existing MG-CFA decision rules are to detect the possible MI distinction due to cultural similarities and discrepancies. For example, NZL, AUS, CAN, and US were considered countries that shared similar cultural paradigms, as were JPN, KOR, and TWN. In contrast, NZL and JPN, for instance, were considered countries that shared different cultural paradigms. For convenience, we termed paired comparisons between similar cultures as "within-culture" comparison, and paired comparisons between different cultures as "cross-culture" comparisons. This broad-stroke terminology does not imply that there are no cultural differences among the "within-culture" countries, or there are no similarities among the "cross-culture" countries. In total, the seven countries constituted 21 planned-comparisons where nine were within-culture and 12 were cross-culture comparisons. Note that, for comparative purpose, five fit indices reported in LISREL were listed, when applicable, in our following MI investigation: χ^2, $\Delta\chi^2$, RMSEA, and CFI, and ΔCFI.

Test of Configural Invariance

Configural invariance investigates whether examinees from different groups employ the same conceptual framework to answer the test items (Cheung & Rensvold, 2002; Horn & McArdle, 1992; Vandenberg & Lance, 2000). In MG-CFA, constraining the number of factor(s) and the pattern of the free and fixed loadings to be the same across groups tests configural invariance. Failure to demonstrate configural invariance indicates that different constructs were measured across groups. Hence, evidence of configural invariance is a prerequisite for MI and further testing is not appropriate if configural invariance does not hold (Cheung & Rensvold, 2002; Horn & McArdle, 1992; Little, 1997; Vandenberg & Lance, 2000). Figure 2 shows the configural model for our TIMSS example. To test configural invariance, this one-factor five-indicator model is constrained to be the same for two countries.

Following Cheung and Rensvold's (2002) recommendation, RMSEA$\leq$ 0.05 was used to evaluate configural model fit. Because model complexity was not a concern (i.e., a simple one-factor five indicator model), CFI was also reported to compliment RMSEA. To distinguish the possible MI difference between "cross-"and "within-"culture comparisons, Table 1 and subsequent tables were organized in a manner that the top nine comparisons were within-culture comparisons and were separated with a line from the bottom part, the 12 cross-culture comparisons. Sample sizes for the 21 comparisons were also reported in Table 1. Appendix A provides the LISREL/SIMPLIS syntax for testing configural invariance with the MACS model.

Figure 2: One-factor Five-indicator Configural Invariance Model for TIMSS



The results showed that all 21 configural models fit the data well; RMSEA ranged from 0.00 (e.g., AUS vs. CAN) to 0.05 (NZL vs. KOR). This good model fit was also supported by the CFI values equal to one for all comparisons. Hence, all 21 comparisons were eligible for further tests of stricter MI. Note that despite indication of good model fit by RMSEA and CFI, χ^2 rejected nine of the configural models including four out of nine within-culture comparisons. If the decision rule had been based on χ^2, it would have suggested termination for further examinations for these comparisons. Rejections of configural invariance by χ^2 were highlighted in bold.

Test of Weak Invariance

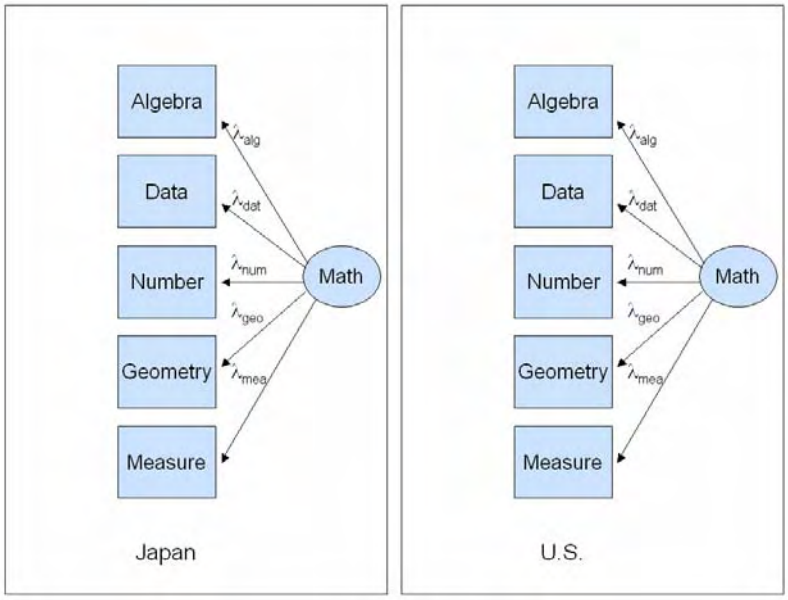
Weak invariance postulates that, for all items, one unit change in the item score is scaled to an equal unit change in the factor score across groups. Often, a substantive researcher's interest is to compare or explain the variation of a construct due to group membership. For such cross-group study to be meaningful, the scale (unit of measurement) of the latent variable should be identical across groups so that the variances derived are on the same metric regardless of group membership. Variance obtained from different units of measurement is not explainable or comparable. In addition to configural constraints, investigating whether the factor loadings are identical across groups tests the equality in item-factor score scaling (see Figure 3). Lack of weak invariance is problematic because the test items are calibrated to the factor scores with different units of measurement across groups. If one unit change in the item score does not result in equal unit change in the factor score across groups, the regression lines are not identical because the slopes are unequal; hence the regression lines are not identical for the groups.

Table 1: Fit Indices for Configural Model

Comparison	*N*	χ^2	p	RMSEA	CFI
AUS vs. NZL	945	11.18	0.34	0.02	1.00
CAN vs. USA	1887	23.17	**0.01**	0.04	1.00
AUS vs. CAN	1278	9.65	0.47	0.00	1.00
AUS vs. USA	1599	16.85	0.08	0.03	1.00
USA vs. NZL	1554	24.70	**0.01**	0.04	1.00
CAN vs. NZL	1233	17.49	0.06	0.04	1.00
JPN vs. KOR	1342	19.78	**0.03**	0.04	1.00
JPN vs. TWN	1309	7.19	0.71	0.00	1.00
TWN vs. KOR	1483	22.50	**0.01**	0.04	1.00
AUS vs. JPN	1079	3.91	0.95	0.00	1.00
AUS vs. KOR	1253	19.21	**0.04**	0.04	1.00
AUS vs. TWN	1220	6.62	0.76	0.00	1.00
USA vs. TWN	1829	20.14	**0.03**	0.03	1.00
USA vs. KOR	1862	32.73	**0.00**	0.05	1.00
USA vs. JPN	1688	17.42	0.07	0.03	1.00
CAN vs. JPN	1367	10.22	0.42	0.01	1.00
CAN vs. KOR	1541	25.52	**0.00**	0.04	1.00
CAN vs. TWN	1508	12.93	0.23	0.02	1.00
TWN vs. NZL	1175	14.46	0.15	0.03	1.00
NZL vs. JPN	1034	11.75	0.30	0.02	1.00
NZL vs. KOR	1208	27.05	**0.00**	0.05	1.00

Note. χ^2 rejection of configural invariance, $p < .05$, were highlighted in bold.

Figure 3: One-factor Five-indicator Weak Invariance Model for TIMSS



Note. λ_{alg}, λ_{dat}, λ_{num}, λ_{geo}, and λ_{mea} represent the factor loadings for Algebra, Data, Number, Geometry, and Measurement.

To help us illustrate this concept, imagine a mathematics test is administered to the US and Japanese students. For simplicity, let us focus on the equation for one-factor and one-item. Equation (1) then becomes,

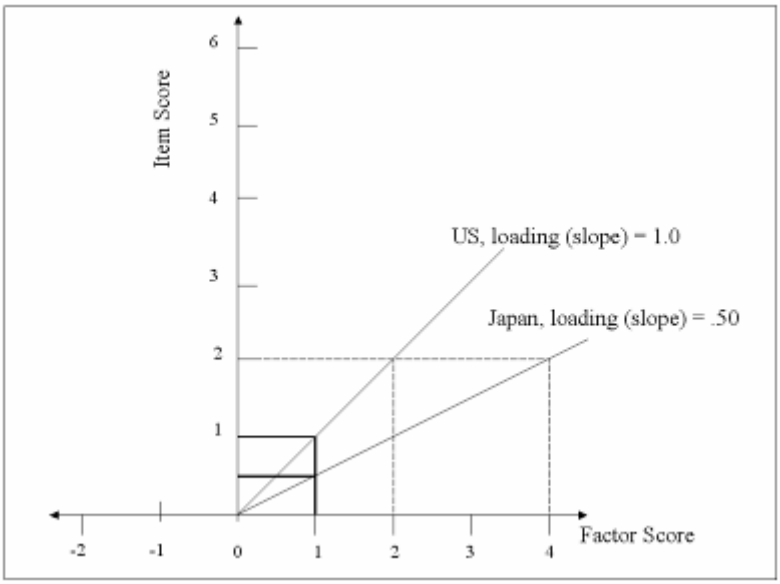
$$y_{ij} = \tau_j + \lambda_j \eta_i + r_{ij} \qquad (2)$$

Rearrange (2) to solve for η_i, assuming $r_{ij} = 0$, namely, no measurement error (we will return to the possible effect of measurement residual in the section on the testing of strict invariance), we get

$$\eta_i = \frac{y_{ij} - \tau_j}{\lambda_j}, \qquad (3)$$

That is, equation (3) shows that an examinee's factor score is equal to the ratio of the difference between the examinee's item score and the item intercept to the item loading. Figure 4 demonstrates the scenario where weak invariance is violated. The example item is hypothesized to be measured on a 0 to 6 scale and the factor is hypothesized to have a location of 0 and scale of 1 (note that the metric of a latent variable is arbitrary). Unequal slopes between the U.S. and Japan are hypothesized to be 1.0 and 0.5 respectively. Illustrated by the bold lines, one can see that one unit of item score is scaled to be one unit of the factor score for the US, whereas for Japan, only 0.5 units of the item score is scaled to be one unit of the factor score. As a result, shown by the dashed lines, for the US students, a score of 2 (with an intercept of 0) is calibrated to have a factor score of 2, however, the same item score is calibrated to have a factor score of 4 for the Japanese students. This unequal factor score calibration with regard to factor loading can be verified by equation (3). Using equation (3), an item score of 2 would be calibrated to a factor score equal to (2-0)/1= 2 for the US students but (2-0)/0.5 = 4 for the Japanese students. Such unequal calibration is, hence, biased against the US students. In this sense, cross-group inequality of factor loadings can be understood as the difference in factor score calibration with regard to the unit of measurement.

Figure 4: Impact of Lack of Loading Equality on MI



Appendix B provides the LISREL/SIMPLIS syntax for testing weak invariance with a MACS model. Note that as discussed earlier, the ΔCFI ≤ -0.01 rule was used to make final decisions about whether weak, strong, and strict MI models hold. To calculate ΔCFI, the CFI of the MI model being tested was compared to that of a one-level less constrained model. For example, the ΔCFI for the

weak MI model was calculated by subtracting the CFI of the configural invariance model from that of the weak invariance model. The results for the 21 weak MI tests were shown in Table 2. Rejections for weak invariance were highlighted in bold. Results showed that all nine within-culture comparisons passed the weak invariance test, and only two (out of nine) cross-culture comparisons were rejected. Note that if χ^2 had been the decision rule for configural invariance instead of RMSEA, many of the comparisons shown to be weak invariant would not have been detected because the investigation on these configure-rejected comparisons had not been allowed to proceed to the weak invariance phase. Similarly, if $\Delta\chi^2$ had been employed for decision-making for weak invariance, it would have rejected almost all the comparisons, 19 of the 21 comparisons including seven out of nine of the within-culture comparisons (Table 2). On the contrary, if the CFI ≥ .90 rule had been employed for decision making for weak invariance, CFI would have confirmed all 21 comparisons (CFI ranged from 0.98 to 1.00). These nearly contradictory conclusions reached by the $\Delta\chi^2$ p-value and CFI ≥ .90 demonstrated that they could be problematic in determining whether weak invariance holds, suggesting that $\Delta\chi^2$ is too strict and CFI is too lenient. Note that although we do not generally recommend this criterion for MI testing, we are referring to CFI ≥ .90 (in this and forthcoming comparisons of $\Delta\chi^2$ and CFI in later sections of this paper) because this is the cut-off commonly referred to in the literature.

Table 2: Fit Indices for Weak Invariance Models

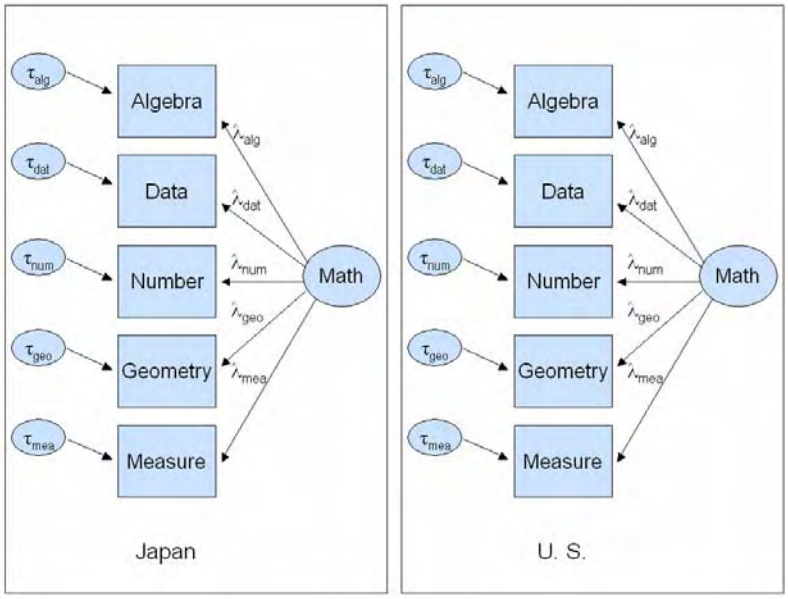
Comparison	χ^2	p	RMSEA	CFI	$\Delta\chi^2$	ΔCFI
AUS vs. NZL	21.96	0.11	0.03	1.00	10.78	0.00
CAN vs. USA	49.65	0.00	0.05	0.99	**26.48**	-0.01
AUS vs. CAN	16.07	0.38	0.01	1.00	6.42	0.00
AUS vs. USA	49.13	0.00	0.05	0.99	**32.28**	-0.01
USA vs. NZL	35.96	0.00	0.04	1.00	**11.26**	0.00
CAN vs. NZL	29.97	0.01	0.04	1.00	**12.48**	0.00
JPN vs. KOR	37.87	0.00	0.05	1.00	**18.09**	0.00
JPN vs. TWN	94.81	0.00	0.09	0.99	**87.62**	-0.01
TWN vs. KOR	64.60	0.00	0.07	0.99	**42.10**	-0.01
AUS vs. JPN	44.97	0.00	0.06	0.99	**41.06**	-0.01
AUS vs. KOR	37.83	0.00	0.05	0.99	**18.62**	-0.01
AUS vs. TWN	36.58	0.00	0.05	1.00	**29.96**	0.00
USA vs. TWN	83.76	0.00	0.07	0.99	**63.62**	-0.01
USA vs. KOR	87.31	0.00	0.07	0.99	**54.58**	-0.01
USA vs. JPN	139.37	0.00	0.10	0.98	**121.95**	**-0.02**
CAN vs. JPN	79.45	0.00	0.08	0.99	**69.23**	-0.01
CAN vs. KOR	51.55	0.00	0.06	0.99	**26.03**	-0.01
CAN vs. TWN	66.40	0.00	0.07	0.99	**53.47**	-0.01
TWN vs. NZL	29.69	0.01	0.04	1.00	**15.23**	0.00
NZL vs. JPN	79.46	0.00	0.09	0.98	**67.71**	**-0.02**
NZL vs. KOR	55.55	0.00	0.07	0.99	**28.50**	-0.01

Note. $\Delta df = 5$, $\chi^2_{0.05}(5, N) = 11.07$
Note. Rejections of weak invariance were highlighted in bold.

Test of Strong Invariance

Strong invariance postulates that, for all items, not only the cross-group loadings but also the intercepts are equal. If the score comparison is to be on the group means of the latent variable, it is necessary to make sure that the centers of the latent variable are scaled identically across groups (Millsap, 1998). This is tested by the equality in the calibration of the mean structure in addition to the variance/covariance structure (i.e., MACS) of the observed variables, which are nonetheless, widely neglected in the MI research practice as discussed earlier. (see Figure 5).

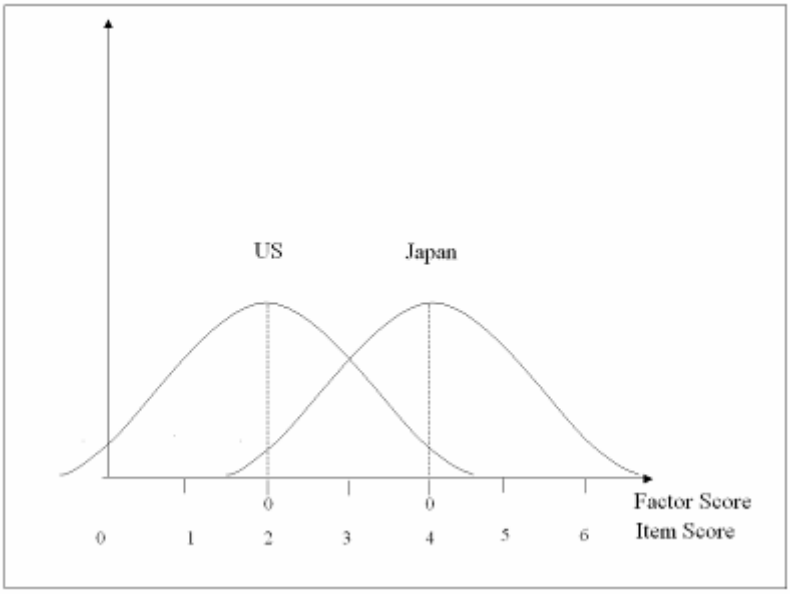
Figure 5: One-factor Five-indicator Strong Invariance Model for TIMSS



Note. τ_{alg}, τ_{dat}, τ_{num}, τ_{geo}, and τ_{mea} represent the intercepts and λ_{alg}, λ_{dat}, λ_{num}, λ_{geo}, and λ_{mea} represent the factor loadings for Algebra, Data, Number, Geometry, and Measurement.

Unequal calibration with regard to the intercept is illustrated in Figure 6, where the hypothetical density functions of factor scores for the US and Japan are intentionally placed on the item scale. For the US, the location of the density function is at 2 on the item scale but at 4 for Japan. Because the location is also the point where the factor score is zero, it is also the intercept of the regression line as shown in Figure 7. Thus, unequal cross-group intercept represents the unequal scaling of factor scores with regard to the location. The impact on factor score comparability resulting from unequal intercepts is shown in Figure 7. For Japanese students, a score of 5 (with a slope/loading of 1.0) is calibrated to have a factor of 1. However, the same item score is calibrated to have a factor score of 3 for the US students. This unequal factor score calibration with regard to the intercept can be verified by equation (3). Using equation (3), an item score of 5 would be calibrated to be factor score equal to (5-2)/1= 3 for the US students but only (5-4)/1 = 1 for the Japanese students. Such unequal calibration is, hence, biased against Japanese students.

Figure 6: Hypothetical Density Function of Factor Score



Note. The US and Japan's factor score densities were put on item scale to show that the centers of the factor scale are not located at the identical position on the item scale.

Figure 7: Impact of Lack of Intercept Equality on MI

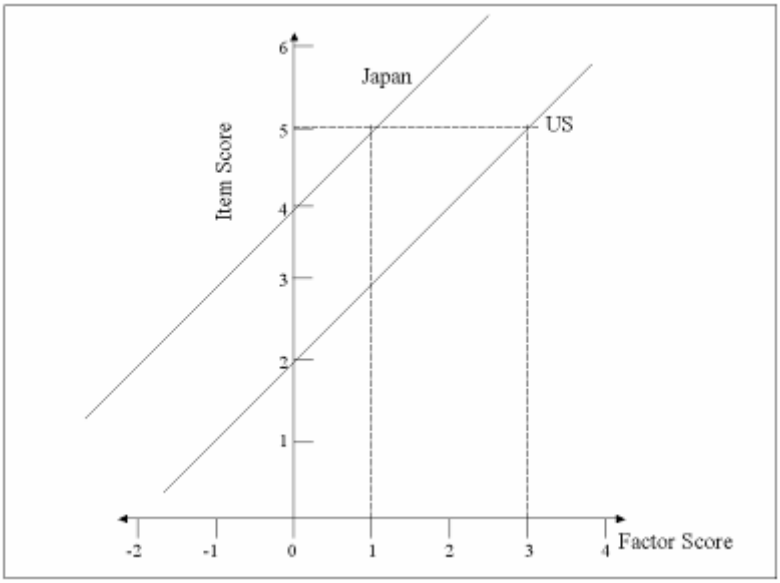
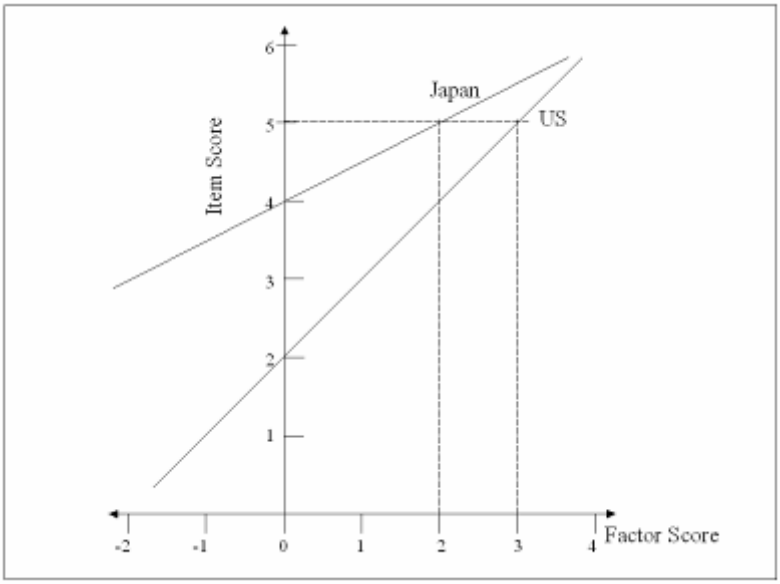


Figure 8 demonstrates the joint impact of both loading and intercept inequality on the factor score calibration (i.e., strong invariance violation). Following the hypothetical specification in Figures 4 and 7, it can be seen that both the intercepts and the slopes are unequal for the U.S. (slope= 1 and intercept = 2) and Japan (slope= 0.5 and intercept= 4). A score of 5 on the item score is calibrated to have a factor score of 2 for the Japanese students, but a factor score of 3 for the US students. This unequal factor score calibration with regard to the intercept and loading can be verified by equation (3). Using equation (3), an item score of 5 would be calibrated to be factor score equal to (5-2)/1= 3 for the US students but only (5-4)/0.5 = 2 for the Japanese students. Such unequal calibration is, hence, biased against the Japanese students. This differential item-factor calibration can also be

understood from an "interaction" perspective shown by the non-parallel regression lines in Figure 8. That is, the item-factor relationship is dependent on the group membership. In this sense, an item score of, say, "5" might mean something quite different with regard to the factor score across groups. In summary, cross-group inequality in the loading or/and intercept implies that, conditional on the factor scores (true score), a student's item score will depend on his/her group membership – i.e., DIF, a consequence that violates Meredith's (1993) definition of MI.

Appendix C provides the LISREL/SIMPLIS syntax for testing strong invariance. Table 3 shows the results for the strong invariance test with the MACS model. Note that strong invariance was not tested for two of the cross-culture comparisons: USA vs. JPN and NZL vs. JPN as indicated by the "--" sign in Table 3 because weak invariance was rejected by $\Delta CFI \leq -0.01$ for these two comparisons in the previous examination, hence MI investigations were terminated at the weak invariance level. Adopting the $\Delta CFI \leq -0.01$ decision rule, seven out of the nine within-culture comparisons passed the strong invariance test; the two exceptions were AUS vs. USA and JPN vs. TWN. However, all the cross-culture comparisons failed the strong invariance test.

Figure 8: Joint Impact of Loading and Intercept Inequality on MI



The large drop in the number of confirmations from weak invariance test (19) to strong invariance test (7) indicated that despite the similarities in the factor loadings, inequality in the regression intercepts was prevalent among the paired comparisons. This indicated that the test was consistently biased against one of the countries in the planned pairs, and this phenomenon was universal for all cross-culture comparisons. If MI had been defined as loading or loading/error equality excluding the examination of intercept equality (i.e., MG-CFA on covariance structure only), biases in the factor score comparison due to unequal calibration in the location (as shown in Figures 7 and 8) would not have been detected; as highlighted in the literature by Zumbo (2003) and Zumbo and Koh (2005). Also note that, as found earlier, $\Delta\chi^2$ and CFI $\geq .90$ rules were problematic in determining strong invariance models; this observation was demonstrated by the findings that almost contradictory conclusions were reached by $\Delta\chi^2$ and CFI $\geq .90$. That is, $\Delta\chi^2$ rejected all 21 comparisons for strong invariance whereas CFI $\geq$ 0.90 rule confirmed 19 comparisons.

Test of Strict Invariance

In regression language, strict invariance dictates that, in addition to intercepts and the slopes, the regression residual variances for all the items are equal across groups (see Figure 9).

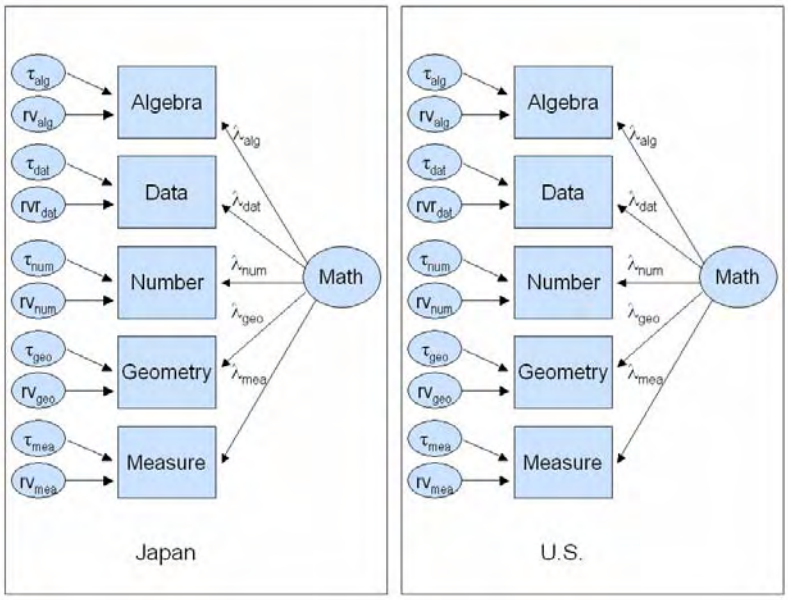
The residual variance is the portion of item variance not attributable to the factor(s). Until recently, it was believed that fixing the residual variances to be equal subsequent to a support for strong invariance is an unnecessarily rigorous requirement for MI (Little, 1997; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000; Widaman, & Reise, 1997). For example, after reviewing the inconsistencies in the literature regarding constraints on the residuals, Vandenberg and Lance (2000) recommended that the evaluation of the residual variance equality be left to the researcher's discretion. The rationale behind this thinking is that, if strong invariance holds, group difference in the residual variances is indicative of only the difference in reliabilities of the observed scores; thus, group difference is compensated if comparison is to be made on the latent variable level. Following this rationale, significant improvement in fit is interpreted as difference in measurement reliability (i.e., random noise) rather than evidence of bias.

Table 3: Fit Indices for Strong Invariance Models

Comparison	χ^2	p	RMSEA	CFI	$\Delta\chi^2$	ΔCFI
AUS vs. NZL	65.64	0.00	0.07	0.99	**43.68**	-0.01
CAN vs. USA	129.83	0.00	0.08	0.98	**80.18**	-0.01
AUS vs. CAN	32.10	0.04	0.03	1.00	**16.03**	0.00
AUS vs. USA	185.41	0.00	0.10	0.97	**136.28**	**-0.02**
USA vs. NZL	80.94	0.00	0.06	0.99	**44.98**	-0.01
CAN vs. NZL	61.62	0.00	0.06	0.99	**31.65**	-0.01
JPN vs. KOR	71.66	0.00	0.06	0.99	**33.79**	-0.01
JPN vs. TWN	175.09	0.00	0.01	0.97	**80.28**	**-0.02**
TWN vs. KOR	87.90	0.00	0.07	0.99	**23.30**	0.00
AUS vs. JPN	201.15	0.00	0.13	0.95	**156.18**	**-0.04**
AUS vs. KOR	207.00	0.00	0.12	0.96	**169.17**	**-0.03**
AUS vs. TWN	170.88	0.00	0.11	0.97	**134.30**	**-0.03**
USA vs. TWN	661.20	0.00	0.18	0.91	**577.44**	**-0.08**
USA vs. KOR	766.33	0.00	0.19	**0.89**	**679.02**	**-0.10**
USA vs. JPN	710.71	0.00	0.19	**0.88**	**571.34**	--
CAN vs. JPN	360.62	0.00	0.15	0.92	**281.17**	**-0.07**
CAN vs. KOR	366.37	0.00	0.14	0.93	**314.82**	**-0.06**
CAN vs. TWN	319.32	0.00	0.13	0.95	**252.92**	**-0.04**
TWN vs. NZL	319.69	0.00	0.15	0.95	**290.00**	**-0.05**
NZL vs. JPN	400.08	0.00	0.18	0.90	**320.62**	--
NZL vs. KOR	420.80	0.00	0.18	0.91	**365.25**	**-0.08**

Note. $\Delta df = 5$, $\chi^2_{0.05}(5, N) = 11.07$
Note. Rejections of strong invariance were highlighted in bold.
Note. "--" indicates that the strong invariance test was not legitimate, because the weak MI did not hold.

Figure 9: One-factor Five-indicator Strict Invariance Model for TIMSS



Note. τ_{alg}, τ_{dat}, τ_{num}, τ_{geo}, and τ_{mea} represent the intercepts, rv_{alg}, rv_{dat}, rv_{num}, rv_{geo}, and rv_{mea} represent the residual variances, and λ_{alg}, λ_{dat}, λ_{num}, λ_{geo}, and λ_{mea} represent the factor loadings for Algebra, Data, Number, Geometry, and Measurement.

However, Deshon (2004) and Lubke and Dolan (2003) maintained that the above statement against the necessity for strict invariance is true if, and only if, the assumption of conditional independence holds. That is, there are no inter-correlations among the item residuals after accounting for the factor scores. Theoretically, the residuals are "assumed" to be conditional independent and are simply the results of unpredictable fluctuations in the measurement process, namely, random errors. Or, more precisely, "IF" the conditional independence assumption holds, an item's residual is neither correlated with those of the other items, nor with the common factors, after conditioning on the factor score. Deshon (2004) and Lubke and Dolan (2003), like earlier psychometricians such as Rozeboom (1966), argued that, in practice, however, it is not uncommon to observe the violation of conditional independence, even to a small degree. They argued that residuals might be the results of both the unpredictable fluctuations and the systematic effects of "unintentionally measured yet un-modeled variable(s)" of one or some particular items (e.g., method effect or minor secondary dimensions). Hence, residual variance of an item consists of not only the random variation, but also the variation due to the effect of unmodeled sources of systematic effects that influence people's item responses (i.e., extraneous variables, for example, difference in the coverage of curriculum or translation effect). Based on Cronbach's (1947) statement on error, Deshon (2004, p. 144) stated that "if the error variances are different across groups, then there are either different variables operating on the measures across groups or the same set of variables operates differently across groups".

Deshon (2004) further argued that common factor analysis does not eliminate or partial out the effects of unmodeled extraneous variable(s). He contended that the belief that systematic item-specific sources of variance are removed from the estimation of the latent variable is based on the ideal assumption that the residuals of items are uncorrelated with each other or the latent variable. Deshon (2003, p.146) stated:

> *These two innocent sounding assumptions of the common factor model are the source of much interpretational ambiguity concerning the meaning of the latent variables. If one believes that the variance in an item response that is not due to the latent variable is completely random noise, then the argument that error variance MI is unnecessary is valid. However, if one adopts Cronbach's (1947) position that the*

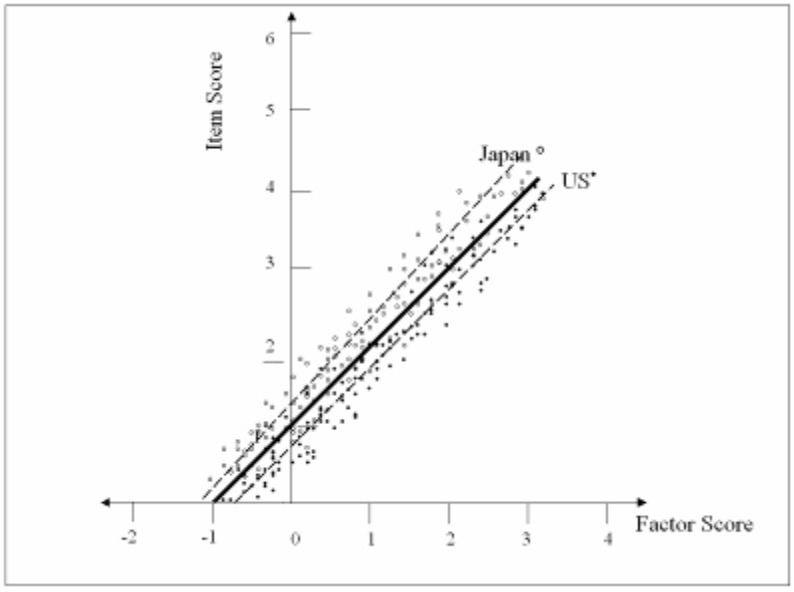
variance not due to the latent variable is actually due to other causative variables that affect item response, then the assumption is almost wrong in every single application of common factor analysis.

Necessity for strict invariance can be easily understood from our item-factor regression notion. Remember that in order to demonstrate the effect of lack of strong invariance, we temporarily ignore the existence of the residual term in equation (1). Now it's appropriate to bring back the complete equation (1) that contains the regression residual term and further partition the residual term into two parts: a) s_{ij}, which is the effect that is unintentionally yet systematically measured by a specific item (or items) and b) r_{ij}, which is the random fluctuations of unreliability, yielding

$$y_{ij} = \tau_j + \lambda_{j1}\eta_{1i} + \lambda_{j2}\eta_{2i} + \ldots \lambda_{jp}\eta_{pi} + s_{ij} + r_{ij}, \quad (4)$$

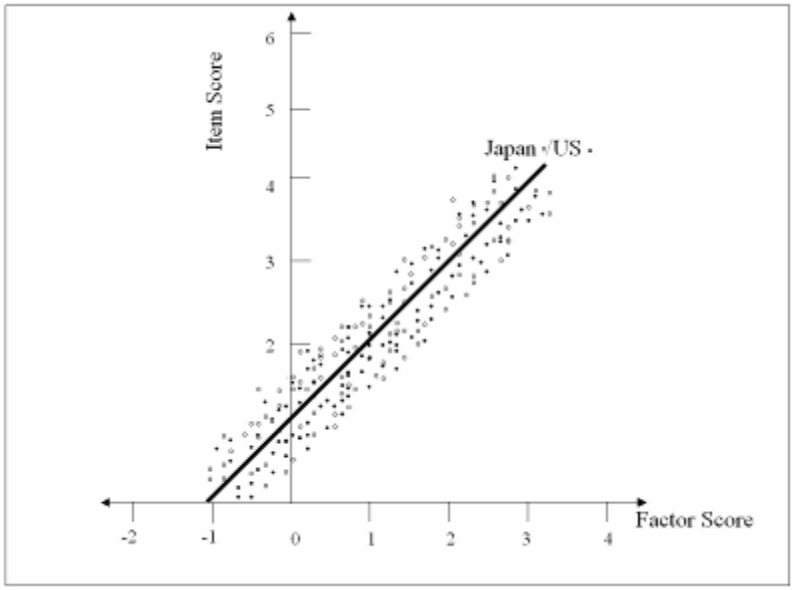
If systematic effect, s_{ij}, is present due to group membership and leads to the mean of the residuals to be higher (or lower) for one group or leads to unequal variation of the residuals between the groups, then such effect will shift the two regression lines away from the identical regression position achieved by strong invariance as a result of the cross-group inequality in the residual mean, the variance, or the joint impact of the two and obfuscate the equality in item-factor calibration. This effect is demonstrated in Figure 10. One can see that the residual values are systematically higher for Japan (indicated by "○") then those of the U.S. (indicated by "●"), as is the variation among the Japanese respondents. As a result, the individual item-factor regression lines indicated by the dotted lines are shifted away from the identical position achieved by strong invariance. Hence, the strong MI condition is obfuscated by the item specific systematic effect.

Figure 10: Impact of Systematic Item-specific Effect on MI



If the item residuals consist of only the random errors, r_{ij} (i.e., the conditional independence assumption is met), they will not obfuscate the MI achieved by strong invariance because the random errors are expected to cancel each other out and result in means of zeros for each group and the residual variances will remain equal for the two groups. Hence, random error has no effect on the cross-group item-factor relationships and the identical calibration lines achieved by strong invariance will remain stable (see Figure 11).

Figure 11: Impact of Random Error Effect on MI



Deshon (2004) argued that regardless of the methods used to estimate the factor scores, common factor analysis does not remove the influence of unwanted systematic effect. Common factor analysis simply reduces the variance of the observed variables to what might have been if such systematic effect did not affect the measured variables. The variance of the measured variable can be reduced but the influence of the unwanted systematic variables cannot be removed. He maintained that strong invariance is sufficient for MI only if the non-random residuals have influenced the common variance of the items and not differentially influenced their specific variances.

In our view, Deshon's (2004) and Lubke and Dolan's (2003) argument against the capacity of common factor analysis to remove unwanted systematic effect can be easily understood from a multi-dimensionality perspective of bias (Ackerman, 1992; Shealy & Stout, 1993). When exploratory factor analysis (i.e., unrestricted factor analysis) is applied separately to the groups, existence of bias is deemed feasible if the item specific variances result in the formation of an extra factor(s) for one group but not the other or for both groups with unequal means and/or variances on the extra factor. This effect of unequal means and/or variances in the extra factor is allowed to freely yield different estimates of the intercepts, loadings, and residual variances across groups because the parameter estimation are conducted separately for each group. However, imagine that a "strong" MG-CFA MI is specified according to the researcher's theory, the number of factors is forced to be the same, and so are the loadings and intercepts. Such constraints will allow item specific effects to reside only in the residual terms and remain undetected if strict MI is not investigated and consequently disguising possible biases in the test scores. In situations where items are sound measures of the construct (i.e., communalities are high) and the model specification is correct for all groups (i.e., low residual variances and uncorrelated errors), testing strict invariance would likely reach the same conclusion as the strong invariance test would. However, such a desirable scenario is not always guaranteed. Hence, a judicious modeling strategy should always incorporate a test of strict invariance as a prudent step for ensuring MI rather than an unnecessarily rigorous requirement.

Appendix D provides the LISREL/SIMPLIS syntax for testing strict invariance with the MACS model. Table 4 shows the results for the strict invariance test. Observe that none of the cross-culture comparisons were tested for strict invariance as indicated by the "--" sign in Table 4 because strong invariance was rejected by the $\Delta CFI \leq -0.01$ rule in the previous analyses. For within-culture comparisons, strict invariance and strong invariance came to the same conclusions. All 7 within-culture comparisons that passed the strong invariance test also passed the strict invariance test. Hence, testing

strict invariance did not alter the final decision about MI (see Table 4). Again, a large number of contradictory conclusions between $\Delta\chi^2$ and CFI $\geq$ .90 were observed as in the weak and strong invariance examinations.

SUMMARY

The purpose of the paper is to decode the meaning of MI and update the practice of MG-CFA. In essence, our purpose is one of knowledge translation from the technical psychometric and statistical literature. We explained why strict invariance is a necessary condition for ensuring MI and why it should always be tested. We demonstrated: (a) why inequalities in the loadings and intercepts have a direct detrimental effect on the item-factor score calibration, and (b) how inequality in the residuals may distort the loading/intercept metric equality. In particular, we stress the necessity for modeling MACS factorial invariance so that the centers of the latent variable are scaled identically for the group mean comparison to be meaningful. Equally important is the testing the existence of group-related systematic effect in the residuals by the strict invariance. Unless residual variances of the measured variables can be clearly shown to be only a reflection of random errors, as a prudent step, equality in the residual terms should always be tested.

Table 4: Fit Indices for Strict Invariance Models

Comparison	χ^2	p	RMSEA	CFI	$\Delta\chi^2$	ΔCFI
AUS vs. NZL	71.81	0.00	0.06	0.99	6.17	0.00
CAN vs. USA	139.47	0.00	0.07	0.98	9.64	0.00
AUS vs. CAN	32.70	0.14	0.02	1.00	0.60	0.00
AUS vs. USA	197.82	0.00	0.10	0.97	**12.41**	--
USA vs. NZL	88.91	0.00	0.06	0.99	7.97	0.00
CAN vs. NZL	69.87	0.00	0.05	0.99	8.25	0.00
JPN vs. KOR	91.35	0.00	0.06	0.99	**19.69**	0.00
JPN vs. TWN	189.87	0.00	0.10	0.97	**14.78**	--
TWN vs. KOR	98.58	0.00	0.06	0.99	10.68	0.00
AUS vs. JPN	286.09	0.00	0.13	0.93	**84.94**	--
AUS vs. KOR	269.74	0.00	0.12	0.95	**62.74**	--
AUS vs. TWN	217.86	0.00	0.11	0.96	**46.98**	--
USA vs. TWN	754.69	0.00	0.17	0.90	**93.49**	--
USA vs. KOR	889.86	0.00	0.18	**0.87**	**123.53**	--
USA vs. JPN	823.39	0.00	0.18	**0.86**	**112.68**	--
CAN vs. JPN	448.51	0.00	0.14	0.91	**87.89**	--
CAN vs. KOR	445.92	0.00	0.14	0.92	**79.55**	--
CAN vs. TWN	377.13	0.00	0.13	0.94	**57.81**	--
TWN vs. NZL	365.40	0.00	0.14	0.94	**45.71**	--
NZL vs. JPN	483.88	0.00	0.18	0.88	**83.80**	--
NZL vs. KOR	485.02	0.00	0.17	0.90	**64.22**	--

Note. $\Delta df = 5$, $\chi^2_{0.05}(5, N) = 11.07$
Note. Rejections of strict invariance were highlighted in bold.
Note. "--" indicates that the strict invariance test was not legitimate, because the strong MI did not hold.

We also discuss that the MG-CFA decision about rejecting or supporting MI should not rely solely on either $\Delta\chi^2$ test or fit indices. Instead, researchers should consider using change in fit indices, in particular, ΔCFI, ΔGamma Hat, or ΔNon-Centrality Index. These conclusions were adopted to guide our example investigation of construct comparability in the scores of the TIMSS mathematics test. Table 5 summarizes the results of the TIMSS MI investigations where within-culture

Table 5. Summary Results of MI for 21 Planned Comparisons

	AUS	NZL	CAN	USA	TWN	KOR
NZL	**Strict**					
CAN	**Strict**	**Strict**				
USA	**Weak**	**Strict**	**Strict**			
TWN	Weak	Weak	Weak	Weak		
KOR	Weak	Weak	Weak	Weak	**Strict**	
JPN	Weak	Configure	Weak	Configure	**Weak**	**Strict**

Note. Results for within-culture comparisons were highlighted in bold.

comparisons were highlighted in bold and the cross-culture comparisons were grouped within a rectangle

What do the results tell us and how should the results of an MI investigation be interpreted? In the TIMSS example, the general pattern observed for the within-culture comparisons is that MI is demonstrated. That is, for within-culture comparisons, the same construct is measured and is measured on the same metric. Hence, if any difference in the factor score is found, one can be assured that such difference is a result of a true difference in the amount of mathematics proficiency rather than measurement artifact. Also, we are assured that comparing and explaining variation is meaningful regardless of the group membership because cross-group variances are assured to be on the same metric. This broad statement does not imply that MI is guaranteed if the comparison is among countries that share similar cultural paradigms. In fact, MI in the AUS vs. USA and JPN vs. TWN comparisons was found to be absent in this study. In other words, although it is very likely that construct comparability does exist among countries that share the same cultural paradigms, MI should never be simply assumed.

For cross-culture MI examinations, only weak invariance, at best, is achieved. This result indicates that intercept invariance does not hold for any of the cross-culture comparisons, hence, the mathematics test, as a whole, was consistently biased against one of the countries in the pairs. One cannot infer that there is true group difference even if the hypothesis test, such as a t-test, is significant because the detected difference might be an artifact of the measurement bias. Any research or policy exercise such as ranking performances or explaining group differences based on such mathematics proficiency scores is not meaningful because mathematics proficiency scores were not measured on the same metric unless some forms of linking or equating, which have their own variation of MI assumptions, is performed before comparison. This is an important point for policy makers, and school effectiveness researchers who value and interpret country rankings. Country rankings, which are commonly found in the media and in policy discussions, are only meaningful if MI has been empirically demonstrated.

Closing Remarks

It is interesting to note that for our TIMSS example, strong and strict invariance reach the same conclusions. This seems to suggest that, if the items communalities are high (e.g., .82, .56, .80, .59, and .79 for the five domains for Taiwan) and the model is correctly specified (indicated, for example, by the good configural fit), tests of strong and strict invariance will likely reach the same conclusions. Readers may, hence, overlook the necessity for strict MI. It is vital to realize that this fortunate coincidence in results is never known a priori. It is not guaranteed that the strong and strict invariance examinations will always come to the same conclusions for other datasets, especially when the model specification of the configural model is uncertain and the communalities of the observed indicators are low. Strict invariance detects potential obstruction of strong invariance due to the item-specific systematic effect. Hence, testing strict invariance should be considered as a prudent step rather than an unnecessarily strict requirement for ensuring MI and should always be employed for MI

investigation. It is apparent when one reads the research literature where in MI is applied that some researchers envision MI as configural or weak invariance, these researchers appear to operate with the principle that strong and strict invariance are unnecessary, and, at best, are psychometric niceties. In this paper, we have shown in detail why this is not the case and why researchers should test for strong and strict invariance.

As the literature suggested, in this study, χ^2 does not provide practical usefulness in testing configural invariance. Likewise, the $\Delta\chi^2$ or CFI $\geq$ 0.90 rules do not appear useful for testing nested MI models. This conclusion is supported by the highly inconsistent (and almost always contradictory) conclusions reached by $\Delta\chi^2$ and CFI $\geq$ 0.90 rules. This contradiction in the MI conclusion should warn researchers that decisions based on either $\Delta\chi^2$ or CFI, as widely applied in today's practice, could be problematic.

On a statistical methodology note, the MG-CFA methods we reviewed and applied were largely based on the maximum likelihood estimation methods, which assume multivariate normal data. Future Monte Carlo studies should be conducted to verify these decision rules. In addition, to our knowledge, no study has investigated the appropriate fit indices for MG-CFA on polychoric correlation matrix for categorical data using alternative estimation methods such as weighted least squares.

In short, this study demonstrates that the success of an MI MG-CFA investigation lies in the researcher's lucid understanding of strict invariance as well as an informed choice of the appropriate fit indices. Cut-off values should be employed carefully in relation to the characteristics of the data such as sample size, complexity in the data structure, and the estimation methods used.

REFERENCES

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multi-dimensional perspective. *Journal of Educational Measurement, 29*, 67-91.

Bandalos, D. L. (2002). The effects of item parcelling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling, 9*, 78-102.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior, 16*, 201-13.

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for equivalence of factor covariance and mean structures: The issue of partial MI. *Psychological Bulletin, 105*, 456-466.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing MI. *Structural Equation Modeling, 9*, 235-55.

Cronbach, L. J. (1947). "Test reliability": Its meaning and determination. *Psychometrika, 12*, 1-16.

Deshon, R. P. (2004). Measures are not invariant across groups with error variance homogeneity. *Psychology Science, 46*, 137-49.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Models, 12*, 343-67.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to MI in aging research. *Experimental Aging Research, 18*, 117-144.

Hu, L., & Bentler, P. (1998). Fit Indices in covariance structure modeling: Sensitivity to underparameterized Model Misspecification*, Psychological Methods, 3*, 424-53.

Jöreskog, K. G., & Sörbom, D. (1999). *LISREL 8 user's reference guide.* Chicago: Scientific Software International.

Kelloway, E. K. (1995). Structural equation modeling in perspective. *Journal of Organizational Behavior, 16,* 215-24.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*, 53-76.

Lubke, G. H., & Dolan, C. V. (2003). Can unequal residual variance across groups mask difference in residual means in the common factor model? *Structural Equation Modeling, 10*, 175-192.

Macnab, D. (2000a). Forces for changes in mathematics in education: The case of TIMSS. *Education Policy Analysis Archives, 8(*15).

Macnab, D. (2000b). Raising standards in mathematics education: Values, vision, and TIMSS. *Educational Studies in Mathematics, 42*, 61-81.

Marsh, H. W. (1994). Confirmatory factor models of factorial invariance: A multi-faceted approach. *Structural Equation Modeling, 1*, 5-34.

Marsh, H. W., Hau, K., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cut-off values for fit indexes and dangers in overgeneralization Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*(3), 320-341.

McArdle, J. J. (1998). Contemporary statistical models of test bias. In J. J. McArdle & R. W. Woodcock (Eds.), *Human abilities in theory and practice* (pp. 157-195). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification, 6*, 97-103.

McGaw, B., & Jöreskog, K. G. (1971). Factorial invariance of ability measures in groups differing in intelligence and socio-economic status. *British Journal of Mathematics and Statistical Psychology*, *24*, 154-168.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 127-43.

Meredith, W. (1993). MI, factor analysis and factorial invariance. *Psychometrika, 58*, 525-43.

Meredith, W. & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement invariance. *Psychometrika, 57(2)*, 289-311.

Millsap, R. E. (1998). Group differences in regression intercept: Implication for factorial invariance. *Multivariate Behavioral Research, 33(3)*, 403-424.

National Institute on Educational Governance, Finance, Policymaking, and Management, (1998). What the Third International Mathematics and Science Study (TIMSS) means for systemic school Improvement. (Report No. GFI-98-9501). Washington, DC: U.S. Government Printing Office.

Rozeboom, W.W. (1966). *Foundations of the theory of prediction*. Homewood, IL: Dorsey.

Schermelleh-Engel, K., Moonsbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance measures. *Methods of Psychological Research Online, 8*, 23-74.

Schmidt, W. H., McKnight, C. C., Cogan, L. S., Jakwerth, P. M., & Houang, R. T. (1999). *Facing the consequences: Using TIMSS for a closer look at U. S. mathematics and science education*. Netherlands: Kluwer Academic Publishers.

Shealy, R. & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Steenkamp, J. E., & Baumgartner, H. (1998). Assessing MI in cross-national consumer research. *Journal of Consumer Research, 25*, 78-90.

Steiger, J. H. (1989). *EzPATH: Causal modeling*. Evanston, IL: SYSTAT.

The International Association for the Evaluation of Educational Achievement; The International Study Center, Lynch School of Education (2001). *TIMSS 1999 International Mathematics Report (Chap. 1)*. Retrieved March 5, 2006 from the TIMSS 1999 Web site: http://timss.bc.edu/timss1999i/math_achievement_report.html

The International Association for the Evaluation of Educational Achievement; The International Study Center, Lynch School of Education. (2004). *TIMSS 2003 International Mathematics Report (Chap. 1)*. Retrieved March 5, 2006 from the TIMSS 2003 Web site: http://timss.bc.edu/PDF/t03_download/T03_M.Chap1.pdf

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1-10.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the MI literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-69.

Widaman, K. F., & Reise, S. P. (1997). Exploring the MI of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281-324). Washington, DC: American Psychological Association.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao and S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 45-79). Elsevier Science B.V.: The Netherlands.

Zumbo, B. D. (2003). Does Item-Level DIF Manifest Itself in Scale-Level Analyses?: Implications for

Translating Language Tests. *Language Testing, 20*, 136-147

Zumbo, B. D., & Koh, K. H. (2005). Manifestation of Differences in Item-Level Characteristics in Scale-Level Measurement Invariance Tests of Multi-Group Confirmatory Factor Analyses. *Journal of Modern Applied Statistical Methods, 4*, 275-282.

Zumbo, B. D., Sireci, S. G., & Hambleton, R. K. (2003, April). Re-visiting exploratory methods for construct comparability and MI: Is there something to be gained from the ways of old? Paper presented at the Annual Meeting of the National Council for Measurement in Education (NCME), Chicago, Illinois

Appendix A: SIMPLIS Syntax for Testing Configural Invariance

**** Note that for multi-group analyses SIMPLIS works on the principle that any**
**** parameters specified in the second group are freely estimated in the second group –**
**** hence they are allowed to be different in the second group.**

Group 1: Japan
Raw Data from File JPN_DOMAIN.PSF
Observed Varibles: ALGEBRA DATA NUMBER GEOMETRY MEASURE
Latent Variables: MATH

Relationships
ALGEBRA = CONST MATH
DATA = CONST MATH
NUMBER = CONST MATH
GEOMETRY = CONST MATH
MEASURE = CONST MATH

Group 2: USA
Raw Data from File USA_DOMAIN.PSF

Relationships
ALGEBRA = CONST MATH
DATA = CONST MATH
NUMBER = CONST MATH
GEOMETRY = CONST MATH
MEASURE = CONST MATH

Set the Error Variance of ALGEBRA Free
Set the Error Variance of DATA Free
Set the Error Variance of NUMBER Free
Set the Error Variance of GEOMETRY Free
Set the Error Variance of MEASURE Free

Path Diagram
End of Problem

Appendix B: SIMPLIS Syntax for Testing Weak Invariance

**** Note that for multi-group analyses SIMPLIS works on the principle that any**
**** parameters specified in the second group are freely estimated in the second group –**
**** hence they are allowed to be different in the second group.**

Group 1: Japan
Raw Data from File JPN_DOMAIN.PSF
Observed Variables: ALGEBRA DATA NUMBER GEOMETRY MEASURE
Latent Variables: MATH

Relationships
ALGEBRA = CONST MATH
DATA = CONST MATH
NUMBER = CONST MATH
GEOMETRY = CONST MATH
MEASURE = CONST MATH

Group 2: USA
Raw Data from File USA_DOMAIN.PSF

Relationships
ALGEBRA = CONST
DATA = CONST
NUMBER = CONST
GEOMETRY = CONST
MEASURE = CONST

Set the Error Variance of ALGEBRA Free
Set the Error Variance of DATA Free
Set the Error Variance of NUMBER Free
Set the Error Variance of GEOMETRY Free
Set the Error Variance of MEASURE Free

Path Diagram
End of Problem

Appendix C: SIMPLIS Syntax for Testing Strong Invariance

**** Note that for multi-group analyses SIMPLIS works on the principle that any**
**** parameters specified in the second group are freely estimated in the second group –**
**** hence they are allowed to be different in the second group.**

Group 1: Japan
Raw Data from File JPN_DOMAIN.PSF
Observed Varibles: ALGEBRA DATA NUMBER GEOMETRY MEASURE

```
Latent Variables: MATH

Relationships
ALGEBRA = CONST MATH
DATA = CONST MATH
NUMBER = CONST MATH
GEOMETRY = CONST MATH
MEASURE = CONST MATH

Group 2: USA
Raw Data from File USA_DOMAIN.PSF
Set the Error Variance of  ALGEBRA  Free
Set the Error Variance of  DATA  Free
Set the Error Variance of  NUMBER  Free
Set the Error Variance of  GEOMETRY  Free
Set the Error Variance of  MEASURE  Free

Path Diagram
End of Problem
```

Appendix D: SIMPLIS Syntax for Testing Strict Invariance

```
** Note that for multi-group analyses SIMPLIS works on the principle that any
** parameters specified in the second group are freely estimated in the second group –
** hence they are allowed to be different in the second group.

Group 1: Japan
Raw Data from File JPN_DOMAIN.PSF
Observed Variables: ALGEBRA DATA NUMBER GEOMETRY MEASURE
Latent Variables: MATH

Relationships
ALGEBRA = CONST MATH
DATA =  CONST MATH
NUMBER = CONST MATH
GEOMETRY = CONST MATH
MEASURE = CONST MATH

Group 2: USA
Raw Data from File USA_DOMAIN.PSF

Path Diagram
End of Problem
```

Author Note

We would like to thank the editor, Dr. Lawrence Rudner, and the two reviewers for feedback that greatly improved this paper.

Citation

Wu, Amery D., Li, Zhen and Zumbo, Bruno D. (2007). Decoding the Meaning of Factorial Invariance and Updating the Practice of Multi-group Confirmatory Factor Analysis: A Demonstration With TIMSS Data. *Practical Assessment Research & Evaluation*, 12(3). Available online: http://pareonline.net/getvn.asp?v=12&n=3

Authors

Correspondence concerning this paper should be addressed to

Bruno D. Zumbo, Professor
University of British Columbia
Scarfe Building, 2125 Main Mall
Vancouver, B.C.
CANADA V6T 1Z

E-mail: bruno.zumbo [at] ubc.ca