International Journal of Psychological Research, 2010. Vol. 3. No. 1.
ISSN impresa (printed) 2011-2084
ISSN electrónica (electronic) 2011-2079

Milfont, T. L., Fischer, R., (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3* (1), 111-121.

Testing measurement invariance across groups: Applications in cross-cultural research.

Probando la invariancia de mediciones entre varios grupos: aplicaciones en la investigación transcultural.

Taciano L. Milfont
Victoria University of Wellington
Ronald Fischer
Victoria University of Wellington

ABSTRACT

Researchers often compare groups of individuals on psychological variables. When comparing groups an assumption is made that the instrument measures the same psychological construct in all groups. If this assumption holds, the comparisons are valid and differences/similarities between groups can be meaningfully interpreted. If this assumption does not hold, comparisons and interpretations are not fully meaningful. The establishment of measurement invariance is a prerequisite for meaningful comparisons across groups. This paper first reviews the importance of equivalence in psychological research, and then the main theoretical and methodological issues regarding measurement invariance within the framework of confirmatory factor analysis. A step-by-step empirical example of measurement invariance testing is provided along with syntax examples for fitting such models in LISREL.

Key words: measurement invariance, cross-cultural research, confirmatory factor analysis, LISREL.

RESUMEN

Los investigadores a menudo comparan grupos de individuos en diferentes variables psicológicas. Cuando se comparan grupos se asume que el instrumento usado para la medición da cuenta de los mismos constructos psicológicos en todos los grupos. Si tal suposición es cierta, las comparaciones son válidas y las diferencias/similitudes entre los grupos pueden ser interpretadas apropiadamente. Si tal suposición no es cierta, las comparaciones e interpretaciones pierden validez. El establecimiento de la invariancia en las mediciones es un prerrequisito esencial para lograr comparaciones apropiadas entre grupos. En este artículo se presenta primero la importancia de la invariancia en investigación psicológica y luego se presentan asuntos teóricos y metodológicos en relación con la invariancia en las mediciones dentro del marco del análisis factorial confirmatorio. Se presenta un ejemplo en LISREL que ejemplifica la prueba de invariancia de mediciones.

Palabras clave: invariancia en las mediciones, investigación transcultural, análisis factorial confirmatorio, LISREL

Article received/Artículo recibido: December 15, 2009/Diciembre 15, 2009, Article accepted/Artículo aceptado: March 15, 2009/Marzo 15/2009

Dirección correspondencia/Mail Address:
Taciano L. Milfont, School of Psychology, Victoria University of Wellington Email: Taciano.Milfont@vuw.ac.nz
Ronald Fischer, School of Psychology, Victoria University of Wellington. Email: Ronald.Fischer@vuw.ac.nz

1. Introduction

Psychological research often compares groups on psychological variables. This is one of the fundamental approaches in cross-cultural research; an instrument that has been found to show adequate psychometric properties in one cultural group is translated and administered to another cultural group. When comparing groups researchers often assume that the instrument (e.g., questionnaires, ability tests) measures the same psychological construct in all groups. Despite its appeal, this assumption is often not justified and needs to be tested.

Testing for equivalence of measures (or measurement invariance) has thus become an important issue in recent years (e.g., Chen, 2008; Cheung & Rensvold, 1999, 2000; Fontaine, 2005; Little, 1997; Steenkamp & Baumgartner, 1998; van de Vijver & Fischer, 2009; van de Vijver & Leung, 1997; van de Vijver & Poortinga, 1982), especially in cross-cultural research because it allows the researcher to check if members of different groups (e.g., female vs. male) or cultures (e.g., Brazilian vs German students) ascribe the same meanings to scale items (for recent applications, see Fischer et al., 2009; Gouveia, Milfont, Fonseca, & Coelho, 2009; Milfont, Duckitt, & Cameron, 2006; Milfont, Duckitt, & Wagner, in press). This paper discusses the issue of measurement invariance and its importance to psychological research in general and cross-cultural research in particular. The paper starts with a brief discussion of equivalence and biases in psychological research. The paper then focuses on measurement invariance, discussing its conceptualization and statistical analyses. A step-by-step empirical example of invariance testing is then provided along with syntax examples for fitting such models in LISREL (using the SIMPLIS command language) (Jöreskog & Sörbom, 1999). LISREL, an abbreviation for linear structural relations, is the pioneering statistical software package for structural equation modeling.

2. Equivalence in cross-cultural research

Any data collected for psychological research may yield unreliable results due to measurement biases. Psychological assessment based on self-reported measures is not different. This is especially important when data is gathered from two or more cultural groups, and when the data is used to compare the groups. In the cross-cultural literature, four levels of equivalence have been distinguished (Fontaine, 2005; van de Vijver & Leung, 1997): functional equivalence (does the construct exist in all groups studied), structural equivalence (are indicators related to the construct in a non-trivial way), metric equivalence (are loading weights identical across groups) and full score or scalar equivalence (are intercepts, that is the origin of measurement scales, identical across groups). A detailed discussion of these types of equivalence is beyond the scope of this article, but interested readers could find discussion of these issues and specific ways to address them elsewhere (Berry, Poortinga, Segall, & Dasen, 2002, Chapter 11; van de Vijver & Leung, 1997, 2000). Briefly, functional (and to some extent structural) equivalence can not be directly tested using statistical methods. Expert judgements and qualitative methods are best to identify these forms of non-equivalence. The statistical method discussed in this paper can only be useful if at least functional equivalence is being met by a test.

If we assume that functional equivalence is being met, are there specific strategies for overcoming and addressing biases? How can one identify specific forms of measurement non-equivalence? Measurement invariance testing within the framework of structural equation modeling can answer these questions because it is a robust procedure for investigating equivalence in multi-group data.

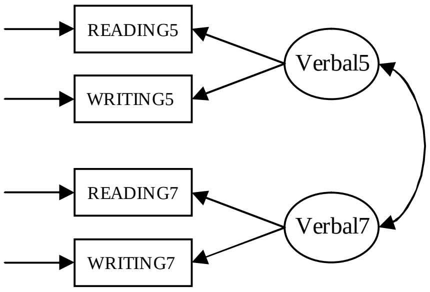
3. Assessing Equivalence

Psychological constructs, such as attitudes, beliefs and values, constitute latent variables that can not be measured directly. As a result, psychological measures function as indicators of the latent construct. In order to meaningfully compare a latent construct across groups, each observed indicator must relate to the latent variable in the same way in all groups. Given that most of the research in cross-cultural psychology focuses on the comparison of psychological constructs across different cultural groups (van de Vijver & Leung, 2000), the issue of equivalence of psychological measures is essential.

Multidimensional scaling, principal component analysis, exploratory factor analysis and confirmatory factor analysis are the four main methods used for assessing equivalence of psychological measures (Fischer & Fontaine, in press). Confirmatory factor analysis (CFA) is one of the most widely used methods to test for measurement invariance and is becoming increasingly popular, and is thus the method discussed in this paper. In brief, CFA is a model testing technique in which a theoretical model is compared with the observed structure in a sample. The individual parameters in CFA models are formally denoted by Greek letters, but this approach will not be used here to facilitate understanding.

CFA models are often graphically represented. Circles or ovals represent latent, unobserved variables, and squares represent the manifest, observed variables. Lines are then used to indicate the relationships between variables. Single-arrow lines pointing to a specific variable represent a hypothesized direct relationship, in which the variable with the arrow pointing to is the dependent variable. Double-arrow lines indicate that the variables are related to each other with no implied direction of the relationship. Figure 1 shows the illustrative model that we will use in this paper. As can be seen, there are two ovals, four squares and five lines linking them. The model indicates that there are four observed variablesð reading and writing in Grade 5 (i.e., READING5 and WRITING5) and reading and writing in Grade 7 (i.e., READING7 and WRITING7)ð , and two latent variablesð verbal ability in Grade 5 and Grade 7 (i.e., Verbal5 and Verbal7). The single-arrow lines from the latent to the observed variables indicate that READING5 and WRITING5 are indicators of the latent variable Verbal5, while READING7 and WRITING7 are indicators of the latent variable Verbal7. The double-arrow line indicates that the latent variables are related. Arrows without origin indicate proportions of error and unexplained variances for each of the four observed variables.

Figure 1. *Theoretical model used in the illustrative example*



It is worth noting that this example is for illustration purposes only. It is strongly advisable that latent factors are measured with at least three indicators (Byrne, 1998; MacCallum, 1995). If less than three indicators are used as in this example, the model is not properly identified (the model is identified in our example because we allow the latent factors to covary and do not estimate any further parameters). Identification issues are complex and are discussed in detail elsewhere (Bollen, 1989; Byrne, 1998; MacCallum, 1995).

3.1. Testing Measurement Invariance across Groups

CFA models are often run with single-sample data. For example, one could collect data from a community sample to test whether the items of a new developed scale comprise good indicators of a given latent construct. In order to assess measurement invariance, multi-group confirmatory factor analyses (MGCFA) are performed. In MGCFA, the theoretical model is compared with the observed structure in two or more samples. Jöreskogôs (1971; 1993) strategy for the assessment of the comparability of factor structures is typically followed to test measurement invariance. In his strategy, nested models are organized in a hierarchical ordering with decreasing numbers of parameters (or increasing degrees of freedom), which entails adding parameter constraints one at a time. These increasingly restrictive models are tested in terms of their fit of the data to the model (Cheung & Rensvold, 1999, 2002; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). Because each new model is nested in the previous model, measurement invariance models become increasingly more restrictive. MGCFA following this approach is widely accepted to be the most powerful and versatile approach for testing measurement invariance (Steenkamp & Baumgartner, 1998).

A typical sequence of models often tested are described and discussed below (for a detailed presentation see Marsh, 1994; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). Before discussing the models, it is important to note that scholars have proposed specific distinctions among the models. One distinction borrowed from model testing (Anderson & Gerbing, 1988) is between ómeasurement invarianceô models---(in a narrower sense) models that assess invariance of construct, factor loading, item intercepts and error variances --- and óstructural invarianceô models ---models that assess invariance of the variances, covariances and means of the latent variables (e.g., Byrne, Shavelson, & Muthén, 1989; Vandenberg & Lance, 2000). Note that the term óstructuralô invariance has a different meaning in this context compared to the cross-cultural literature on equivalence. As noted above, structural equivalence in the cross-cultural literature assess whether indicators are related to the construct in a non-trivial way (see, e.g., Fontaine, 2005). We follow the distinction between measurement and structural invariance to organise the specific models below. Measurement invariance needs to be tested for cross-group comparisons (especially for mean comparisons); structural invariance is optional and researchers need to decide whether the further restrictions are theoretical meaningful. We follow the general succession of tests proposed by Vandenberg and Lance (2000). They also provide a very good flowchart of the logical sequencing for assessing cross-group invariance. Figure 2 provides a visual aid to better understand the models discussed here.

Figure 2a. *Configural invariance (same structure across groups)*

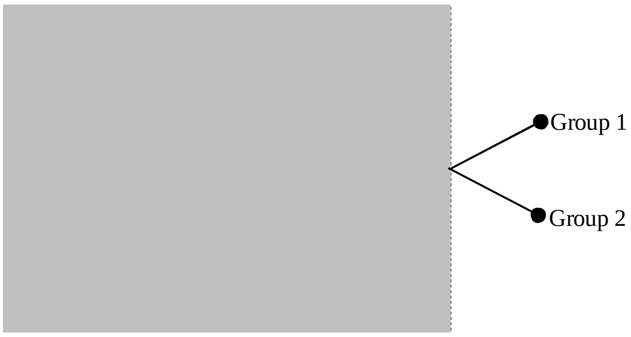


Figure 2b. *Metric invariance (same factor loadings across groups)*

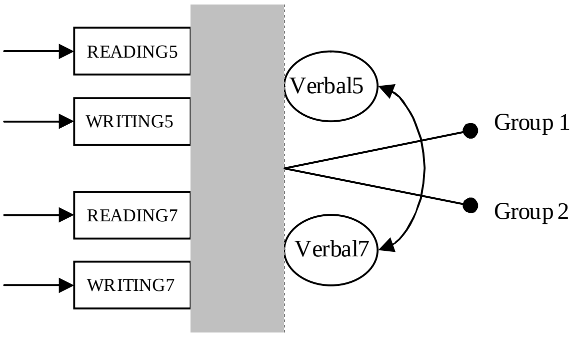


Figure 2c. *Scalar invariance (same item intercepts across groups)*

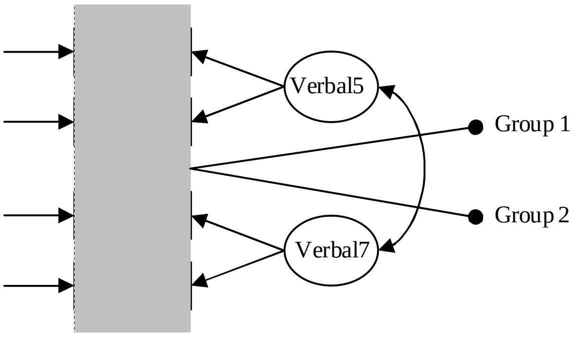


Figure 2d. *Error variance invariance (same error variance across groups)*

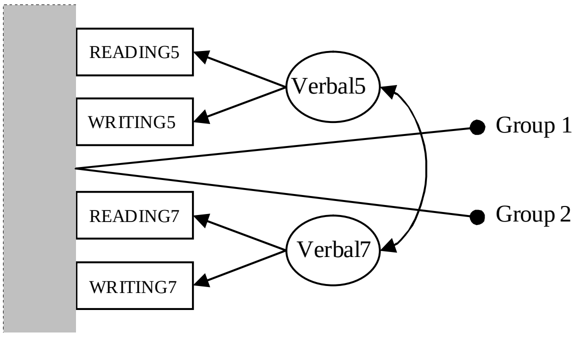


Figure 2e. *Factor variance invariance (same factor variance across groups)*

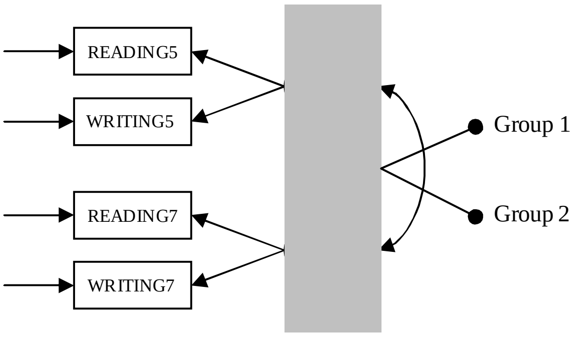


Figure 2f. *Factor covariance invariance (same factor covariance across groups)*

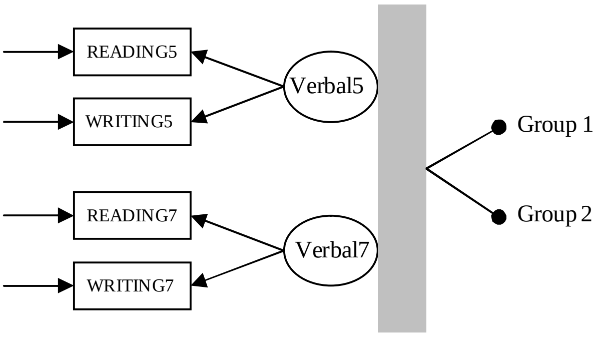
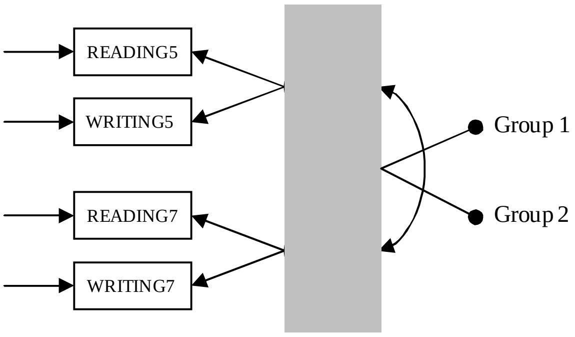


Figure 2g. *Factor mean invariance (same factor mean across groups)*



3.1.1. Tests of aspects of measurement invariance

Models that test relationships between measured variables and latent constructs are measurement invariance tests. There are four common models that fall in this category: configural, metric, scalar and error variance invariance.

Model 1 (Configural invariance). This model is the first step to establish measurement invariance, and is satisfied if the basic model structure is invariant across groups, indicating that participants from different groups conceptualize the constructs in the same way. Configural invariance can be tested by running individual CFAs in each group. However, even if the model fits well in each group, it is still necessary to run this step in MGCFA, since it serves as the comparison standard for subsequent tests (also known as the baseline model). This model is tested by constraining the factorial structure to be the same across groups (see Figure 2a).

Model 2 (Metric invariance). This model tests if different groups respond to the items in the same way; that is, if the strengths of the relations between specific scale items and their respective underlying construct are the same across groups. If metric invariance is satisfied, obtained ratings can be compared across groups and observed item differences will indicate group differences in the underlying latent construct. Research has suggested that at least partial metric invariance must be established before continuing in the sequence of tests (Vandenberg & Lance, 2000). This model is tested by constraining all factor loadings to be the same across groups (see Figure 2b).

Model 3 (Scalar invariance). Scalar, or intercept, invariance is required to compare (latent) means. Establishing scalar invariance indicates that observed scores are related to the latent scores; that is, individuals who have the same score on the latent construct would obtain the same score on the observed variable regardless of their group membership. This model is tested by constraining the intercepts of items to be the same across groups (see Figure 2c). This is the last model necessary to compared scores across groups. All additional tests are optional and may be theoretically meaningful in specific contexts.

Model 4 (Error variance invariance). To test if the same level of measurement error is present for each item between groups, all error variances are constrained to be equal across groups (see Figure 2d).

3.1.2. Tests of aspects of structural invariance

Models concerning only the latent variables are structural invariance tests. There are three common models that fell in this category: factor variance, factor covariance and factor mean invariance. These models are not necessarily nested, for example, it is possible to test for factor mean invariance (see Model 7 below) straight after testing Model 3 (intercept or scalar invariance).

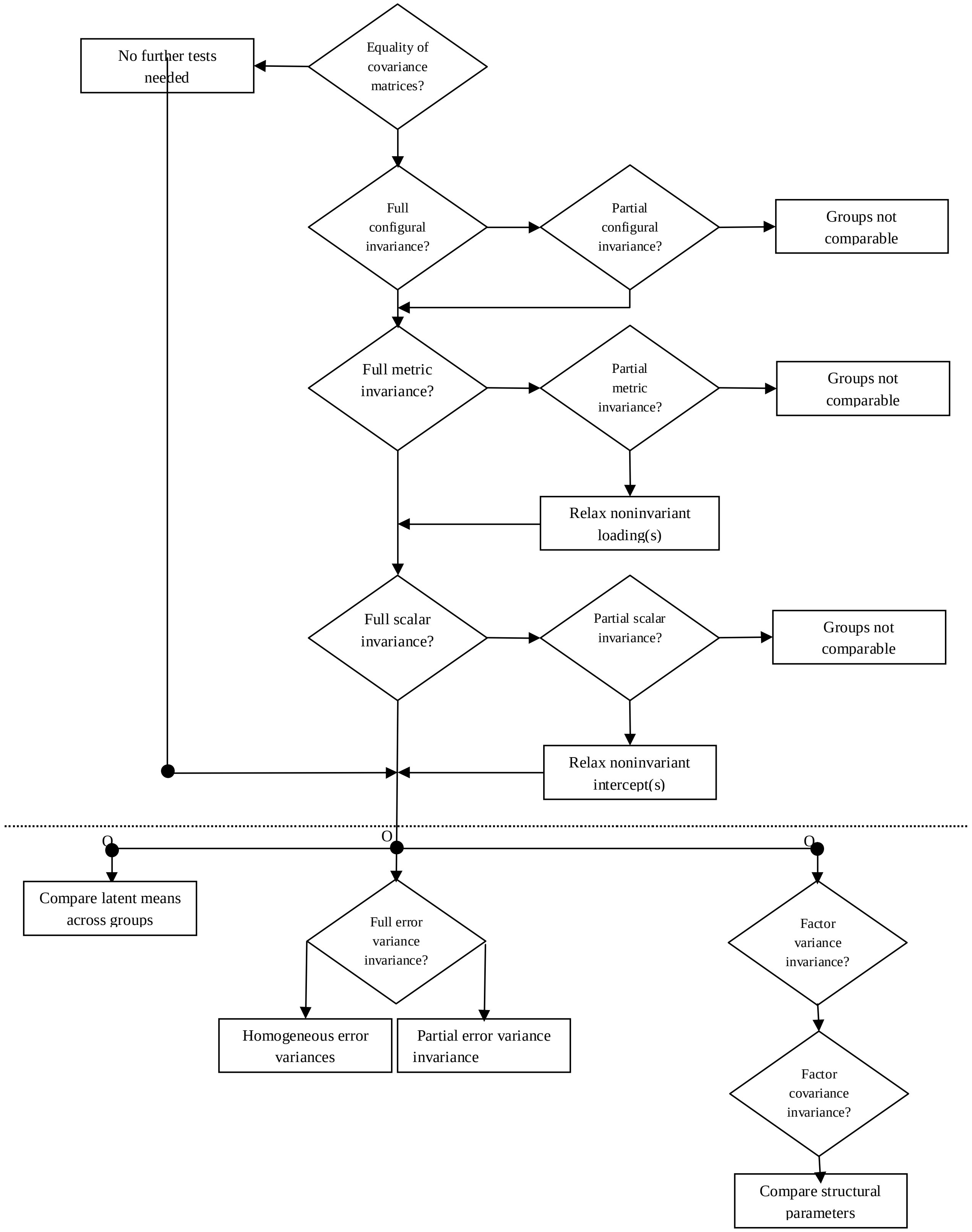
Model 5 (Factor variance invariance). Invariance of factor variance indicates that the range of scores on a latent factor do not vary across groups. This model is tested by constraining all factor variances to be the same across groups (see Figure 2e).

Model 6 (Factor covariance invariance). The stability of the factor relationships across groups is assessed in this model. The model thus implies that all latent variables have the same relationship in all groups. This model is tested by constraining all factor covariances to be the same across groups (see Figure 2f).

Model 7 (Factor mean invariance). Invariance of latent factor mean indicates that groups differ on the underlying construct(s). This model is tested by constraining the means to be the same across groups (see Figure 2g).

Figure 3 presents a flowchart of the logical sequencing of steps for assessing the degree of cross-group invariance adapted from Vandenberg and Lance (2000). As can be seen, only the measurement models are organized in a hierarchical ordering with increasing constraints from one model to the next. Each model is nested in the previous model, and measurement tests become increasingly restrictive. As a result, a model is only tested if the previous model in the hierarchical ordering has been shown to be equivalent across groups. Structural models, in contrast, are not hierarchical or sequential.

Figure 3. *Flowchart of the logical sequencing for assessing cross-group invariance (adapted from Vandenberg & Lance, 2000, p. 56)*


Equality of covariance matrices?
No further tests needed
Full configural invariance?
Partial configural invariance?
Groups not comparable
Full metric invariance?
Partial metric invariance?
Groups not comparable
Relax noninvariant loading(s)
Full scalar invariance?
Partial scalar invariance?
Groups not comparable
Relax noninvariant intercept(s)
O
O
O
Compare latent means across groups
Full error variance invariance?
Homogeneous error variances
Partial error variance invariance
Factor variance invariance?
Factor covariance invariance?
Compare structural parameters

3.2. Differences between full and partial invariance

The models discussed above address *full* measurement invariance because they assess whether each given element (i.e., factor loadings, item intercept, factor variance) is equal in all groups. However full measurement invariance is unlikely to hold in practice. To address the too strict and unrealistic goal that invariance restrictions must hold for all parameters across groups, Byrne et al. (1989) introduced to concept of *partial* measurement invariance, in which only a subset of parameters in a model is constrained to be invariant while another subset of parameters is allowed to vary across groups. Hence, partial measurement invariance may allow appropriate cross-group comparisons even if full measurement invariance is not obtained. Partial measurement invariance can be assessed in two cases: (1) when measures are invariant across some but not all groups, or (2) when some but not all of the parameters are invariant across groups (Vandenberg & Lance, 2000).

Given the lack of clear-cut criterion for using partial measurement invariance, Vandenberg and Lance (2000) have made some recommendations. They argue that configural invariance and (at least partial) metric invariance need to be established before testing any further partial invariance model. They further argue that partial metric invariance is permitted only if parameters relaxed to vary across groups are a minority of indicators (see also van de Vijver & Poortinga, 1982), and when there is strong theoretical and empirical (i.e., cross-validation evidence) bases. Milfont et al. (2006) provides a practical illustration of partial measurement invariance testing in cross-cultural research.

3.3. How to compare the models: Using goodness-of-fit indices

As discussed above, CFA is a model testing technique in which a theoretical model is compared with the observed structure in a sample. Goodness-of-fit indices are used to determine the degree to which the theoretical model as a whole is consistent with the empirical data. These indices thus indicate how well the empirical data ófitô the proposed theoretical model. LISREL (as well as other available software like Amos and EQS) provides several indices to assess how well an a priori hypothesized model fits the sample data. The likelihood ratio test (also called chi-square or G^2 test) is an objective model fit index, and has been traditionally used as a goodness-of-fit statistic in structural equation modeling. However, its sensitivity to sample size and its underlying assumption that the model fits the sample data perfectly has long been recognized as problematic (e.g., Bentler & Bonett, 1980; Browne & Cudeck, 1993). It has thus been recommended that this statistic should be used as a measure of fit rather than a test statistic (Jöreskog, 1993).

Several fit indices, or subjective model fit indices, have thus been developed to overcome limitations of the likelihood ratio test (for reviews, see Bentler & Bonett, 1980; Hu & Bentler, 1995; Kaplan, 2000; Mulaik et al., 1989). These fit indices can be categorized into absolute or incremental fit indices. While the former measure how well an a priori model reproduces the sample data, the latter assess improvement in fit by comparing a target model with a more constrained nested model (Hu & Bentler, 1999). Detailed considerations of these indices are beyond the scope of this article but are covered at length elsewhere (Bollen, 1989; Hu & Bentler, 1995; Mulaik et al., 1989). Numerous fit indices consider different aspects of fit, and it has been recommended that researchers should report multiple fit indices in structural equation modeling studies (Hu & Bentler, 1995; Thompson, 2000).

The absolute indices used here to evaluate overall model fit were: the normed chi-square, or the chi-square to degrees of freedom ratio (G^2/*df*) (Wheaton, Muthén, Alwin, & Summers, 1977), the root mean square error of approximation (RMSEA) (Steiger & Lind, 1980), and a standardized version of Jöreskog and Sörbomôs (1981) root mean square residual (SRMR). A G^2/*df* ratio of 3:1 or less indicates good fit (Carmines & McIver, 1981); RMSEA and SRMR values close to .06 and .08 respectively indicate acceptable fit (Hu & Bentler, 1999), and RMSEA values in the range of .08 to .10 indicate mediocre fit and above .10 indicate poor fit (Browne & Cudeck, 1993; MacCallum, Browne, & Sugawara, 1996). Furthermore, the difference in chi-square between two nested models (i.e., G^2 difference test), the comparative fit index (CFI) (Bentler, 1990). the expected cross-validation index (ECVI) (Browne & Cudeck, 1989), and a consistent version of Akaikeôs (1987) information criterion (CAIC) (Bozdogan, 1987) were used as incremental fit indices to calculate improvements over competing models. Significant results for the G^2 difference test indicate that the model with smaller G^2 has a statistically better fit. This test, however, has the same limitations as the overall likelihood ratio test so that with large samples very trivial differences yield a significant test result. Therefore, the G^2 difference test was used only as indicative of significant improvements. CFI values close to .95 indicate acceptable fit (Hu & Bentler, 1999). Lower ECVI and CAIC values reflect the model with the better fit (Garson, 2003). In addition, 90% confidence intervals (90%CI) were also reported for both RMSEA and ECVI, following MacCallum et al.ôs (1996) guidelines.

These absolute and incremental fit indices are often used to compare an unconstrained model with one having measurement invariance constraints. In addition, Cheung and Rensvold (2002) has suggested three specific incremental indices for testing measurement invariance. These indices are based on the differences in Bentlerôs (1990) comparative fit index (CFI), Steigerôs (1989) gamma hat (GH), and McDonaldôs (1989) non-centrality index (NCI) that are obtained when comparing nested models. If, in the sequence of the invariance tests, two nested models show a decrease in the value of CFI, GH and NCI greater than or equal to .01, .01, and .02 in magnitude, respectively, the more restrictive model should be rejected (Cheung, 2005; Cheung & Rensvold, 2002). LISREL output does not give the GH or NCI, but these indices can be easily calculated using a free online program (Pirritano, 2005).

4. Testing Measurement Invariance across Groups: A practical example

Given the pedagogical focus of this paper on measurement invariance, we selected a publicly available and simple model for illustrative analysis. The data used to provide an illustration of measurement invariance testing are taken from the LISREL manual (Jöreskog & Sörbom, 1999, p. 62), which is available to all users. The data is based on scores on the ETS Sequential Test of Educational Progress for two groups of boys who took the test in both Grade 5 and Grade 7. The groups were created based on whether or not the boys were in the academic curriculum in Grade 12. A sample of 373 formed the ôboys academicô group (i.e., BA) and a sample of 249 formed the ôboys non-academicô group (i.e., BNA). The model is the same as depicted in Figure 1. READING5 and WRITING5 are indicators of the latent variable Verbal5 that represents boysô verbal ability at Grade 5, and READING7 and WRITING7 are indicators of the latent variable Verbal7 that represents boysô verbal ability at Grade 7. Although this is a longitudinal rather than a cross-cultural data, the same steps described below can be used in cross-cultural research. And although this example is based on only two groups, the process of measurement invariance testing is similar if more than two groups are compared. If the number of comparisons becomes larger, different approaches become useful (Fontaine & Fischer, in press; Selig, Card, & Little, 2008).

The LISREL syntaxes for each of the model tested are given in Appendix A. There are specific publications available that describe the specific steps in more detail (e.g., Byrne, 1998; Jöreskog & Sörbom, 1999). To make it easier to actually run the measurement invariance syntaxes used in our example, we have included details explaining what we have done for each step in the actual syntaxes. Anyone who has the software can copy the syntaxes we provide and run them easily in LISREL.

Before comparing the groups, it is important to make sure that the hypothetical structure provides good fit for both groups. Thus, the first step is to test whether the proposed two-factor model fits the empirical data from each group. Results show excellent model fit for the BA group (G^2 = .86; *df* = 1; G^2/*df* = .86; RMSEA = .00, 90%CI = .00-.13; SRMR = .003; CFI = 1.00) as well as for the BNA group (G^2 = .66; *df* = 1; G^2/*df* = .66; RMSEA = .00, 90%CI = .00-.16; SRMR = .004; CFI = 1.00), indicating that the two-factor model of verbal ability is supported in both groups.

The second step is to move from single-group CFA to MGCFA in order to cross-validate the two-factor model across the two groups. Model 1 tested whether the proposed structure (Figure 1) would be equal across the two groups. As excellent fit of the two-factor structure had been established independently for each group earlier, one could expect that configural invariance would be supported. The fit indexes confirmed this. As can bee seen in Table 1, Model 1 provided excellent fits to the data, indicating that the factorial structure of the construct is equal across groups. Note that the G^2 value reported for this model is simply the sum of the G^2 values for the model tested separately for each sample, as reported above.

As the configural invariance was supported, the factor pattern coefficients were then constrained to be equal to test for metric invariance. Model 2 had good fit indices (e.g., G^2/*df* < 3; RMSEA < .08; CFI > .95;), but the chi-square test was significant, indicating that the imposition of constraints (equal factor loadings across groups) resulted in statistically significant decreases in the fit of Model 2 compared to Model 1. As signed earlier, however, this test has limitations. Considering the other comparative fit indices (e.g., æCFI, æGH and æNCI), the overall results indicate the viability of constraining the factor loading to be the same across the groups.

The scalar invariance model (Model 3) and error variance invariance model (Model 4) also provide excellent fits to the data. As can be seen in Table 1, the overall goodness-of-fit indices and the tests of differences in fit between adjacent models (Model 3 vs. Model 2, and Model 4 vs. Model 3) support measurement invariance. Support for scalar invariance indicates that the latent means can be meaningfully compared across groups. Support for error variance invariance indicates that the four observed variables are invariant across groups, having no measurement bias.

Table 1. *Fit indices for invariance tests*

Model	G^2 (*df*)	G^2/*df*	RMSEA (90%CI)	SRMR	æ G^2 (æ*df*)	CFI (æCFI)	GH (æGH)	NCI (æNCI)	ECVI (90%CI)	CAIC	Comparison	Decision
Model 1: Full configural invariance	1.52 (2)	.76	.00 (.00-.10)	.004	ï ï	1.00 (ï ï)	1.00 (ï ï)	1.00 (ï ï)	.061 (.00-.10)	135.31	ï ï	Accept
Model 2: Full metric invariance	8.65 (4)	2.16	.061 (.00-.12)	.044	7.13* (2)	1.00 (.00)	.9963 (.0037)	9963 (.0037)	.066 (.058-.086)	127.57	Model 1 vs. Model 2	Accept
Model 3: Full scalar invariance	9.96 (6)	1.66	.046 (.00-.095)	.042	1.31 (2)	1.00 (.00)	.9968 (ï .0005)	9968 (ï .0005)	.087 (.068-.095)	173.48	Model 2 vs. Model 3	Accept
Model 4: Full error variance invariance	22.07 (10)	2.21	.062 (.026-.098)	.060	12.11* (4)	.99 (.01)	.9904 (.0064)	9903 (.0065)	.094 (.065-.11)	155.86	Model 3 vs. Model 4	Accept
Model 5: Full factor variance invariance	30.39 (12)	2.53	.070 (.040-.10)	.17	8.32* (2)	.99 (.00)	.9854 (.0050)	9853 (.0050)	.10 (.067-.12)	149.32	Model 4 vs. Model 5	Accept
Model 6: Full factor covariance invariance	37.33 (13)	2.87	.078 (.049-.11)	.18	6.64** (1)	.99 (.00)	.9866 (ï .0012)	9865 (ï .0012)	.11 (.072-.13)	48.82	Model 5 vs. Model 6	Accept
Model 7: Full factor mean invariance	172.58 (15)	11.51	.18 (.16-.21)	.070	135.25 (2)	.90 (.09)	.8874 (.0992)	8808 .1057)	.32 (.25-.38)	269.21	Model 6 vs. Model 7	Reject

The analyses above support the measurement invariance of the two-factor model across the two groups. Analyses were then performed to assess the structural invariance. It is worth noting again that structural models are not hierarchical, so that the model testing sequence is arbitrary. However, we used a sequence often used in the literature (Vandenberg & Lance, 2000). The first model tested the invariance of the factors variance (Model 5). As can be seen in Table 1, constraining the factors variance to be equal across the two groups substantially increased the SRMR. Overall, however, these constraints did not significantly worsen model fit as compared to Model 4. The equality of factors covariance was then tested (Model 6). The goodness-of-fit indices and the tests of differences in fit between Models 6 and 5 support the invariance of the covariance between the latent variables. This indicates that the covariance between boyôs verbal ability at Grade 5 and Grade 7 is equal across groups. Support for Model 5 and 6 respectively indicates that factor correlations and factor covariances are identical across groups.

The final test of structural invariance was to constraint the latent factor means to be equal across groups. By all standards, Model 7 provided a very poor fit to the data, rejecting the imposition of factor mean invariance. This indicates that the latent factor means differ across the two groups. Close inspection of the output from Model 3 indicates that the latent means for the non-academic group is below the mean of the academic group in both grades, which is expected (Jöreskog & Sörbom, 1999, p. 70).

CONCLUSIONS

This paper provides an overview of the issue of invariance in measurement. We used a publicly available two-sample data as a case study in an effort to illustrate this issue. Using the explanations we provide as well as the figures and LISREL syntaxes, the reader should be able to understand and conduct his own analyses to evaluate measurement invariance. When comparing groups an assumption is made that the instrument (e.g., questionnaires, tests) measures the same psychological construct in all groups. If this assumption holds, the comparisons are valid and differences/similarities between groups can be meaningfully interpreted. However, if this assumption does not hold comparisons and interpretations may not be meaningful. Researchers should explicitly evaluate measurement invariance and then distinguish between different levels of similarity or equivalence.

REFERENCES

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52,* 317-332.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*, 411-423.

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

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

Berry, J. W., Poortinga, Y. H., Segall, M. H., & Dasen, P. R. (2002). *Cross-cultural psychology: Research and applications* (2nd ed.). Cambridge: Cambridge University Press.

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

Bozdogan, H. (1987). Model selection and Akaike's information criteria (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24*, 445-455.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum.

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

Carmines, E. G., & McIver, J. D. (1981). Analyzing models with unobserved variables: Analysis of covariance structures. In G. W. Bohinstedt & E. F. Borgatta (Eds.), *Social measurement: Current issues* (pp. 65-115). Beverly Hills, CA: Sage.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*, 1005-1018.

Cheung, G. W. (2005). Cheung & Rensvold (2002) critical values. Retrieved March 15, 2006, from the Structural Equation Modeling Discussion Group at http://bama.ua.edu/cgi-bin/wa?A2=ind0506&L=semnet&P=R45491&D=0&I=1&T=0.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*, 1-27.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology, 31*, 188-213.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255.

Fischer, R., Ferreira, M. C., Assmar, E., Redford, P., Harb, C., Glazer, S., et al. (2009). Individualism-collectivism as descriptive norms: Development of a subjective norm approach to culture measurement. *Journal of Cross-Cultural Psychology, 40*, 187-213.

Fischer, R., & Fontaine, J. R. J. (in press). Methods for investigating structural equivalence In D. Matsumoto & F. J. R. Van de Vijver (Eds.), *Cross-cultural research methods*. New York: Oxford University Press.

Fontaine, J. R. J. (2005). Equivalence. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 803-813). San Diego: Academic Press.

Fontaine, J. R. J., & Fischer, R. (in press). Multilevel structural equivalence. In D. Matsumoto & F. J. R. van de Vijver (Eds.), *Handbook of cross-cultural research methods*. Oxford: Oxford University Press.

Garson, G. D. (2003). PA 765 Statnotes: An online textbook. Retrieved February 07, 2004, from http://www2.chass.ncsu.edu/garson/pa765/statnote.htm.

Gouveia, V. V., Milfont, T. L., Fonseca, P. N., & Coelho, J. A. P. M. (2009). Life satisfaction in Brazil: Testing the psychometric properties of the satisfaction with life scale (SWLS) in five Brazilian samples. *Social Indicators Research, 90*, 267-277.

Hu, L.-T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76-99). Thousand Oaks, CA: Sage.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409-426.

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294-316). Newbury Park, CA: Sage.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by the method of maximum likelihood*. Chicago: National Educational Resources.

Jöreskog, K. G., & Sörbom, D. (1999). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Lincolnwood, IL: Scientific Software International.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, CA: Sage.

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

MacCallum, R. C. (1995). Model specification: Procedures, strategies, and related issues. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 16-36). Thousand Oaks, CA: Sage.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149.

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

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

Milfont, T. L., Duckitt, J., & Cameron, L. D. (2006). A cross-cultural study of environmental motive concerns and their implications for proenvironmental behavior. *Environment and Behavior, 38*, 745-767.

Milfont, T. L., Duckitt, J., & Wagner, C. (in press). The higher order structure of environmental attitudes: A cross-cultural examination. *Interamerican Journal of Psychology*.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105*, 430-445.

Pirritano, M. (2005). MS Excel formulas for computing McDonald's non-centrality index and Steiger's gamma hat. Retrieved Retrieved March 15, 2006, from http://www.unm.edu/~mpirrita/SEM_Fit_Indexes/Fit_Indexes.xls.

Selig, J. P., Card, N. A., & Little, T. D. (2008). Latent variable structural equation modeling in cross-cultural research: Multigroup and multilevel approaches. In F. J. R. Van de Vijver, D. A. Van Hemert & Y. H. Poortinga (Eds.), *Individuals and cultures in multilevel analysis* (pp. 93-120). Mahwah, NJ: Erlbaum.

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

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

Steiger, J. H., & Lind, J. C. (1980, June). *Statistically based tests for the number of common factors.* Paper presented at the Psychometric Society Annual Meeting, Iowa City, IA.

Thompson, B. (2000). Ten commandments of structural equation modeling. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 261-284). Washington, DC: American Psychological Association.

van de Vijver, F. J. R., & Fischer, R. (2009). Improving methodological robustness in cross-cultural organizational research. In R. S. Bhagat & R. M. Steers (Eds.), *Handbook of culture, organizations, and work* (pp. 491-517). Cambridge: Cambridge University Press.

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage.

van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. *Journal of Cross-Cultural Psychology, 31*, 33-51.

van de Vijver, F. J. R., & Poortinga, Y. H. (1982). Cross-cultural generalization and universality. *Journal of Cross-Cultural Psychology, 13*, 387-408.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70.

Wheaton, B., Muthén, B., Alwin, D. F., & Summers, G. F. (1977). Assessing the reliability and stability in panel models. In D. R. Heise (Ed.), *Sociological methodology 1977* (pp. 84-136). San Francisco: Jossey-Bass.