Unclassified **EDU/WKP(2014)2**

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development **17-Jun-2014**

English - Or. English

DIRECTORATE FOR EDUCATION AND SKILLS

EDU/WKP(2014)2
Unclassified

Cancels & replaces the same document of 12 June 2014

Evaluating Measurement Invariance of TALIS 2013 Complex Scales

A Comparison between Continuous and Categorical Multiple-Group Confirmatory Factor Analyses

This paper has been prepared by:

- Deana Desa, Research and Analysis Unit, IEA Data Processing and Research Center

Julie Belanger
Julie.BELANGER@oecd.org

JT03359510

Complete document available on OLIS in its original format

English - Or. English

OECD EDUCATION WORKING PAPERS SERIES

Comment on the series is welcome, and should be sent to edu.contact@oecd.org.

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

---

www.oecd.org/edu/workingpapers

EVALUATING MEASUREMENT INVARIANCE OF TALIS 2013 COMPLEX SCALES

A COMPARISON BETWEEN CONTINUOUS AND CATEGORICAL MULTIPLE-GROUP CONFIRMATORY FACTOR ANALYSES

Deana Desa

Research and Analysis Unit, IEA Data Processing and Research Center[1]

Email of the correspondence author: deana.desa@iea-dpc.de

[1] The author would like to acknowledge Eugene Gonzalez, Steffen Knoll, Plamen Mirazchiyski, Andres Sandoval-Hernandez, Agnes Stancel-Piatak and the other members of the Research and Analysis Unit (RandA) at the IEA Data Processing and Research Center for the work and preparation of this study. The author would like to thank Ralph Carstens and Mark Cockle from the IEA DPC for helpful comments on previous versions of this paper. This paper was developed as part of the presentation for the 2013 TALIS Analysis Expert Group Meeting in Hamburg, Germany.

ABSTRACT

This paper evaluates measurement invariance of complex scales from a social survey using both a continuous approach and a categorical approach to help inform future decisions in choosing the most appropriate methods to perform the validation of complex scales. In particular, continuous and categorical approaches are compared for constructing and validating 11 complex scales across 23 countries participating in the first round of the OECD Teaching and Learning International Survey (TALIS). Two invariance testing approaches were compared – 1) continuous multiple-group confirmatory factor analysis; 2) categorical multiple-group confirmatory factor analysis. Latent variable modelling was employed to account for the complex structure of the relationships between many items in each scale. The performance of the models is reported and illustrated based on the evaluation of the level of measurement invariance. All of the scales established configural and metric levels of invariance from both approaches, and three scales established scalar invariance from the categorical approach, allowing for a meaningful mean score comparison across countries. Limitations of the models compared in this study and future considerations for construction and validation of scaling complex scales are discussed.

Contents

INTRODUCTION

Large scale surveys are often intended for population comparisons such as in cross-country or cross-cultural comparisons on constructs measured by contextual scales (e.g., attitudes, beliefs, values, behaviours, and socio-demographic characteristics). Large scale surveys such as the OECD Teaching and Learning International Survey (TALIS) require careful translations or adaptions of the questionnaire items into different education systems and different languages, need to have consistent response rates with comparable complex sampling plans, and need to use comparable country-by-country data collection techniques. These represent several challenges in international surveys, particularly in dealing with the unique cultures and dissimilarities of many different countries.

The same contextual questionnaire items are employed across a large number of countries where responses from all of these items are used for the construction and validation of complex scales. These steps are undertaken to ensure that the questionnaire items answered by the survey respondents are *sample-free* (Wright, 1968), that the set of questionnaire items answered by the survey respondents in one country behave in a similar manner when answered by other respondents in different countries. The questionnaire items with consistently different results according to independent grouping conditions (i.e. respondents belonging to different countries) are said to have different measurement properties (e.g., reliability or other item statistics such as item-total correlation), and thus may show systematic inaccuracy in the observed item responses. Items with unusually low item statistics can be, as necessary, improved by examining any flaws or inaccuracies (e.g., scoring options, rewording the text of the items and different categories in the response options) or may be discarded from further analysis by deeming them to show insufficient evidence of having good measurement properties.

The procedures carried out before the scales are constructed, usually using latent variable modellings—from framing the questionnaire items based on their underlying theoretical background, to a standard data collection—are necessary steps, but not sufficient to guarantee the requirement of measurement properties of the measured items (i.e. observed variables/items) describing complex scales. A “good” scale preserves the meaning and should function in the same way across countries, that is, the scale measures what it purports to measure regardless of whom we choose to measure with them. In this case, therefore, the measurement properties of the scale remain unchanged when the survey respondents from different countries respond to the items, allowing the scale to be compared across countries. Any difference can then be attributed to the way cultures or beliefs affect the scales (e.g., different socioeconomic background).

The key concern when making such a large number of country comparisons is to ensure that the measurement of the latent constructs of the scales is invariant cross-nationally. That is, meaningful cross-national comparisons of the scales require that the questionnaire items used to operationalise or describe the underlying latent construct are measurement invariant across countries. Comparisons of simple statistics for individual observed items (e.g., individual variable mean, percentage, and percentile) per se do not require such a restriction. However, comparisons of the means of scales

combining several observed items require different levels of measurement invariance (levels of invariance are described later). Lower levels of invariance allow for comparisons of correlations between scales, while higher levels of invariance are necessary for comparisons of means of the constructed scales across groups. Measurement invariance testing is applied within latent modellings framework using either structural equation modelling (SEM) or item response theory (IRT). The application of IRT is widely used for surveys or tests with binary and polytomous observed variables (i.e., categorical data) whereas there is very little empirical validation of the benefits of the categorical over the continuous modelling using the confirmatory factor analysis within SEM. Both CFA and IRT have different utility for representing observed variables and testing theoretical hypotheses with one general goal that they are used to predict or estimate psychological and mental constructs based on the observed information about the respondents' responses and the characteristics of the questionnaires or test items.

The focus of this paper is to examine further, and to provide empirical findings, on the performance of measurement invariance for complex scales developed in TALIS when different methodological approaches are used. The findings are meant to help inform future discussions about the most appropriate ways to perform the validation of scales in surveys such as TALIS.

Definition of Contextual Complex Scale as Latent Construct

It is not uncommon to have tens or hundreds of questionnaire items in educational and psychological surveys. The basic advantage of developing scales from these observed variables is that each set of variables covers the different characteristics of the items, and that these can be used to understand the variability that exists between them. This set of variables is allocated to factors called latent constructs. A complex scale is defined as a latent construct simply because it does not have a perfect measure of the construct, but several questionnaire items may have answers that attempt or intend to measure it. It is dubious that a single item might capture or successfully cover the full meaning of the elicited complex behaviour or attitude. Latent constructs are therefore defined as psychological or mental constructs using advanced statistical methodologies such as confirmatory factor analysis (CFA, Jöreskog, 1969; Long, 1983) and item response theory (IRT, Lord and Novick, 1968; Lord, 1980). The former is focused on and used in this study to allow for a comparison between two different approaches within the same CFA framework. The description of a CFA model is introduced here.

Figure 1 shows examples of questionnaire items answered by teachers in many countries (e.g. 23 countries in TALIS 2008) where responses to these items are categorised from strongly disagree to strongly agree. The items are used to describe a scale defined as a latent construct, namely *classroom disciplinary climate* (CCLIMATE).

Figure 1. Questionnaire items for classroom disciplinary climate in TALIS 2008

How strongly do you agree or disagree with the following statements about this <target class>?

Please mark one choice in each row.

		Strongly Disagree	Disagree	Agree	Strongly Agree
a)	When the lesson begins, I have to wait quite a long time for students to <quieten down>.	☐1	☐2	☐3	☐4
b)	Students in this class take care to create a pleasant learning atmosphere.	☐1	☐2	☐3	☐4
c)	I lose quite a lot of time because of students interrupting the lesson.	☐1	☐2	☐3	☐4
d)	There is much noise in this classroom.	☐1	☐2	☐3	☐4

The latent construct of classroom disciplinary climate is represented by a latent regression model using a confirmatory factor analysis (CFA) approach. The CFA model for the scale is a combination of structural and measurement models, that is written as

$$\boldsymbol{X} = \boldsymbol{\tau_y} + \Lambda_y \mathbf{CCLIMATE} + \boldsymbol{\varepsilon} \ .$$

The observed responses, $\boldsymbol{X} = (x_1, x_2, x_3, x_4)$, for the four items describing the underlying scale are predicted in the structural model above based on the strength of the association between the observed variables and the targeted latent scale. These item-factor relationships are generally called factor loadings denoted as $\boldsymbol{\Lambda_y} = (\lambda_1, \lambda_2, \ldots, \lambda_p)$. The standardised loading is interpreted in a similar way to the standardised regression coefficient that is within a standard deviation unit increase or decrease between the observed responses and the estimated scale scores. The estimated means of responses of the variables are called the intercepts and denoted as $\boldsymbol{\tau_y} = (\tau_1, \tau_2, \tau_3, \tau_4)$, and the estimates are computed when the scale score is located at zero on its linear continuum. Measurement errors influence the observed responses (e.g., random distractions to respondent or unwanted systematic errors). From the latent regression model above, the term is expressed in the residual for each observed variable as $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)$. The measurement model of CFA model above has three components specified from its loadings, intercepts and residual variances. The structural model of CFA model is then derived for the latent construct (i.e. CCLIMATE) that is described by the mean and variance components of the construct.

For the continuous distributed observed responses, a means and covariances structure (MACS) is employed to compute the scale such that the modelling complies with the assumption of multivariate normally distributed observed responses (Little, 1997; Sörbom, 1974). For the categorical observed

responses (i.e., ordered categorical data), there is one additional parameter for the measurement model that is called the latent thresholds parameter denoted as $\boldsymbol{\vartheta}_1 = (\vartheta_{11}, \vartheta_{21}, \vartheta_{31})$, $\boldsymbol{\vartheta}_2 = (\vartheta_{12}, \vartheta_{22}, \vartheta_{32})$, $\boldsymbol{\vartheta}_3 = (\vartheta_{13}, \vartheta_{23}, \vartheta_{33})$ and $\boldsymbol{\vartheta}_4 = (\vartheta_{14}, \vartheta_{24}, \vartheta_{34})$. The thresholds parameter represents the expected degrees for endorsement in each item from one response category to another category and therefore it is used explicitly for items with categorical responses. This categorical modelling is a generalisation of a two-category response model with one threshold. Note that there are three thresholds in a four-category item. The first threshold represents the expected value at which a survey respondent would be most likely to choose Disagree instead of Strongly Disagree. The second threshold represents the expected value at which a survey respondent would be mostly likely to choose Agree instead of Disagree, and the third threshold in the expected value of choosing Strongly Agree rather than Agree. Whenever categorical modelling is applied it is not only estimating the same parameters as in the continuous modelling but it is also estimating the probability of endorsement of response categories through its thresholds parameter.

These measurement model parameters of the CCLIMATE scale (loadings, intercepts, error variances and thresholds), for the continuous or categorical modelling, are used in testing the measurement invariance of the scale when the scale is defined to measure the same construct across different countries. Figure 2 shows simple illustrations of (a) the CFA model and (b) a regression plot for one country (model and plot for all other compared countries followed the same structure with the assumption of measurement invariance). The arrows in Figure 2(a) represent the relationships between the observed variables for Item 1 (I1) to Item 4 (I4) with the latent construct of CCLIMATE, and the linear dotted line in Figure 2(b) displays the regression line predicting observed item responses (*y-axis*) on estimated latent construct CCLIMATE (*x-axis*). The loading for each observed variable is computed using any two points on the regression line as the ratio between the differences on the y-axis and the differences on the x-axis, that is, $loadings = a/b$. This gives the slope of the regression line. The mean of the observed responses is the point where the regression line crosses the y-axis that is when CCLIMATE located at zero on the x-axis. This point is called the intercept of the regression line. The thresholds parameter is illustrated in Figure 2(c), where each of the ϑ_1, ϑ_2 and ϑ_3 estimate is inferred as transitional distribution from response categories 1 to 2, 2 to 3 and 3 to 4 (i.e, $\boldsymbol{x}^*$), respectively. The thresholds are estimated for every one variable of CCLIMATE. Thresholds parameter is also interpreted as the expected change of the locations on the latent construct continuum (on the y-axis) influenced by the strength of the item-factor association (i.e., $threshold = loading \times location$). The size of the thresholds is not necessarily a constant distance but increase across response categories.

In the continuous or categorical CFA modelling, CCLIMATE represents the underlying latent construct of classroom disciplinary climate where the estimated scale scores from the models are continuous and normally distributed. The presence of nuisance factors (e.g. gender and languages) can be regarded as group differences or differences in the CFA modelling of the scale where they do not mainly influence the scores of the scale but do affect the observed responses. Two steps are considered for the modelling of the CCLIMATE scale. In the first step, CFA models are evaluated in each homogeneous group (in this case a CFA within each country), where there are different estimates of item parameters (i.e., loadings, intercepts and thresholds) per country. For the second step, the

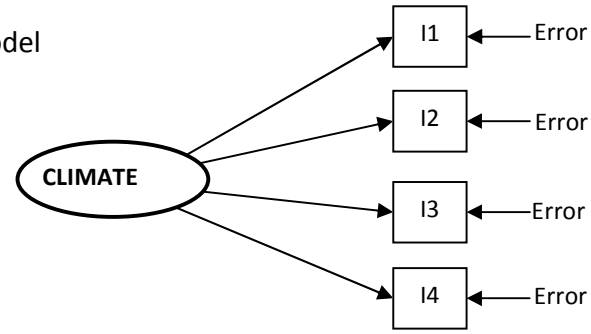
parameters are now restricted or imposed to have measurement invariance. It involves a multiple-group comparison of the same hypothesised structural model of classroom disciplinary climate, and evaluates at different levels of invariance where all countries are simultaneously analysed using the multiple-group CFA model. Any country differences are due to their differences in common scales means, variances, and correlations.

It is worth noting that the categorical responses of ordinal-scaled data, for example disagreement (coded 1 and 2) and agreement (coded 3 and 4) categories in the CCLIMATE scale or the frequency categories (e.g. coded as 0 for never to 6 for every day) for other scales, are often treated as if they were continuous and analysed assuming a normal distribution. Results from research practice have shown that an analysis of ordered-categorical data with this assumption is problematic due to distortion of the factor structure for different groups (Lubke and Muthén, 2004). A substantial number of studies has focused on applying CFA and MGCFA models without assuming normality but rather non-continuous distribution of the observed data (i.e., categorical ordered-scaled).

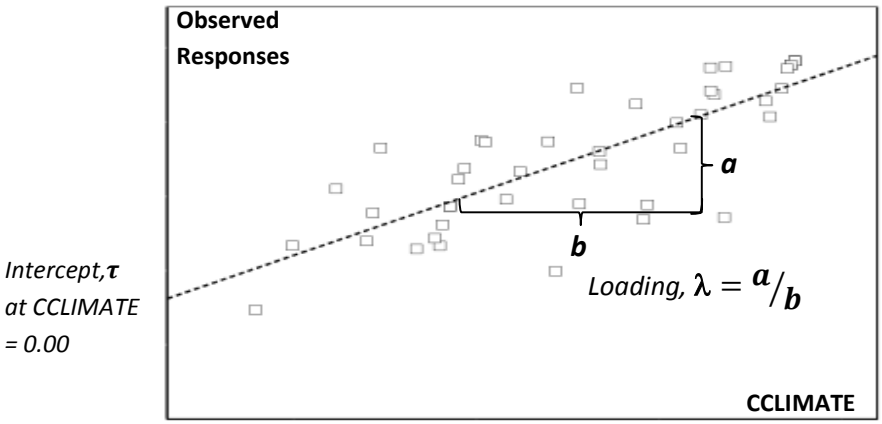
Cross-country similarities or differences on the contextual scale ought to be meaningfully compared when the requirement for measurement invariance is met. Failing to establish the invariance of the questionnaire items limits the interpretation of classroom disciplinary climate or other complex scale and most likely would lead to erroneous conclusions about the constructs. In other words, for this example, it is important to consider whether teachers from different countries or economies read or interpret each statement of the classroom disciplinary climate items similarly. When continuous or categorical CFA and MGCFA models are used the relationship of the items on the underlying structural latent construct as well as the meaning of the scale defined from the items are evaluated and it is expected that approximate comparable correlations of the items and comparable scale score means are revealed across the different populations. These measurement properties can be tested and evaluated using further statistical methods explained in the following section of this paper.

Figure 2. Illustrations of a CFA Model

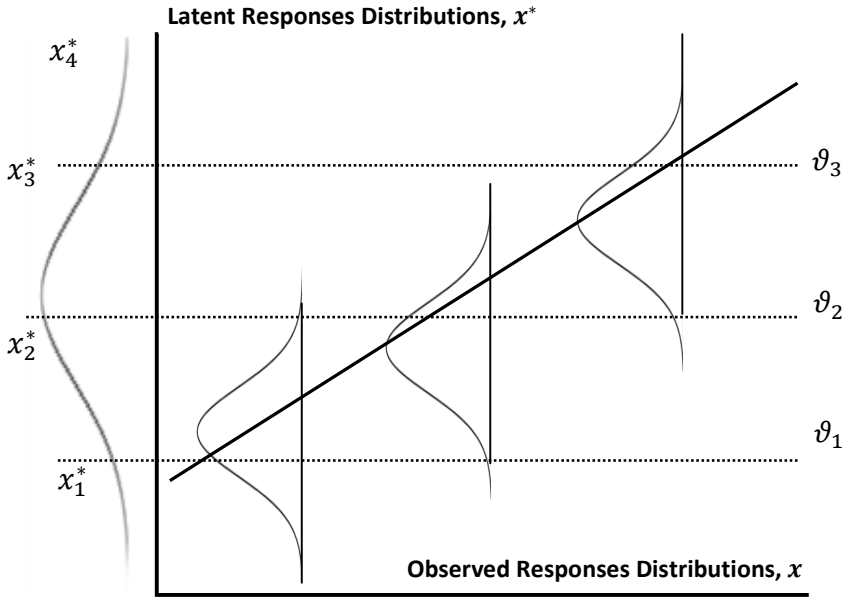
a. A CFA model



b. Latent response variable regression



c. Thresholds parameter for one variable



Continuous and Categorical Confirmatory Factor Analysis

Large scale assessment data normally consist of a mix of continuous variables and categorical variables. A common practice for psychological and educational contextual complex scales construction and measurement invariance testing is to use linear confirmatory factor (CFA) analysis and linear multiple-group confirmatory factor analysis (MGCFA, see Little and Slegers, 2005; Meredith, 1993; Sörbom, 1974) of the means and covariances of the observed questionnaire items. This tests measurement invariance assuming that continuous observed variables follow a multivariate normal distribution. CFA assuming a multi-normal distribution of the observed variables has been shown to be sufficiently robust against violations of the assumption of continuous observed variables if the categorisation (e.g., Likert-scale items, constructed responses, ordered categorical responses) is based on at least five response categories or if the data show a lack of normal distribution of the item responses (e.g., Babaku et al. 1987; Curran et al. 1996; Muthén and Kaplan 1985; DiStefano 2002).

CFA and MGCFA are usually estimated using the maximum likelihood (ML) estimator. It is well suited for CFA and MGCFA modelling when the observed variables are continuously distributed. Although this estimator works well, its performance based on simulation studies showed that the estimation of the model parameters can result in problems such as inflated Type I error rate, biased results and low reliability when observed variables are explicitly nominal (e.g. yes or no and true or false) or ordered categorical (e.g., “strongly disagree”, “disagree”, “agree”, and “strongly agree” or “never”, “seldom”, “quite often” and “very often”) but are treated as continuous (Lubke and Muthén, 2004; Millsap and Yun-Tein, 2004). A study by Lubke and Muthén (2004) has cautioned against applying the continuous approach for ordinal categorical response variables. The study observed that ML estimation on these different thresholds (i.e., response categories) among the indicators (i.e., observed variables) leads to heavily inflated chi-square values and low parameter coverage for the factor loadings. This means that the modelling fails to retain model-data agreement and produces more estimation errors. Therefore a robust estimator to non-normality of the observed distributions from maximum likelihood (MLR) is recommended for the continuous variables modelling and a robust weighted least-squares estimation of mean- and variance-adjusted (WLSMV) is used for the categorical variables modelling where both MLR and WLSMV produce stable estimates with robust standard errors and adjusted chi-square statistics (Muthén and Asparouhov, 2002; Millsap and Yun-Tein, 2004; Muthén et al., 1997).

Modelling observed categorical variables through WLSMV improves the agreement between the CFA and MGCFA models with the categorical observed variables and it preserves the factor structure across populations over the traditional ML estimation. MLR does not estimate the thresholds parameter and it is well suited for the categorical observed variable using the continuous CFA and MGCFA models where only the loadings and intercepts are estimated, whereas WLSMV estimates loadings, intercepts and thresholds in the categorical CFA and MGCFA models. An alternative estimator for the categorical modelling is called maximum likelihood parameter estimates with standard errors and a mean- and variance-adjusted (MLMV) that is also robust to non-normality. MLMV does not allow for analysis with complex sampling design limiting its use for this present study.

Measurement Invariance for Cross-Country Comparisons

Configural, metric, scalar, and strict invariance are the most commonly tested levels of measurement invariance. A scale (and statistics based on it, such as a mean) is comparable across countries or groups only if scalar measurement invariance is achieved. These levels of measurement invariance are hierarchically structured (nested) levels so that metric invariance requires configural, scalar requires configural and metric invariance, and strict invariance requires scalar, metric and configural invariance. At the basic level of invariance, configural invariance, all groups have common factors and items. Figure 2(a) in the previous section illustrates this requirement of the configural level of invariance for one country where the same item-factor structure of the latent construct of CCLIMATE is also expected for other countries if the configural level of invariance holds. To have the common factors having the same meaning across groups and the same measurement unit suggests that metric invariance (also known as weak invariance) is established. This implies equal strength of the association between items and factors for all countries. See Figure 2(b) for the illustration of the regression line for one country. For multiple groups there are multiple latent regression lines with the same slopes implying that the loadings are equivalent across countries. Metric level of invariance is considered the minimum level of invariance if making comparisons of the relationship between factors across countries and observed variables (Byrne, 2008; Gregorich, 2006). At the scalar (or strong) invariance level, the intercepts are all equal and thus all items indicate the same cross-cultural differences in latent means. Scalar (or strong) level of measurement invariance is the minimum level of invariance required if we intend to perform a valid cross-country comparison of the scale scores (i.e., means comparisons). It is not good practice to examine mean score differences of the developed latent construct with scalar non-invariance because scalar non-invariance yields inconsistencies in the interpretation of the meaning of the score of the latent construct across different countries. Lack of equivalent interpretations of the mean score occurs whenever not all of the observed means can be predicted from the observed variables. Most likely this introduces additional bias of these particular variables and therefore leads to a biased estimation of the latent mean. Thus item-factor structure with different intercepts should be assumed across countries.

Finally, strict invariance, as the strongest level of invariance, implies that the conditional variance of the response is invariant across groups. Strict invariance requires that, in addition to equal factor loadings and intercepts, the residual (specific factor plus error variable) variances are equivalent across groups. Meredith (1993) argued that strict invariance is a necessary condition for a fair and equitable comparison. However, from the 1990s to date, the governing norm reflected in research practice is that metric (or weak) invariance, or strong invariance at best, constitutes sufficient evidence for measurement invariance and allows for cross-country comparisons (Little, 1997; Horn and McArdle, 1992; Steenkamp and Baumgartner, 1998).

For ordered-categorical variables (e.g., four-point Likert-scale responses), Muthén and Asparouhov (2002) and Millsap and Yun-Tein (2004) proposed a categorical MGCFA approach with the solution for full measurement invariance. Full invariance requires that the intercepts, factor loadings, residual variances, and the thresholds are held equal across groups. The approach described from the two studies also solves the difficulties of model specification and identification in multiple groups CFA where there are a lot of ways to free and fix the model parameters correctly. For example, at the configural level of invariance, the first factor loadings and all residual variances are fixed to 1.0, and the factor means are fixed to zero, therefore, all other factor loadings, factor variances and thresholds are freely estimated across groups. Other way, as suggest in Glockner-Rist and Hoijtink (2003) for assessing the thresholds in categorical items, is to assess the invariance of the thresholds while holding the factor loadings constant across groups. Once it has been controlled for non-invariant thresholds, metric invariance can be assessed. Finally, the equality of residual variances can be tested when the items have the same quality as measures of the latent variable in all countries. Different levels of invariance levels with different model specification and indentification the categorical observed variables can be analysed using the Theta parameterisation implemented in Mplus 7.1 (Muthén and Muthén, 1998-2012). An example of the syntaxes for testing measurement invariance (e.g., common factor and items at the configural level of invariance) using the continuous and categorical approaches is enclosed in the Appendix.

METHODS

This paper compares the levels of measurement invariance established (i.e., configural, metric and scalar) in each complex scale using both continuous and categorical MGCFA approaches. A database containing survey responses from 73 100 teachers and 4 362 school principals in the 23 countries who participated in TALIS 2008 was examined. As mentioned previously, the motivation for this study was a proposal within an OECD expert paper from Rutkowski and Svetina (2013) where the authors examined two complex scales and compared measurement invariance of the scale in five countries from the TALIS 2013 field trial data using both continuous and categorical approaches. The focus of this paper is, therefore, to examine further and to provide empirical findings on the performance of measurement invariance for the complex scales developed in TALIS 2008 when continuous and categorical approaches are used in order to help inform future discussions about the most appropriate ways to perform the validation of scales in surveys such as TALIS. There are 10 teacher scales and 1 principal scale examined in this paper. The scales were initially developed from teacher and principal questionnaires using a continuous MGCFA approach, as described in the TALIS 2008 Technical Report (OECD, 2010).[2] The scales examined in this paper are as follows:

- Classroom disciplinary climate (CCLIMATE)
- Structuring (TPSTRUC)
- Teacher-student oriented (TPSTUD)
- Enhanced activities teaching practice (TPACTIV)
- Teacher-student relations (TSRELAT)
- Teacher constructivist beliefs (TBCONS)
- Direct transmission beliefs (TBTRAD)
- Teacher self-efficacy (SELFEF)
- Exchange and co-ordination (TCEXCHAG)
- Professional collaboration (TCCOLLAB)
- Principal constructivist beliefs (PBCONS)

[2] Teacher and principal questionnaires can also be found in TALIS 2008 Technical Report (OECD, 2010).

The questionnaires or items used to operationalise each of the aforementioned scales are all of categorical response format. The items for the each of the scales are detailed in Table 1. Necessary data preparation such as reverse coding the questionnaire items to have the same direction as the rest of the items and collapsing categories with low response rates was undertaken. There is no standard rule at where to collapse categories but any one category with zero response rates, even for only one country, should be merged to the closest category. In this study, any response rates less than 1% was considered very low. These preliminary steps are necessary to ascertain all countries have the same number of response categories and were sufficiently populated in all categories of the scale items. For example, four items - T31G, T31H, T31I, and T31J - are the items for the TSRELAT scale with four response categories “Strongly Agree”, “Agree”, “Disagree”, and “Strongly Disagree”. There were less than 1% response rates for the “Strongly Disagree” category for items T31G, T31H, and T31I. Therefore, categories “Disagree” and “Strongly Disagree” were collapsed into one category to represent the disagree response. Other scale items examined in this study where response categories were collapsed from the observed categories are reported in Table 1. Details about the other items associated with the constructed scales and other technical documentation on the data preparation can be obtained from Chapter 11 of the TALIS 2008 Technical Report (OECD, 2010).

Table 1. **Questionnaire items for complex scales**

Scale	Item Number and Item Statement	Response Categories
CCLIMATE	T43.a) When the lesson begins, I have to wait quite a long time for students to <quieten down>. T43.b) Students in this class take care to create a pleasant learning atmosphere. T43.c) I lose quite a lot of time because of students interrupting the lesson. T43.d) There is much noise in this classroom.	1 "Strongly Disagree" 2 "Disagree" 3 "Agree" 4 "Strongly Agree"
TPSTRUC	T42.b) I explicitly state learning goals. T42.c) I review with the students the homework they have prepared. T42.h) At the beginning of the lesson I present a short summary of the previous lesson. T42.i) I check my students' exercise books. T42.m) I check, by asking questions, whether or not the subject matter has been understood.	1 "Never or hardly ever" 2 "In about one-quarter of lessons" 3 "In about one-half of lessons" 4 "In about three-quarters of lessons" 5 "In almost every lesson"
TPSTUD	T42.d) Students work in small groups to come up with a joint solution to a problem or task. T42.e) I give different work to the students that have difficulties learning and/or to those who can advance faster. T42.f) I ask my students to suggest or to help plan classroom activities or topics. T42.n) Students work in groups based upon their abilities.	1 "Never or hardly ever" 2 "In about one-quarter of lessons" 3 "In about one-half of lessons" 4 "In about three-quarters of lessons" 5 "In almost every lesson"
TPACTIV	T42.j) Students work on projects that require at least one week to complete. T42.o) Students make a product that will be used by someone else. T42.q) I ask my students to write an essay in which they are expected to explain their thinking or reasoning at some length. T42.s) Students hold a debate and argue for a particular point of view which may not be their own.	1 "Never or hardly ever" 2 "In about one-quarter of lessons" 3 "In about one-half of lessons" 4 "In about three-quarters of lessons" 5 "In almost every lesson"
TSRELAT	T31.g) In this school, teachers and students usually get on well with each other. T31.h) Most teachers in this school believe that students' well-being is important. T31.i) Most teachers in this school are interested in what students have to say. T31.j) If a student from this school needs extra assistance, the school provides it.	1 "Strongly Disagree" * 2 "Disagree" * 3 "Agree" 4 "Strongly Agree"
TBCONS	T29.d) My role as a teacher is to facilitate students' own inquiry. T29.f) Students learn best by finding solutions to problems on their own. T29.i) Students should be allowed to think of solutions to practical problems themselves before the teacher shows them how they are solved. T29.l) Thinking and reasoning processes are more important than specific curriculum content.	1 "Strongly Disagree" * 2 "Disagree" * 3 "Agree" 4 "Strongly Agree"
TBTRAD	T29.a) Effective/good teachers demonstrate the correct way to solve a problem. T29.g) Instruction should be built around problems with clear, correct answers, and around ideas that most students can grasp quickly. T29.h) How much students learn depends on how much background knowledge they have – that is why teaching facts is so necessary. T29.k) A quiet classroom is generally needed for effective learning.	1 "Strongly Disagree" * 2 "Disagree" * 3 "Agree" 4 "Strongly Agree"
SELFEF	T31.b) I feel that I am making a significant educational difference in the lives of my students. T31.c) If I try really hard, I can make progress with even the most difficult and unmotivated students. T31.d) I am successful with the students in my class. T31.e) I usually know how to get through to students.	1 "Strongly Disagree" * 2 "Disagree" * 3 "Agree" 4 "Strongly Agree"
TCEXCHAG	T30.c) Discuss and decide on the selection of instructional media (e.g. textbooks, exercise books). T30.d) Exchange teaching materials with colleagues.	1 "Never" 2 "Less than once per year"

Scale	Item Number and Item Statement	Response Categories
	T30.e) Attend team conferences for the age group I teach. T30.f) Ensure common standards in evaluations for assessing student progress . T30.g) Engage in discussion about the learning development of specific students.	3 "Once per year" 4 "3-4 times per year" 5 "Monthly" 6 "Weekly"
TCCOLLAB	T30.h) Teach jointly as a team in the same class. T30.i) Take part in professional learning activities (e.g. team supervision). T30.j) Observe other teachers' classes and provide feedback. T30.k) Engage in joint activities across different classes and age groups (e.g. projects). T30.l) Discuss and coordinate homework practice across subjects.	1 "Never" 2 "Less than once per year" 3 "Once per year" 4 "3-4 times per year" 5 "Monthly" 6 "Weekly"
PBCONS	P32.d) The role of teachers is to facilitate students' own inquiry. P32.f) Students learn best by finding solutions to problems on their own. P32.l) Thinking and reasoning processes are more important than specific curriculum content. P32.i) Students should be allowed to think of solutions to practical problems themselves before the teacher shows them how they are solved.	1 "Strongly Disagree" * 2 "Disagree" * 3 "Agree" 4 "Strongly Agree"

Note. T: Teachers Questionnaire, P:Principal Questionnaire, *Collapsed categories

The Models and Model Estimation

Note that the operational calibration sampling method was used for the TALIS 2008 measurement invariance analysis (OECD, 2010). The calibration samples are "test samples" which represent *probability proportional to size* (PPS) samplings of 1 000 teachers and 150 principals. Calibration samples are typically used for estimating parameters and fixing the same estimated parameters to the other comparable groups. Normally, this calibration process is practical and efficient for incomplete test designs[3]. However, TALIS 2008 implemented a complete survey design where all of the questionnaires items were administered to all of the teachers or principals. For this paper, and with currently available software, the entire TALIS 2008 dataset is used (rather than a calibration sample) for testing the measurement invariance of the aforementioned scales, with equal contribution from each of the participating country using a weighting variable in the analysis (computation of the weights variable is explained later).

Given that there is missing data on observed variables, a model-based approach is used. The assumption is that missing data on observed variables is missing at random (MAR): a variable with missing data (e.g., variable with omitted [related to non-understanding] or not reached [related to fatigue or non-cooperation]) which has a systematic relationship with the missingness in the dependent variable, and where missingness may not completely be caused by a random process (Graham, 2012; Schafer and Graham, 2002). In other words, any observations with missing data under the MAR mechanism will have the same statistical behaviour (i.e., probability distribution) as in the observations

[3] In the incomplete test designs, survey or test items are groups in blocks. Only subsets of the total items are administered. Not all items are tested to all subjects or respondents.

that are not missing. In this present study, MLR and WLSMV are used as the estimators for the continuous and categorical modellings. The estimators take into account the complex design of TALIS 2008 as well as give a proper treatment for the missing data observed in the models. That is, in the model-based approach, missing data is treated using MLR for the continuous modelling and WLSMV with univariate and bivariate probit regression techniques for all the cases with data on the latent construct and its categorical observed variables for the categorical modelling. The model-based procedure is performed for handling MAR missing data where the observations with missing responses are not strictly removed (e.g., by listwise or pairwise deletion) but preserved in the analysis by taking into account the probability distribution of missingness.

Sampling weights were used to account for the unequal selection probabilities of the observations in the sample because TALIS adopts a complex, two-stage, survey design (OECD 2008). This means that, for each country, teachers were randomly selected (second stage) from the list of randomly selected schools (first stage). If the appropriate sampling weights are not used when performing analyses, this can translate into a biased estimate of sampling error. For this reason, sampling weights were used in the measurement invariance modelling to correctly estimate the corresponding standard errors. Also where there are differences in the number of teachers or principals sampled between countries the estimation weights variable is provided in the database for the purpose of country-level estimate. In the measurement invariance analysis, where all countries are simultaneously analysed, the given weights are rescaled so that each of the participating country will contribute equally in the estimate. The rescaled weights are not provided in the TALIS database but easily computed from the given final weights. The mechanics on rescaling the weights using SPSS macros are explained in further details in Gonzalez (2012). To accommodate the sampling weights in both categorical and continuous model and to avoid biased estimates of the standard error, pseudo-maximum likelihood (PML) is implemented in Mplus for both MGCFA modellings with MLR and WLSMV estimators (Asparouhov, 2005). PML yields consistent and unbiased estimates by maximising the weighted log-likelihood of the measurement invariance MGCFA models. This study has observed an identical result up to 1 000 digits of precision when using either the final weights or the rescaled weights for the MGCFA invariance analysis in Mplus.

For the continuous approach, linear mean structure modelling is applied, meaning the categorical responses of the scales' items are treated and estimated as continuous, assuming multivariate normally distributed responses. For the categorical approach, Theta parameterisation (Muthén and Muthén, 1998-2012) is applied where ordered-categorical responses are used as they are observed in the dataset. There are two steps of model specification in Theta parameterisation. First, at the configural level of invariance, the residual variances are constrained to 1.0 and the factor means are constrained to zero allowing for the unequal estimations of thresholds and loadings of the observed categorical responses in all groups. The second step is specified for the higher levels of invariance where thresholds and loadings are estimated equally across groups. Also, the residual variances are restricted to 1.0 and factor means are restricted to zero in one group but they are freely estimated in the others. These two specifications for invariance testing are used to avoid misspecification in the categorical modelling when simultaneous multiple group comparison is analysed (see Muthén and Asparouhov, 2002 for an excellent primer on parameterisation). The categorical modelling with Theta parameterisation is similar

to partial credit model (PCM) within the item response theory framework (e.g.,Muraki, 1982; Muraki, 1990; Masters and Wright, 1997) which is used for cognitive and contextual outcomes such as in PISA 2009 assessment data (e.g. OECD, 2012).

Model-data Fit

CFA and MGCFA are hypothesised testable models to predict the variability that is observed between the variables by creating a latent construct. To evaluate what was modelled from the observed variables, the fit of the baseline model (i.e., the congeneric model as defined in Graham, 2006) and each model of invariance, we used the following acceptable criteria: non-significant chi-square test with *p* degrees of freedom (χ_p^2); *Comparative Fit Index* greater than 0.90; *Tucker-Lewis Index* greater than 0.90; *Root Mean Square Error Approximation* less than 0.08; *Standardised Root Mean Square Residual* less than 0.06; and *Weighted Root-Mean-Square Residual* less than 0.90, that is: *CFI* ≥ 0.90, *TLI* ≥ 0.90, *RMSEA* ≤ 0.08, *SRMR* ≤ 0.06 and *WRMR* ≤ 0.90 (Yu, 2002; MacCallum et al., 1996; Byrne, 2008; Hu and Bentler, 1998, 1999; Steiger, 1990).

The evaluation of measurement invariance for each scale is based on the criteria in Chen (2007) as reported in the TALIS 2008 Technical Report (OECD, 2010), that is to evaluate the hypothesis of equal loadings and, additionally, equal thresholds. We used the standard changes in the absolute values of the fit indices. An MGCFA model is viewed as invariant based on the absolute changes in the fit indices between a higher level of invariance to a lower level, that is |Δ*CFI*| ≤ 0.010, |ΔTLI| ≤ 0.010, |Δ*RMSEA*| ≤ 0.010, and |ΔSRMR| ≤ 0.005, |ΔWRMR| ≤ 0.005. In the TALIS Technical Report (OECD, 2010), either the criterion for *CFI* or *RMSEA* is used for deciding whether the level of invariance is established. These strict cut-offs should only be used when comparing two groups (two countries in this case), and are here only used as a rough orientation as they are rather conservative measures for a large number of countries comparison. We followed the same criteria to keep the consistency in making the decision about the level of invariance achieved by each scale in both continuous and categorical approaches. Noteworthy of the *RMSEA* index is that it is not a good indicator for model comparisons as the index is very sensitive to the number of parameters in the model but insensitive to the sample size (Brown, 2006). Thus, *RMSEA* favours parsimony modelling, penalises a more complex model and thus cannot be regarded as infallible. Change in *CFI* provided the best performance when comparing nested models, because the index is independent of sample size and model complexity (Cheung and Rensvold, 2002). We observed from the report that change in *CFI* and/or *RMSEA* which is close enough to the absolute difference of 0.010 (e.g., less than 0.014) is also acceptable for practical decisions of the level of measurement invariance. A more lenient criterion is suggested in Rutkowski and Svetina (2013) when comparing a large number of groups (e.g., 10, 20 or more), that is to relax the more stringent cut-offs for the fit indices (e.g., *CFI*, TLI, and *RMSEA* absolute changes less than 0.02). Therefore, these criteria are used for this study to evaluate the comparison.

It is worth noting that the reported fit indices between the two approaches are not directly comparable because of non-identical modelling approaches (i.e., continuous vs. categorical). The fit indices are separately examined for each modelling to evaluate construct validity of the scale within the

structural equation modelling framework. Thus, the levels of invariance established are independently evaluated from either the categorical or continuous approach. The match of achieved invariance can be an indication of the evidence that the categorical or continuous MGCFA modelling is preferred, and is a focal interest of this study.

A summary of the methods for this present study is presented in Table 2 to show the comparison of modellings implemented using Mplus 7.1 for continuous and categorical MGCFA.

Table 2. Comparison on scaling approaches for TALIS 2008 complex scales

	Continuous Modelling	**Categorical Modelling**
Analysis Type	Measurement invariance (MGCFA)	Measurement invariance (MGCFA)
Observed Variables	Categorical scale	Categorical scale
Missing Data Treatment	Model based approach for missing at random distribution	Model based approach for missing at random distribution
Model Estimator	MLR	WLSMV
Variables Specification	Treated as and assumed continuous	Categorical
Sampling Consideration in Mplus	Complex sampling design, sampling weights with stratification and cluster variables	Complex sampling design, sampling weights with stratification and cluster variables
Analysis Specification	Specified as continuous MEAN STRUCTURE	THETA PARAMETERIZATION
Mode-data Fit Criteria	CFI, TLI, RMSEA, SRMR	CFI, TLI, RMSEA, WRMR
Models Comparison*	$\Delta\ CFI$ and $\Delta\ RMSEA$	$\Delta\ CFI$ and $\Delta\ RMSEA$
Invariance Level Achieved	Total Number of Scales = 11	Total Number of Scales = 11
Configural	11	11
Metric	11	11
Scalar	0	3

Note. *For two groups comparison, Chen (2007) recommends to view models as invariant if absolute change in *CFI* ≤ .01, absolute change in *RMSEA* ≤ .01 and absolute change in SRMR ≤ .005. Tables 3 to 13 present the results of the model-data fit indices for each scale. More lenient criteria are used with respect to a large number of countries comparison (i.e. 23 countries.)

RESULTS

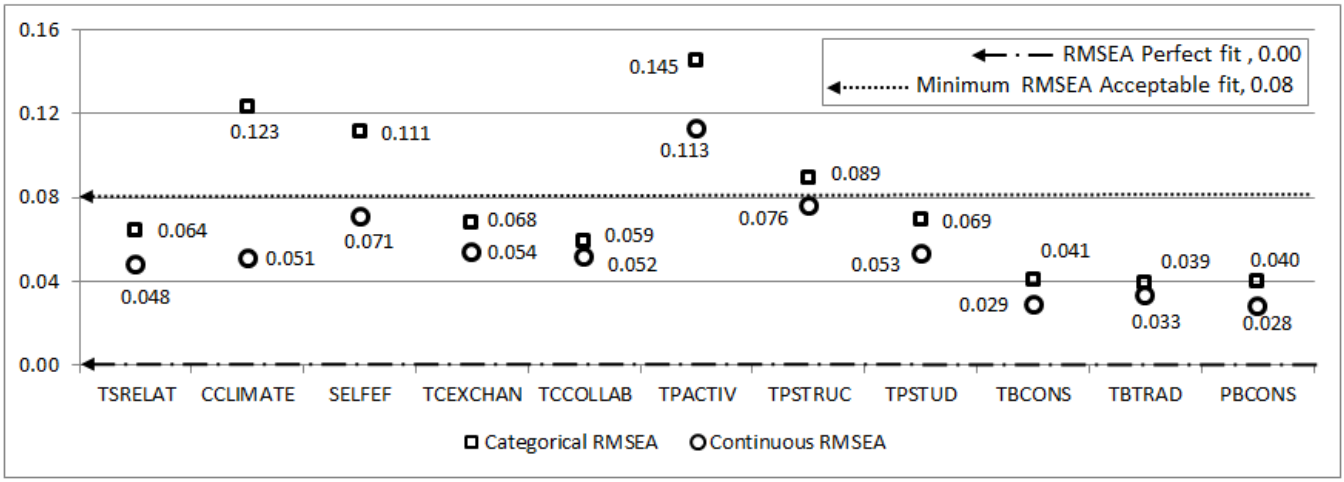
This paper evaluates the performance of measurement invariance of TALIS 2008 complex scales when using either the continuous or the categorical MGCFA approach. The result describes the performance of both approaches for constructing complex scales.

Overall Comparison

To better illustrate the performance between continuous and categorical MGCFA approaches in constructing and validating complex scales, we compared the distribution of the model-data fit indices across 11 scales examined in this paper. Figures 3, 4, and 5 illustrate *RMSEA* and *CFI* for configural, metric, and scalar invariance levels for both the continuous and categorical approaches. The y-axis shows the distribution of *RMSEA* or *CFI*, and the scales represented by their names are separately positioned on the x-axis. The dash-dotted lines in the figures illustrate the cut-off for exact or perfect fits where *RMSEA=0.00* and *CFI=1.00*. Different model parameters are estimated depending on the use of either the continuous or the categorical approach (e.g., Millsap and Yun-Tein, 2004; Muthén and Asparouhov, 2002 for modelling details), direct comparison of the fit indices between continuous and categorical approaches was not intended. Therefore the model-data fits indices presented are separately evaluated.

We observed acceptable patterns of the *RMSEA* index in most cases from the continuous approach as well as from the categorical approach, where any values located at or below the round dotted line shows that *RMSEA* tends to stay within the acceptable range (i.e., *RMSEA* ≤ 0.08). Mediocre criterion is represented by any *RMSEA* value less than or equal to 0.10. When the continuous approach is applied, one scale at the configural level, and nine scales at the scalar level showed *RMSEA* falling outside the acceptable (or mediocre range). From 11 scales, the index is larger than the acceptable or mediocre cut-offs when the categorical approach is applied, that is three scales at the configural level, five scales at the metric level, and eight scales at the scalar level. This is most likely a result of its sensitivity to a large number of parameters in the scales (i.e., four to six items in the scales and three to five response options for each item). These findings are presented in Tables 3 to 13 and illustrated in the Figures 3 to 5. For the evaluation of *CFI*, none of the scales showed poor model-data fit at the configural level (i.e., all had *CFI* ≥ 0.90) for either approach. The acceptable criterion and the observed *CFI* are illustrated in Figures 3 to 5 (represented by the round dotted line in each *CFI* plot) where *CFI* for all scales was observed to be greater or equal to 0.90. For the continuous approach, one scale at metric level and nine scales at scalar level displayed poor fit. From the categorical approach, two scales at metric invariance and eight scales at scalar invariance displayed poor fit (i.e., *CFI*<0.90).

Figure 3. RMSEA and CFI comparisons for configural invariance



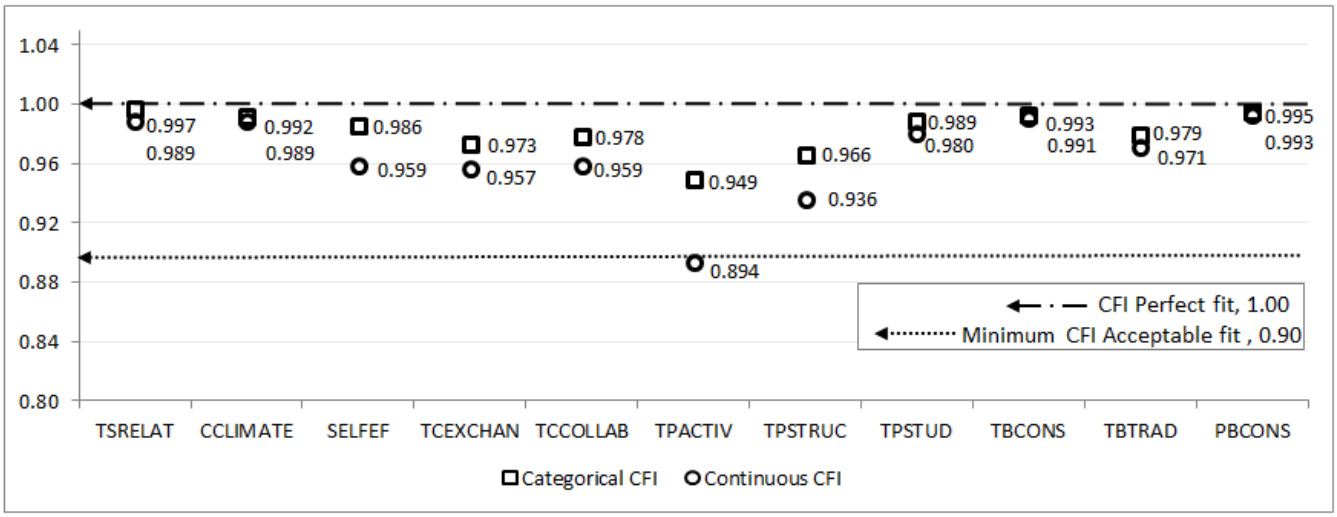
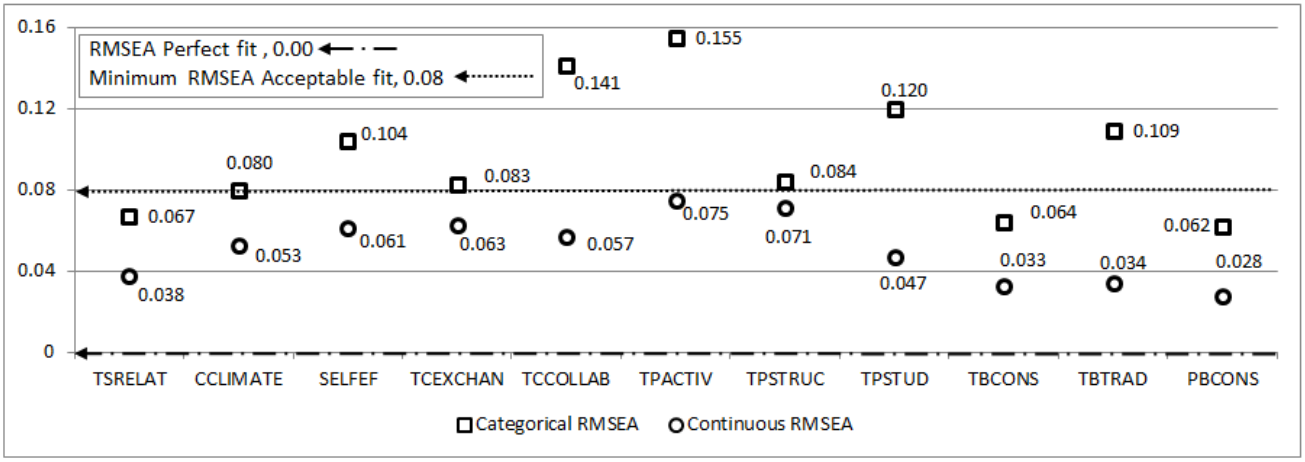


Figure 4. RMSEA and CFI comparisons for metric invariance



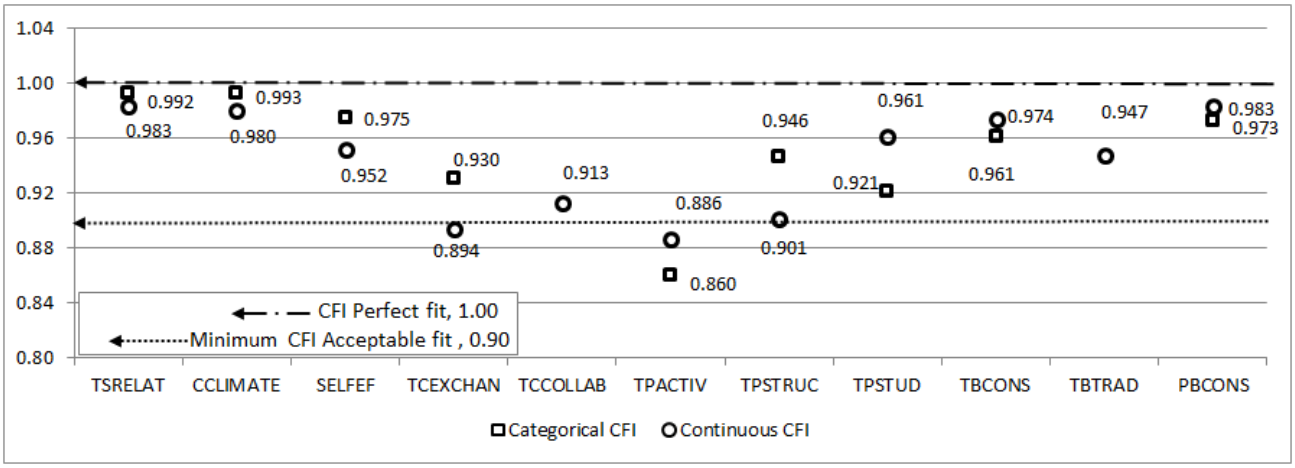
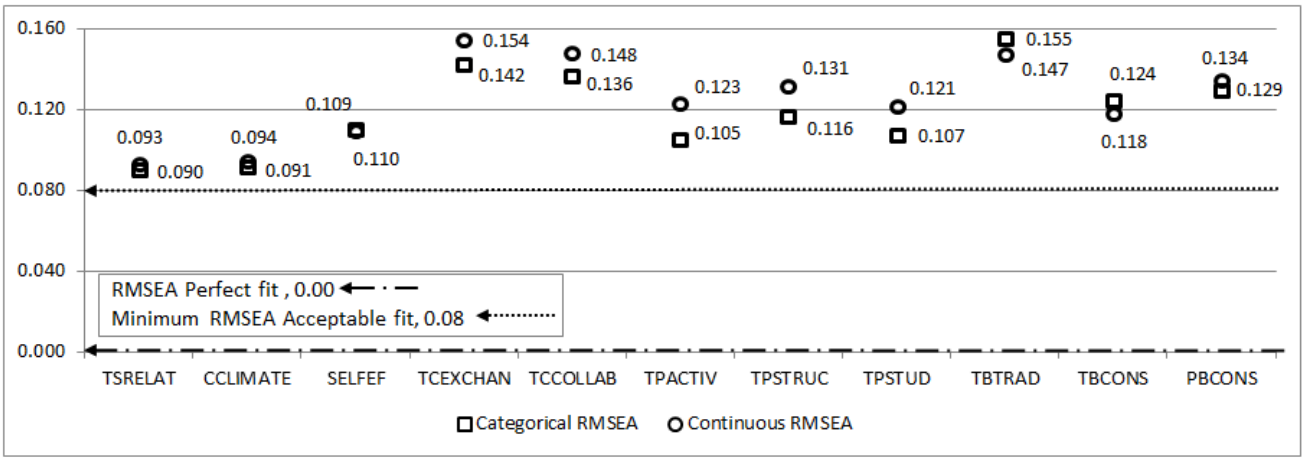
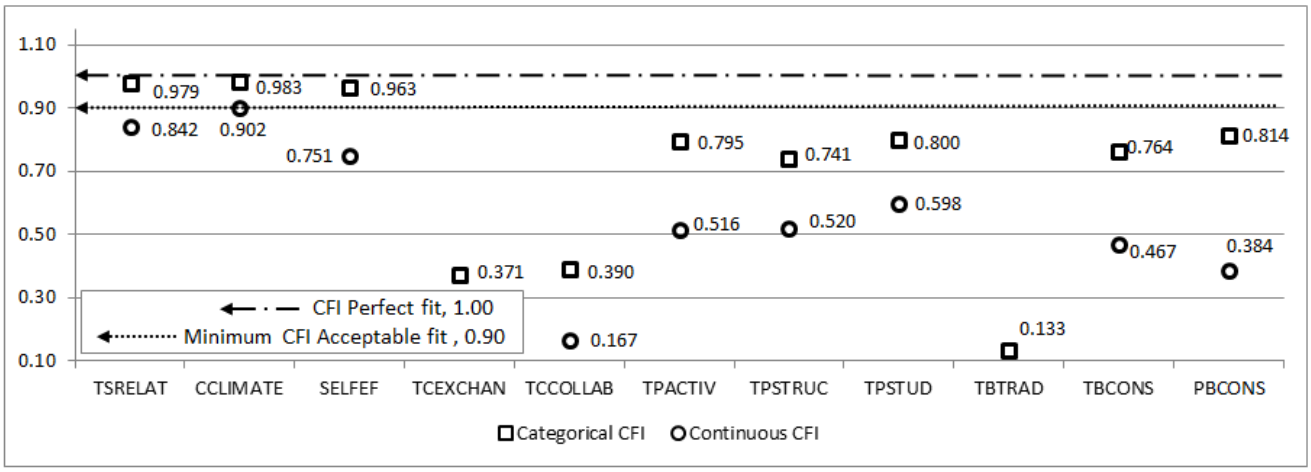


Figure 5. RMSEA and CFI comparisons for scalar invariance





Measurement Invariance for Individual Scales

Changes in the fit indices for measurement invariance testing for each tested scale are presented in Tables 3 to 13. In these tables, numbers in bold show acceptable changes in the model fit at a level of measurement invariance. This means the scale has established the particular level of invariance. The values in italics should be interpreted as close to establishing the particular invariance level. Details for these model-data fit indices for each scale and their absolute changes between levels of invariance can be seen in the aforementioned tables.

The interpretation of measurement invariance starts with the configural invariance level as the baseline model. At the configural level, this study observed that all countries have common factors and items for all 11 scales with acceptable model-data fits from both approaches. At metric invariance, all of the 11 scales established metric invariance on the basis of more lenient cut-offs for *CFI* and/or *RMSEA* with respect to the very large number of countries involved in the analysis.

Following the hierarchical order of the measurement invariance testing, all of the 11 scales from continuous and nine scales from categorical MGCFA can be further tested at higher level of invariance. It has been observed that none of the teacher or principal scales established scalar invariance when the continuous approach is applied. From the categorical approach, the highest level of invariance for TBTRAD and TCCOLLAB is at the configural level. Also three scales from the operationalised teacher scales—TSRELAT, CCLIMATE, and SELFEF—observed from the categorical approach have established scalar invariance.

Table 3. Teacher-student relations (TSRELAT)

Invariance Level	Continuous				Categorical			
	CFI	*RMSEA*	Δ*CFI*	Δ*RMSEA*	*CFI*	*RMSEA*	Δ*CFI*	Δ*RMSEA*
Configural	0.989	.048	-	-	0.997	.064	-	-
Metric	0.983	.038	**.006**	**.010**	0.992	.067	**.005**	**.000**
Scalar	0.842	.093	.141	.055	0.979	.090	**.013**	**.008**

Table 4. Classroom disciplinary climate (CCLIMATE)

Invariance Level	Continuous				Categorical			
	CFI	*RMSEA*	Δ*CFI*	Δ*RMSEA*	*CFI*	*RMSEA*	Δ*CFI*	Δ*RMSEA*
Configural	.989	.051	-	-	.999	.123	-	-
Metric	.980	.053	**.009**	**.002**	.993	.080	**.001**	**0.009**
Scalar	.902	.094	.078	.041	.983	.091	**.010**	**0.002**

Table 5. Self-efficacy (SELFEF)

Invariance Level	Continuous				Categorical			
	CFI	*RMSEA*	Δ*CFI*	Δ*RMSEA*	*CFI*	*RMSEA*	Δ*CFI*	Δ*RMSEA*
Configural	.959	.071	-	-	.986	.111	-	-
Metric	.952	.061	**.007**	**.010**	.975	.104	**.011**	**.007**
Scalar	.751	.109	.201	.048	.963	.110	**.012**	**.006**

Table 6. Direct transmission beliefs (TBTRAD)

Invariance Level	Continuous				Categorical			
	CFI	*RMSEA*	Δ*CFI*	Δ*RMSEA*	*CFI*	*RMSEA*	Δ*CFI*	Δ*RMSEA*
Configural	.971	.033	-	-	.979	.039	-	-
Metric	.947	.034	.024	**.001**	.731	.109	.248	.070
Scalar	.000	.147	.947	.113	.133	.155	.337	.046

Table 7. Teacher constructivist beliefs (TBCONS)

Invariance Level	Continuous				Categorical			
	CFI	*RMSEA*	Δ*CFI*	Δ*RMSEA*	*CFI*	*RMSEA*	Δ*CFI*	Δ*RMSEA*
Configural	.991	.029	-	-	.993	.041	-	-
Metric	.974	.033	*.017*	**.004**	.952	.064	.032	*.023*
Scalar	.467	.118	.507	.085	.764	.124	.197	.060

Table 8. Principal constructivist beliefs (PBCONS)

Invariance Level	Continuous				Categorical			
	CFI	*RMSEA*	Δ*CFI*	Δ*RMSEA*	*CFI*	*RMSEA*	Δ*CFI*	Δ*RMSEA*
Configural	.993	.028	-	-	.995	.040	-	-
Metric	.983	.028	**.010**	**.000**	.973	.062	*.022*	*.022*
Scalar	.384	.134	.458	.106	.814	.129	.112	.067

Note. *Italicized values* should be interpreted as being very close to the acceptable change

Table 9. Teaching practices: structuring (TPSTRUC)

Invariance Level	Continuous				Categorical			
	CFI	*RMSEA*	Δ*CFI*	Δ*RMSEA*	*CFI*	*RMSEA*	Δ*CFI*	Δ*RMSEA*
Configural	.936	.076	-	-	.966	.089	-	-
Metric	.901	.071	.035	**.005**	.946	.084	*.020*	**.005**
Scalar	.520	.131	.381	.060	.741	.116	.205	.03

Note. *Italicized values* should be interpreted as being very close to the acceptable change

Table 10. Teaching practices: teacher-student oriented (TPSTUD)

Invariance Level	Continuous				Categorical			
	CFI	*RMSEA*	Δ*CFI*	Δ*RMSEA*	*CFI*	*RMSEA*	Δ*CFI*	Δ*RMSEA*
Configural	.980	.053	-	-	.989	0.069	-	-
Metric	.961	.047	*.019*	**.006**	.921	0.120	.068	.051
Scalar	.598	.121	.363	.074	.800	0.107	.121	0.013

Table 11. Teaching practices: enhanced activities (TPACTIV)

Invariance Level	Continuous				Categorical			
	CFI	*RMSEA*	Δ*CFI*	Δ*RMSEA*	*CFI*	*RMSEA*	Δ*CFI*	Δ*RMSEA*
Configural	0.894	.113	-	-	.949	.145	-	-
Metric	.886	.075	**.008**	.038	.860	.155	.089	**.010**
Scalar	.516	.123	.370	.048	.795	.105	.065	.050

Table 12. Co-operation: exchange and co-ordination (TCEXCHAG)

Invariance Level	Continuous				Categorical			
	CFI	*RMSEA*	Δ*CFI*	Δ*RMSEA*	*CFI*	*RMSEA*	Δ*CFI*	Δ*RMSEA*
Configural	.957	.054	-	-	.973	.068	-	-
Metric	.894	.063	.063	**.009**	.930	.083	.043	**.015**
Scalar	.093	.154	.801	.091	.371	.142	.559	.059

Table 13. Professional collaboration (TCCOLLAB)

Invariance Level	Continuous				Categorical			
	CFI	*RMSEA*	Δ*CFI*	Δ*RMSEA*	*CFI*	*RMSEA*	Δ*CFI*	Δ*RMSEA*
Configural	.959	.052	-	-	.978	.059	-	-
Metric	.913	.057	.046	**.005**	.790	.141	.188	.082
Scalar	.167	.148	.746	.091	.390	.136	.400	.005

CONCLUSION AND DISCUSSION

The purpose of this paper is to evaluate two different modelling techniques for constructing complex scales - the practice followed in TALIS 2008 using the continuous MGCFA modelling and an advancement of categorical MGCFA modelling in order to inform future decisions regarding the most appropriate methodology to be used in the construction of complex scales in surveys such as TALIS. The highest level of invariance established from the continuous approach is at the metric level. This is due to, at the scalar level of invariance, response categories in the scale items being treated as linear when using the continuous approach, and therefore non-invariant thresholds from the categories not being controlled for across groups. In other words, the characteristics of each response category (e.g., a respondent's tendency to choose one category or the other) in an item is not explicitly accounted for. The findings from this study imply that an acceptable level of invariance (to permit for a scale score means comparison across countries) cannot be established for all scales when using the continuous approach. This conclusion is consistent with the literature in large-scale assessments, that is, for most complex scales only metric invariance is established.

Therefore, comparisons of scale mean values based on continuous models across countries should then be made with a careful interpretation, as the mean scores may have a slightly different meaning due to the differences in the latent factor structure. For example, a score for the enhanced learning activities teaching practice scale in Country A, say 10 points, may or may not correspond to the same score of 10 for enhanced learning activities teaching practice scale in Country B. Thus, the levels of teachers' practice in enhanced learning activities between these countries are not straightforwardly comparable or explainable based on the mean scores. This could possibly be as a result of ambiguous wording, poor translation of the questionnaire items or other unexplained nuisances that can be legitimately regarded as important factors in the designs of the questionnaire items and survey.

In contrast, differences in non-metric invariance scale scores can be due to differences in the underlying metrics between countries or due to true differences in the countries' cultural backgrounds. The metric level of invariance of scales is notably a prerequisite for comparing the relationships (i.e., correlations) between two or more scales across countries. Differences in the associations between scales, for example the correlations between structuring (TPSTRUC), teacher-student oriented (TPSTUD) and enhanced activities (TPACTIV) teaching practices across Country A and Country B, are allowed to be directly compared because of the similarities in the underlying factor-item relationships (i.e. metric) of the scales.

This study showed that the categorical approach empirically resolved the complication caused by the continuous modelling approach when constructing the following complex scales: teacher-student relations (TSRELAT); classroom disciplinary climate (CCLIMATE); and teacher self-efficacy (SELFEF). The categorical approach is methodologically the most appropriate approach to be applied for scales with categorical observed variables, and empirically observed from TSRELAT, CCLIMATE, and SELFEF. Each of these scales consists of four measured items with four response categories in each item. Note also two out of four response categories for TSRELAT and SELFEF were collapsed, so the final number of categories in the analysis was three, which in turn affects the degree of asymmetry of the response categories distribution. When there are a very few number of response categories, the continuous modelling demonstrated a problem establishing the scalar level of invariance. The establishment of scalar invariance for these scales when using the categorical MGCFA allows for an unequivocal interpretation of the scale means comparison across countries. It is advisable that the categorical approach be considered in scaling and validating these scales to assure measurement invariance. Thus, interpretations regarding the degree of achieved invariance are improved and hence more credible. Furthermore, given that group comparisons are distorted when continuous MGCFA models are applied to inherently categorical data (Lubke and Muthén, 2004), fewer inferential errors around group comparisons can be made if a categorical approach is chosen. To summarise, when the distribution of the variables appropriately fits the MGCFA modelling, estimation errors can be minimised and model-data agreement is more likely to be attained.

LIMITATIONS AND SUGGESTIONS

The application of different modellings governed by the observed data is not new but often overlooked. For example, there are different statistics used for computing the association between the two variables: Pearson's correlation coefficient is used for two continuous variables where t-test is used for testing the hypothesis; Spearman's correlation is used for two categorical variables where the assumptions of normal theory with a sufficient sample size is applied for testing the hypothesis; and polychoric correlation coefficient is used for two ordered-categorical variables where chi-square test is used in the hypothesis testing. The application of statistical methodologies that are governed by the distributions of the observed data (i.e., continuous vs. categorical) has triggered decades of debate (Borgatta & Bohrnstedt, 1980; Lubke and Muthén, 2004; Muthén & Kaplan, 1985; Steven, 1946, 1959; Townsend & Ashby, 1984; Zumbo & Zimmerman, 1993). The intention of this paper is not to resolve the debate into measurement invariance testing methodologies but to offer a heuristics starting point for deciding which type of modelling to used in common practice for international surveys.

The decision to use either continuous or categorical MGCFA approaches for measurement invariance is crucial when scale means are to be compared across countries. This study suggests that three scales from categorical MGCFA modelling are more promising for cross-country comparisons. All other scales only established the metric level of invariance and closer attention should be paid as to why these scales behave differently. The questionnaire items designed with four-point versus five-point categories and with different extremities (i.e., agreement versus frequency responses) which are dubious in determining their behaviours or response distributions when using either of the MGCFA modellings. The present study observed three complex scales that are accountable to which MGCFA modelling technique is more likely to retain the distributional assumptions (i.e., continuous vs. categorical) of the observed questionnaire items. However, the strict assumption that is imposed by the scalar invariance level for all of the other complex scales is not solvable by either of these modellings and yet to be further examined. Further research based on simulation studies, where there is the availability of software (e.g., Mplus, LISREL, SAS) to appropriately model both continuous and categorical MGCFA, could further examine invariance testing using different numbers of items per scale, numbers of response categories per item, numbers of groups, as well as cut-offs for model-data fit indices, to allow for a better understanding of the behaviours of these scales and the robustness of the approaches.

The cut-offs used to determine the level of measurement invariance are widely used for continuous outcome variables for two groups comparison. However, there is not much evidence in the literature for clear cut-offs when comparing a large number of groups, and particularly in categorical MGCFA. Thus, more research is needed on determining clear model-fit cut-offs for measurement invariance for the categorical approach. For example, Rutkowski and Svetina (2013) reported a more liberal *RMSEA* and a more stringent *CFI* should be expected, thus suggesting that more relaxed cut-offs ought to be used when the number of groups compared is greater than two in the continuous modelling. Alternatively, the Satorra-Bentler test (Hu and Bentler, 1999; Bentler and Satorra, 2010) for the continuous approach or DIFFTEST (Muthén and Muthén, 1998-2012) for the categorical approach could be applied to test for

significant measurement invariance. When the *p*-value of the tests is less than 0.05 it indicates that the more restrictive model did not provide a poorer fit compared to the less restrictive model, which implies the measurement invariance hypothesis at a higher level could not be rejected.

These standard approaches; continuous or categorical, using multiple-group confirmatory factor analysis with strict or exact equality constraints, are rather conservatives and could be too cumbersome to be practical for the analysis of many groups particularly in international surveys where there can be a large number of non-invariant measurement parameters. Studies in Asparouhov and Muthén (2013), Finch and French (2008) and Muthén and Asparouhov (2013) urged for the application of Bayesian structural equation modelling measurement invariance analysis and multiple group factor analysis alignment.[4] These are radically different methods that can be used to estimate the scales where there is no assumption of measurement invariance and yet can estimate the factor mean and variance parameters in each group while discovering the optimal measurement invariance pattern in most groups as well as identifying non-invariance pattern which may occur in a small number of groups. The methods follow the same logic as an exploratory approach where non-invariant, partial non-invariant and invariant parameters across groups are informed and used in the MGCFA modelling. According to these studies, the number of measurement non-invariance parameters and the amount of measurement non-invariance is minimal and factor means and variances are optimally estimated. In other words, there is no presence of exact or strict invariance assumed but measurement invariance is simplified and tolerated across different groups compared. Despite their practical appeal for international surveys, the methods offer valuable alternatives that are optimal solutions for evaluating measurement invariance for a large number of countries comparison, permits for cross-country comparisons, and improves interpretability of complex scales across different countries.

[4] Bayesian structural equation modelling measurement invariance and multiple group factor analysis alignment are available in Mplus 7.1 for continuous and binary variables and their extensions to ordered-categorical variables are currently under development.

REFERENCES

Asparouhov, T. (2005), Sampling weights in latent variable modelling. Structural equation modeling 12, 3. pp. 411-434.

Asparouhov, T. and Muthén, B. (2013), Multiple Group Factor Analysis Alignment. Mplus Web Notes.

Babaku, E. and Ferguson, C. E. and Jöreskog, K. G. (1987), The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. Journal of Marketing Research 37, pp. 72-141.

Bentler, P. and Satorra, A. (2010), Testing Model Nesting and Equivalence. Psychological Methods 15, 2, pp. 111-123.

Borgatta, E.F., and Bohrnstedt, G.W. (1980), Level of measurement-Once over again. Sociological Methods and Research, 9, 147-160.

Brown, T. A. (2006), Confirmatory Factor Analysis. New York. Guilford Publication.

Byrne, B. M. (2008), Testing for multigroup equivalence of a measuring instrument: A walk through the process. Psicothema 20, 4, pp. 872-882.

Chen, F. F. (2007), Sensitivity of goodness of fit indexes to lack of measurement invariance. Structural Equation Modeling 14, 3, pp. 464-504.

Cheung, G. W. and Rensvold, R. B. (2002), Evaluating goodness-of-fit indexes for testing measurement invariance. Structural equation modelling, 9, 2, pp. 233-255.

Curran, P. J. and West, S. G. and Finch, J. F. (1996), The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. Psychological methods 1, 1, pp. 16.

DiStefano, C. (2002), The Impact of Categorization with Confirmatory Factor Analysis. Structural Equation Modeling 9, 3, pp. 327-346.

Finch, W. H., & French, B. F. (2008). Using exploratory factor analysis for locating invariant referents in factor invariance studies. Journal of Modern Applied Statistical Methods, *7*(1), 18.

Glockner-Rist, A. and Hoijtink, H. (2003), The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. Structural Equation Modeling 10, 4, pp. 544-565.

Gonzalez, E. J. (2012), Rescaling sampling weights and selecting mini-samples from large-scale assessment databases. pp. 117-134.

Graham, J. M. (2006), Congeneric and (Essentially) Tau-Equivalent Estimates of Score Reliability: What They Are and How to Use Them. Educational and Psychological Measurement 66, 6, pp. 930-944.

Graham, J. W. (2012), Analysis of Missing Data. in Missing Data: Analysis and Design (Statistics for Social and Behavioral Sciences. New York: Springer Science.

Gregorich, S.E. (2006), Do self-report instruments allow meaningful comparisons across diverse population groups. Testing measurement invariance using the confirmatory factor analysis framework. Medical Care. Vol. 44, 11/3. pp. S78–94.

Horn, J. L. and McArdle, J. J. (1992), A practical and theoretical guide to measurement invariance in aging research. Experimental Aging Research 18, 3-4, pp. 117-144.

Hu, L. and Bentler, P. M. (1998), Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. Psychological methods 3, 4, pp. 424.

Hu, L. and Bentler, P. M. (1999), Cutoff criteria for fit indexe s in covariance structure analysis: Coventional criteria versus new alternatives. Structural Equation Modeling 6, 1, pp. 1-55.

Jöreskog, K. G. (1969), A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34(2), 183-202.

Little, T. D. (1997), Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. Multivariate Behavioral Research 32, 1, pp. 53-76.

Little, T. D. and Slegers, D. W. (2005), Factor analysis: Multiple groups. Encyclopedia of statistics in behavioral science.

Long, J. S. (1980), Confirmatory factor analysis. Beverly Hills, CA; Sage.

Lord, F. M. (1980), Applications of Item Response Theory to Practical Testing Problems. New Jersey: Lawrence Erlbaum.

Lord, F. M., and Novick, M. R. (1968), Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley.

Lubke, G. and Muthén, B. (2004), Applying Multigroup Confirmatory Factor Models for Continuous Outcomes to Likert Scale Data Complicates Meaningful Group Comparisons. Structural Equation Modeling. 11, pp. 514-534.

MacCallum, R. C. and Browne, M. W. and Sugawara, Hazuki M. (1996), Power analysis and determination of sample size for covariance structure modeling. Psychological methods. Vol. 1, 2, pp. 130-149.

Masters, G., & Wright, B. (1997), The partial credit model. In W.J. van der Linden and R.K. Hambleton (eds.), Handbook of modern item response theory. New York/Berlin/Heidelberg. Springer.

Meredith, W. (1993), Measurement invariance, factor analysis and factorial invariance. Psychometrika 58, 4, pp. 525-543.

Millsap, R. E. and Yun-Tein, J. (2004), Assessing factorial invariance in ordered-categorical measures. Multivariate Behavioral Research 39, 3, pp. 479-515.

Muthén, B. and Asparouhov, T. (2002), Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. Mplus web notes 4, 5, pp. 1-22.

Muthén, B. and Asparouhov, T. (2013), BSEM Measurement Invariance Analysis. Mplus Web Note.

Muthén, B. and Kaplan, D. (1985), A Comparison of Some Methodologies for the Factor Analysis of Non-Normal Likert Variables. British Journal of Mathematical and Statistical Psychology 38, pp. 171-189.

Muthén, L. K. and Muthén, B. O. (1998-2012), Mplus User's Guide 7. Los Angeles, CA: Muthén & Muthén.

Muthén, B. and du Toit, S.H.C. and Spisic, D. (1997), Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes . Unpublished Technicl Report.

Muraki, E. (1982), A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16 (2), 159-176.

Muraki, E. (1990), Fitting a polytomous item response model to Likert-type data. Applied Psychological Measurement, 14 (1), 59-71.

OECD (2012), *Scaling Outcomes (Technical Report)*. OECD Publishing. Paris.

OECD (2010), *TALIS 2008 Technical Report (Technical Report)*. OECD Publishing. Paris.

Rutkowski, L. and Svetina, D. (2013), Assessing the Hypothesis of Measurement Invariance in the Context of Large-Scale International Surveys. Educational and Psychological Measurement.

Schafer, J. L. and Graham, J. W. (2002), Missing Data: Our View of the State of the Art. Psychological Methods 7, 2. pp. 147-177.

Steenkamp, Jan-Benedict E. M. and Baumgartner, H. (1998), Assessing Measurement Invariance in Cross-National Consumer Research. Journal of Consumer Research 25, 1, pp. 78-107.

Steiger, J. H. (1990), Structural Model Evaluation and Modification: An Interval Estimation Approach - Multivariate Behavioral Research - Volume 25, Issue 2. Multivariate Behavioral Research 25, 2, pp. 173-180.

Stevens, S.S. (1946), On the theory of scales of measurement. Science, 103, 677-680.

Stevens, S.S. (1959), Measurement, Psychophysics and Utility. In Churchman, G.W. and Ratoosh, P. (Eds). Measurement: Definitions and theories. New York, Wiley.

Townsend, J. T. and Ashby, F. G. (1984), Measurement Scales and Statistics: The Misconception Misconceived, Psychological Bulletin, 96, pp. 394-401.

Sörbom, D. (1974), A general method for studying differences in factor means and factor structure between groups. British Journal of Mathematical and Statistical Psychology 7, 2, pp. 229-239.

Wright, B. D. (1968), Sample-free test calibration and person measurement. Paper presented at the National Seminar on Adult Education Research. Chicago.

Yu, C. (2002), Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. Unpublished Dissertation. University of California Los Angeles.

Zumbo, B. D., & Zimmerman, D. W. (1993), Is the selection of statistical methods governed by level of measurement? Canadian Psychology, 34, 390-400.

APPENDIX

Example syntax for continuous (configural) invariance testing (MGCFA) in Mplus 7.1.

```
TITLE: Configural Model

DATA:  FILE IS "D:\temp\inputData.dat";

VARIABLE:
    NAMES ARE id1 id2 id3 wgt1 strata clus
    Item1 Item2 Item3 Item4;

    USEVARIABLES ARE Item1 Item2 Item3 Item4;
    !CATEGORICAL ARE Item1 Item2 Item3 Item4;

    GROUPING IS id3 (1=g1 2=g2 3=g3 4=g4 5=g5 6=g6 7=g
    8=g8 9=g9 10=g10 11=g11 12=g12 13=g13 14=g14 15=g15 16=g16 17=g17
    18=g18 19=g19 20=g20 21=g21 22=g22 23=g23);

    MISSING ARE Item1 Item2 Item3 Item4 (999999);
    WEIGHT IS wgt1;
    STRATIFICATION IS strata;
    CLUSTER IS clus;

ANALYSIS:
    TYPE IS COMPLEX;
    !PARAMETERIZATION=THETA;
    !ESTIMATOR = WLSMV;
    ITERATIONS = 100000;
    CONVERGENCE = 0.000001;
    MODEL = CONFIGURAL;

MODEL:
    scale by Item1 Item2 Item3 Item4;

OUTPUT:
   sampstat standardized modindices tech1 tech4;

PLOT: TYPE= PLOT1 PLOT2;
!SAVEDATA: DIFFTEST=configuralDC.dif;
```

Example syntax for categorical (configural) invariance testing (MGCFA) in Mplus 7.1.

```
TITLE: Configural Model

DATA:  FILE IS "D:\temp\inputData.dat";

VARIABLE:
    NAMES ARE id1 id2 id3 wgt1 strata clus
    Item1 Item2 Item3 Item4;

    USEVARIABLES ARE Item1 Item2 Item3 Item4;
    CATEGORICAL ARE Item1 Item2 Item3 Item4;

    GROUPING IS id3 (1=g1 2=g2 3=g3 4=g4 5=g5 6=g6 7=g
    8=g8 9=g9 10=g10 11=g11 12=g12 13=g13 14=g14 15=g15 16=g16 17=g17
    18=g18 19=g19 20=g20 21=g21 22=g22 23=g23);

    MISSING ARE Item1 Item2 Item3 Item4 (999999);
    WEIGHT IS wgt1;
    STRATIFICATION IS strata;
    CLUSTER IS clus;

ANALYSIS:
    TYPE IS COMPLEX;
    PARAMETERIZATION=THETA;
    ESTIMATOR = WLSMV;
    ITERATIONS = 100000;
    CONVERGENCE = 0.000001;
    MODEL = CONFIGURAL;

MODEL:
    scale by Item1 Item2 Item3 Item4;

OUTPUT:
   sampstat standardized modindices tech1 tech4;

PLOT: TYPE= PLOT1 PLOT2;
!SAVEDATA: DIFFTEST=configuralDC.dif;
```