Structural Equation Modeling: A Multidisciplinary Journal

ISSN: 1070-5511 (Print) 1532-8007 (Online) Journal homepage: https://www.tandfonline.com/loi/hsem20

Multiple-Group Invariance with Categorical Outcomes Using Updated Guidelines: An Illustration Using M*plus* and the lavaan/semTools Packages

Dubravka Svetina, Leslie Rutkowski & David Rutkowski

To cite this article: Dubravka Svetina, Leslie Rutkowski & David Rutkowski (2019): Multiple-Group Invariance with Categorical Outcomes Using Updated Guidelines: An Illustration Using M*plus* and the lavaan/semTools Packages, Structural Equation Modeling: A Multidisciplinary Journal, DOI: 10.1080/10705511.2019.1602776

To link to this article: https://doi.org/10.1080/10705511.2019.1602776

Published online: 29 Apr 2019.

Submit your article to this journal

Article views: 310

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=hsem20

Structural Equation Modeling: A Multidisciplinary Journal, 0: 1–20, 2019
© 2019 Taylor & Francis Group, LLC
ISSN: 1070-5511 print / 1532-8007 online
DOI: https://doi.org/10.1080/10705511.2019.1602776

Routledge
Taylor & Francis Group

Check for updates

Multiple-Group Invariance with Categorical Outcomes Using Updated Guidelines: An Illustration Using M*plus* and the lavaan/semTools Packages

Dubravka Svetina, Leslie Rutkowski, and David Rutkowski
Indiana University Bloomington School of Education

Meaningful comparisons of means or relationships between latent constructs across groups require evidence that measurement is equivalent across the studied groups– a property known as measurement equivalence or invariance (ME/I). Methods typically involve an evaluation of increasingly stringent models via confirmatory factor analysis, a typical assumption of which is continuous observed variables. When that assumption is not met – as is often the case in many surveys – alternative methods that directly model the categorical nature of the data exist. Although well established, categorical ME/I models pose a number of complexities and various recommendations for their evaluation. To that end, we describe the current state of categorical ME/I and demonstrate an up-to-date method for model identification and invariance testing. In the tutorial, we exemplify a common approach to establishing ME/I via multiple-group confirmatory factor analysis using M*plus* and the lavaan and semTools packages in R.

Keywords: measurement invariance, model identification, categorical variables, M*plus*, lavaan, semTools in R

INTRODUCTION

In cross-cultural operational and academic research, meaningful latent construct comparisons across multiple populations are of frequent interest. Examples of these sorts of constructs include physical self-perception (e.g., Hagger, Biddle, Chow, Stambulova, & Kavussanu, 2003), cognitive-emotional regulation (e.g., Megreya, Latzman, Al-Attiyah, & Alrashidi, 2016), and educational achievement (OECD, 2014, 2016). For example, Hagger et al. (2003) examined the appropriateness of Fox and Corbin's (1989) hierarchical multidimensional model of physical self-perception among different (outside Western Europe) cultures by comparing adolescents from Great Britain, Russia, and Hong Kong. Megreya et al. (2016) focused their study on examining the psychometric properties of the cognitive emotion regulation questionnaire. Specifically, the authors were interested in its tenability to reflect the hypothesized nine-factor structure of cognitive emotion regulation in four different Arabic-speaking Middle Eastern countries (university students from Egypt, Saudi Arabia, Kuwait, and Qatar were considered). Additionally, the authors examined gender differences on each of the nine subdomains (e.g., self-blame, acceptable, rumination, etc.). Lastly, international large-scale assessments (ILSAs), such as the Trends in International Mathematics and Science Study (TIMSS; Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009), compare dozens of populations in terms of educational achievement in Mathematics and Science as well as other non-achievement domains (e.g., values, beliefs, and attitudes toward teaching profession). In all of these contexts, a criterion for comparing scale scores is that the latent variable is understood and measured equivalently across all groups/countries. This property is referred to as *measurement invariance* (Meredith, 1993) or *lack of bias* (Lord, 1980).[1]

In a cross-cultural context, a common approach for establishing evidence of measurement equivalence or invariance (ME/I) is through multiple-group confirmatory factor analysis (MG-CFA; Horn & McArdle, 1992; Jöreskog, 1971; Meredith, 1993). This method – a straight-forward extension of CFA –

Correspondence should be addressed to Dubravka Svetina dsvetina@indiana.edu Counseling & Educational Psychology, Indiana University Bloomington School of Education, 201 N. Rose Ave, Bloomington 47405-1006, United States

relies on a set of hierarchical tests to impose increasingly restrictive equality constraints on parameters of interest across comparison groups. To the degree that equivalence is achieved, increased comparative inferences are possible. Until recently, most methodological research in this area relied on two-group comparisons (Chen, 2007; Cheung & Rensvold, 2002; French & Finch, 2006; Meade, Johnson, & Braddy, 2008); however, recent research has pointed to the fact that comparisons across more than two groups create added complexity (Asparouhov & Muthén, 2014; Rutkowski & Svetina, 2014, 2017; Svetina & Rutkowski, 2017). Given the relative recentness of these developments, a paucity of instructional information is available for applied cross-cultural researchers that are working with many groups.

To that end, the main purpose of this paper is to provide a didactic illustration of MG-CFA for establishing measurement invariance that would be appropriate for contexts with many groups with the emphasis of using Wu and Estabrook (2016) approach to model identification and ME/I testing. We exemplify the MG-CFA approach to conducting ME/I using M*plus* and the lavaan and semTools packages in R. A secondary goal is to survey current methodological approaches to establishing ME/I. This paper is organized as follows. The next section provides *Background* on how ME/I is defined, with an emphasis on utilizing Wu and Estabrook (2016) model identification and constraints for analyzing measurement invariance. Model fit and evaluation are also discussed. The following section, *Tutorial*, provides researchers and practitioners a step-by-step guide for conducting ME/I analyses utilizing M*plus* 7.2 (Muthén & Muthén, 1998), the lavaan (Rosseel, 2012) and the semTools (Jorgensen, Pornprasertmanit, Schoemann, & Rosseel, 2018) packages in R (R Core Team, 2018) for ordered categorical outcomes. The next section discusses alternative approaches to ME/I, with a focus on tests of partial invariance as potential next steps when ME/I at the scale level is not supported. Lastly, we offer a few concluding remarks regarding methodological considerations in conducting ME/I, as well as document briefly some of the newest developments in ME/I literature.

BACKGROUND

Defining and establishing ME/I

We first begin with the general factor model given by $\mathbf{\Sigma} = \mathbf{\Lambda\Phi\Lambda}' + \mathbf{\Theta}$, where $\mathbf{\Sigma}$ represents the covariance matrix of the observed variables, $\mathbf{\Lambda}$ represents a matrix of factor loadings that express the strength of the relationship between the vector of latent variables, ξ, with associated covariance matrix $\mathbf{\Phi}$, to the arbitrary vector of observed variables, $\boldsymbol{Y}$. Finally, $\mathbf{\Theta}$ represents the covariance matrix of the measurement errors for Y. The mean structure is included as ν. Then, the observed variables' means can be represented by $\mathrm{E}(Y) = \mathrm{E}(\nu + \mathbf{\Lambda}\boldsymbol{\xi} + \boldsymbol{\epsilon})$. With the usual assumption that $\mathrm{E}(\boldsymbol{\epsilon}) = 0$ and $\mathbf{E}(\xi) = \boldsymbol{\kappa} = 0$, then $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\nu}$. This model is easily generalizable to the multiple population context by permitting separate covariance matrices for each population/group. In other words, $\mathbf{\Sigma}^{(g)}$with mean structure $\boldsymbol{\nu}^{(g)},\ g = 1, \ldots,\ G$.

The general approach for establishing ME/I is that if the null hypothesis, $\mathrm{H}_0 :\ \mathbf{\Sigma}^{(1)} = \mathbf{\Sigma}^{(2)} = \cdots = \mathbf{\Sigma}^{(G)}$ is rejected, a series of hierarchically nested tests follow. The first of these is one of the same form or *configural invariance* (Horn & McArdle, 1992; Horn, McArdle, & Mason, 1983). Here, the number and pattern of parameters are assumed equal across groups; however, the *values* of the parameters are assumed different within identification constraints. Researchers use a chi-square test of model fit, supplemented by several model fit indices, discussed subsequently, to evaluate the tenability of this hypothesis. The typical next test in the hierarchy is one of the equal loadings, otherwise known as *metric invariance* or *weak factorial invariance* (Meredith, 1993). The null hypothesis of metric invariance is $H_0 : \mathbf{\Lambda}^{(1)} = \mathbf{\Lambda}^{(2)} = \ldots = \mathbf{\Lambda}^{(G)}$. In other words, the pattern and value of factor loadings are equivalent across populations. The traditional test is an overall chi-square test and a chi-square difference test. These formal hypothesis tests are supplemented by overall and incremental fit indices. The usual last test in the hierarchy is that of *scalar* or *strong factorial invariance*. Here, in addition to equal loadings, the intercepts are assumed equal. The null hypothesis in this case is $H_0 : \mathbf{\Lambda}^{(1)} = \mathbf{\Lambda}^{(2)} = \ldots = \mathbf{\Lambda}^{(G)},\ \boldsymbol{\nu}^{(1)} = \boldsymbol{\nu}^{(2)} = \ldots = \boldsymbol{\nu}^{(G)}$. As before, a decision is made based on overall and incremental chi-square tests and fit indices, the details of which we discuss subsequently.[2]

In the former description, the distribution of observed variables, $\boldsymbol{Y}$ is assumed multivariate normal; however, in many surveys, observed variables are binary (yes/no), Likert-scaled, or otherwise ordinal in nature. Ignoring the categorical distribution of these variables in a CFA context can have severe consequences on parameters, model fit, and cross-group comparisons (Beauducel & Herzberg, 2006; Lubke & Muthén, 2004; Muthén & Kaplan, 1985). When the normality assumption is untenable, alternative estimators are available. Of these, the diagonally weighted least

[1] Through an item response theory (IRT) framework, measurement invariance is also known as an absence of *differential item functioning* (Hambleton & Rogers, 1989; Mellenbergh, 1994; Swaminathan & Rogers, 1990); however, we do not emphasize the IRT perspective here.

[2] Although some scholars advocate for *strict factorial invariance* (Meredith, 1993) or equality of residual variances as a condition for comparing latent means (Deshon, 2004; Lubke & Dolan, 2003), in practice, this level of invariance is rarely pursued given that scalar invariance supports cross-group comparisons of manifest (or latent) variable means on the latent variable of interest (Hancock, 1997; Little, 1997; Thompson & Green, 2006).

squares (DWLS) and variants are commonly implemented (Muthén & Asparouhov, 2002). We describe the categorical multiple-group approach next.

Based on the work of Millsap (2011), Muthén and Christoffersson (1981), Muthén and Asparouhov (2002), and others, the methods of establishing ME/I for categorical observed variables are well established. We start with a $p \times 1$ vector of observed variables, $\boldsymbol{Y}$ that take discrete ordered values $0,\ 1,\ 2, \ldots, C$. It is assumed that for each $Y_j,\ j = 1,\ 2, \ldots,\ p$ there is an underlying continuous latent response variable, Y_j^* the value of which determines the observed category of Y_j. And Y_j^* is related to Y_j through a set of $C + 1$ threshold parameters, $\boldsymbol{\tau}_j = \left(\tau_{j0}, \tau_{j1}, \ldots, \tau_{jC+1}\right)$ where $\tau_{j0} = -\infty$ and $\tau_{jC+1} = \infty$. The probability that Y_j = c is given as:

$$P\left(Y_j = c\right) = P\left(\tau_{jc} \leq Y_j^* \leq \tau_{jc+1}\right) \tag{1}$$

for $c = 0,\ 1, \ldots, C$. The model for the vector of latent response variables is given as:

$$\boldsymbol{Y}^* = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\epsilon} \tag{2}$$

where factor loadings and residuals are defined as in the normal case. Here, however, $\boldsymbol{\nu}$ is a vector of latent intercept parameters. The mean and covariance structure of this model is the same as the normal case: $\mathrm{E}(\mathbf{Y}^*) = \boldsymbol{\nu}, \mathrm{Cov}(\mathbf{Y}^*) = \boldsymbol{\Sigma}^* = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$. With this specification, $\mathrm{E}(\mathbf{Y}^*) = \boldsymbol{\nu}$ is assumed to be zero for identification purposes. A further identification restriction is that the latent response variables have unit variance, which implies that $\boldsymbol{\Theta}$ is estimated as a remainder (Millsap, 2011, p. 128; Muthén & Asparouhov, 2002). When $C > 1$, which we assume here, the correlation between the latent response variables is a polychoric correlation.

As in the normal context, the categorical factor model can be similarly extended to the multiple-group case by allowing for separate thresholds and covariance matrices of the latent response variables for each population, that is $\boldsymbol{\tau}^{(k)}$and $\boldsymbol{\Sigma}^{*(k)}$, with $k = 1,\ 2, \ldots,\ K$ (note that $\boldsymbol{\nu}^{(k)} = \mathbf{0}$ for all k). Similarly, when data are ordinal, the tests are for "baseline" invariance, equal slopes, and equal slopes and thresholds (analogous to configural, metric, and scale invariance, respectively). Again, overall and difference chi-square tests are used to evaluate the tenability of the respective invariance hypotheses.

Unpacking Wu and Estabrook (2016)

In their 2016 article, Wu and Estabrook discuss the identification issues in CFA for ordered categorical outcomes within the context of invariance testing across groups. Their approach differs from current practices of conducting ME/I, where current practice is to first establish a baseline model and subsequently impose increasing parameter restrictions. According to Wu and Estabrook, the current approach is not optimal because it is dependent on the way the baseline model is identified with respect to the scales of latent continuous responses, which further could imply restrictions on what/how parameters are constrained to equality and ultimately may lead to different conclusions. The authors are particularly concerned with how the threshold model is identified. Given the popularity of Likert-type items on assessments of various constructs, the treatment of data as categorical becomes even more important to consider and model. Thus, our goal is to focus on selected solutions Wu and Estabrook proposed in terms of model identification and testing for ME/I.[3] According to Wu and Estabrook, this also implies that after establishing configural invariance, threshold invariance is tested first, followed by invariance testing for loadings – an approach that differs slightly from operational ME/I whereupon establishing the configural invariance, equal loadings followed by invariance testing in intercepts/thresholds are considered. We utilize Wu and Estabrook's language in the tutorial for convenience purposes, including reference to model specification for equal thresholds (Proposition 4) and equal thresholds and loadings (Proposition 7). Specifically, we first identify the model using delta parameterization such that diag($\boldsymbol{\Phi}$) = $\mathbf{I}$, $\boldsymbol{\kappa} = \mathbf{0}$, $\boldsymbol{\nu} = \mathbf{0}$, and diag($\boldsymbol{\Sigma}$) = $\mathbf{I}$ (equation 7 in Wu & Estabrook).[4] This is equivalent to a baseline model specification that allows for $\mathbf{T}^{(g)}$, $\boldsymbol{\Lambda}^{(g)}$, and $\boldsymbol{\Theta}^{(g)}$ to differ across groups, while $\boldsymbol{\nu}$ remains fixed to 0. Subsequent steps in ME/I testing are presented through Proposition 4 (threshold invariance), with model identification restrictions per equation 15 in Wu and Estabrook ($\boldsymbol{\nu}^{(1)} = \mathbf{0}$, diag ($\boldsymbol{\Sigma}^{(1)}$) = $\mathbf{I}$, and for all groups: $\boldsymbol{\kappa}^{(g)} = \mathbf{0}$ and diag($\boldsymbol{\Phi}^{(g)}$) = $\mathbf{I}$). And followed by Proposition 7 (model with threshold and loading invariance for three or more ordered categories), with model identification restrictions per equation 19 (diag ($\boldsymbol{\Sigma}^{(1)}$) = $\mathbf{I}$, $\boldsymbol{\nu}^{(1)} = \mathbf{0}$, diag($\boldsymbol{\Phi}^{(1)}$) = $\mathbf{I}$, and for all groups, $\boldsymbol{\kappa}^{(g)} = \mathbf{0}$). We illustrate these steps in fitting a single factor model with four indicators that are treated as categorical (with 4 categories) using M*plus* and lavaan and semTools packages in the *Tutorial* section.

Model fit evaluation

The first step in establishing ME/I is to examine the fit of a baseline model – that is, for each population/group, a test

[3] Readers are directed to the original article for complete detail on all conditions posited by Wu and Estabrook (2016).

[4] As presented in Figure 1 in Wu and Estabrook, there are six different ways that the model can be identified as a baseline model while remaining statistically equivalent. Superscript $^{(g)}$ denotes different parameters for different groups. For purposes of our tutorial, we select one path to identify the model and test for threshold followed by loadings invariance.

of same form should be adequate prior to examining more restrictive models of invariance (i.e., constraining thresholds and/or loadings to be equal across the groups). The reason is that if the baseline models do not fit, proceeding with more restrictive models is not meaningful. Assuming the baseline model fit is reasonable, a typical approach to evaluate the tenability of ME/I is to conduct chi-square difference tests for models with equal thresholds and/or loadings. However, research has shown the sensitivity of the chi-square (difference) test to sample size (Bagozzi, 1977; Bentler & Bonett, 1980) and inflated Type I error rate (Yuan & Chan, 2016). More so, even for a short, four-item scale administered in 20 countries/groups, 570 pair-wise comparisons in factor loadings alone are possible. Compounding this with typical sample sizes in the thousands in cross-country surveys, detecting misfit with the chi-square is highly likely (Bentler & Bonett, 1980). Thus, in addition to the chi-square difference test, researchers have recommended using model fit indices such as (change in) Comparative Fit Index (CFI; Bentler, 1990) or (change in) the root-mean-square error of approximation (RMSEA; Steiger & Lind, 1980) between the models. Specifically, Cheung and Rensvold (2002) recommended a change in CFI to be equal to or greater than −.010 as evidence of non-invariance.[5]

Further, Chen (2007) suggested that changes in CFI (ΔCFI) equal to or greater than −.010 supplemented by a change in the RMSEA (ΔRMSEA) less than or equal to .015 were indicative of non-invariance when sample sizes were equal across groups and larger than 300 in each group. Chen also recommended ΔCFI not less than −.005 and ΔRMSEA at least as small as .010 when sample sizes were unequal and each sample size was smaller than 300.

In a series of studies, Rutkowski and Svetina examined the normal and categorical models for ME/I in the context of a large number of groups (up to 20) with unidimensional (Rutkowski & Svetina, 2014, 2017) and multidimensional constructs (Svetina & Rutkowski, 2017). Across their studies, the authors made several recommendations as well as cautionary notes when using the indices. Their findings suggest that currently available model fit measures (i.e., chi-square, RMSEA, CFI, and TLI and their difference/changes in statistics) may be insufficient across different settings. This is largely documented in their recommendations (and adjustments) across the studies and contexts under consideration (large sample sizes, large number of groups, underlying model and data characteristics being aligned or misaligned). For example, in 2014, assuming normal model for their analyses, Rutkowski and Svetina recommended to adjust the typical criteria to evaluate ME/I in large number of groups setting to consider the change in RMSEA of .030 (.010) for evaluating metric (scalar) invariance, and −.020 (−.010) for changes in CFI for the same evaluations, respectively. Further, in 2017, recognizing the disjuncture between the generating models (ordered categorical) and the analytic models (normal) in operational settings, the authors investigated the performance of the fit indices in evaluating ME/I when the generating and analytic models aligned (i.e., both were categorical). Their recommendations were slightly adjusted such that for slopes, changes in CFI greater than or equal to −.004 and changes in RMSEA less than or equal to .050, and for slopes and thresholds, changes in CFI greater than or equal to −.010 in magnitude and changes in RMSEA less than or equal to .010 were considered.

Recent studies, such as those by Finch and French (2018) and Kim, Cao, Wang, and Nguyen (2017), further extend the conversation regarding the cutoff values by emphasizing newer approaches for testing ME/I alongside their own recommendations (see *Discussion*).

With that, we provided Table 1 as a summary of established and emerging recommendations of evaluating ME/I in the literature over the last several decades.[6] This table is meant to guide an analyst to adopt recommendations for conducting ME/I analyses by considering various aspects in empirical investigations (i.e., Chen, 2007; Cheung & Rensvold, 2002), as well as those that attempted to mimic contexts of ILSAs (e.g., Rutkowski & Svetina, 2014; Svetina & Rutkowski, 2017).

As noted in Table 1, researchers have proposed various recommendations at different levels of evaluation, when investigating ME/I (i.e., not one size fits all). For example, Rutkowski and Svetina (2017) proposed recommendations for cutoff scores (e.g., ΔRMSEA) for metric invariance and scalar invariance separately, while other authors reported changes in relative model fit more generally. Additionally, there seems to be a lack of agreement among scholars as to which recommendation to adopt. While most of these studies were comprehensive, the limitation of simulation study designs naturally limits generalizability across all contexts. Further, while methodological recommendations exist as to how to conduct such analyses appropriately, criteria are based on situations that vary. More so, as Raykov, Marcoulides, and Millsap (2012) illustrated, in some cases, the overall fit indices may not be "sensitive to location violations of individual parameter constraints…" (p. 721). Most importantly, we recognize that these recommendations do not apply to all contexts; thus, a researcher should be aware that these recommendations are nonconforming to all situations and should report their choice to evaluation of model fit and findings. Nonetheless, in order

[5] This finding was supported by French and Finch (2006) in the multivariate normal context.

[6] We are not including all aspects of individual studies. For example, Cheung and Rensvold (2002) studied impact of factor variances, strengths of factor loadings, and factor correlations-aspects not necessarily examined in other studies (which included other aspects).

TABLE 1
Selected Examples of Varying Cutoff Values for Testing ME/I under Various Approaches

Source	*Approach*	*# Groups*	N	*# Factors*	*Distribution*	*Overall recommendations**
Chen (2007)	MG-CFA	2	150, 250, or 500 per group	1	Normal	ΔCFI ≥-.005, ΔRMSEA ≤ .010 ΔCFI ≤ −.005 or −.010 for CFI, ΔRMSEA ≥ .010 or .015 ΔGamma hat ≤ −.005 or −.008. ΔSRMR ≥ .025 or .030 for metric invariance testing ΔSRMR ≥ .005 or .010 for intercept and residual variance invariance testing
Cheung and Rensvold (2002)	MG-CFA	2	150 or 300 per group	2 or 3	Normal	ΔCFI ≥-.010, ΔGamma hat ≥-.001, ΔMcDonald's NCI ≥-.02
French and Finch (2006)	MG-CFA	2	150/150 150/500 or 500/500	2 or 4	Normal	ΔCFI less than −.01 or chi-square difference of $p < .05$ or .01 (with use of maximum likelihood)
French and Finch (2006)	MG-CFA	2	150/150 150/500 or 500/500	2 or 4	Ordinal	$\Delta\chi^2$ at .05 (but low power)
Finch and French (2018)	Equivalence testing	2	100, 200, 400, 600, 1000, 1500, or 2000 per group	1		ε0[+] (Equivalence) for some value of RMSEA Excellent fit: < 0.01 Close fit: 0.01–0.05 Fair fit: 0.05–0.08 Mediocre fit: 0.08–0.10 Poor fit: 0.10+
Kim et al. (2017)[†]	MG-CFA ML-CFA ML- FMM Bayesian Alignment	25 or 50	50, 100, or 1000 per group	1	Normal	MG CFA, ΔCFI with the cutoff of .01 ML CFA, ΔCFI with the cutoff of .01 BIC with total sample size Bayesian, the PPP of .05 and 95% CI
Rutkowski and Svetina (2014)	MG-CFA	10 or 20	Varied from 600 to 6,000 per group	1	Normal	ΔRMSEA ≤ .03 and ΔCFI ≥-.020 for metric; ΔRMSEA ≤ .01 and ΔCFI ≥-.010 for scalar
Rutkowski and Svetina (2017)	MG-CFA	10 or 20	Varied from 600 to 6,000 per group	1	Ordinal	ΔRMSEA ≤ .05 in conjunction with sig. Δχ2 and ΔCFI ≥-.004 for metric ΔRMSEA ≤ .01 in conjunction with sig. Δχ2 and ΔCFI ≥-.004 for scalar
Svetina and Rutkowski (2017)	MG-CFA	10 or 20	750 to 6,000 per group	2 or 5	Ordinal	ΔRMSEA ≤ .05 in conjunction with significant Δχ2 for metric ΔRMSEA ≤ .01 and ΔCFI ≥-.002 for scalar [for 3 or fewer dimensions]

Notes. *We recommend going to original sources for more nuanced recommendations the authors provided in their studies. Given different approaches, some of these studies or recommendations are not necessarily interchangeable, and should not be used as ultimate rules but rather guides to inform decisions regarding ME/I.

†In their article, Kim et al. provide synthesis/comparison of five approaches in testing for ME/I across many groups, including their strengths and weaknesses. Methods studied included: MG CFA (multiple group confirmatory factor analysis); ML CFA (multilevel CFA); ML FMM (multilevel factor mixture modeling); Bayesian (Bayesian approximate); and Alignment (alignment optimization).

[+] $\varepsilon 0 = \frac{df(RMSEA_0)^2}{m}$, where $RMSEA_0$ is a maximum tolerated RMSEA value; m is the number of groups; df is degrees of freedom for model (Yuan & Chan, 2016).

to meaningfully compare groups, establishment of ME/I is warranted.

TUTORIAL

In this section, we work through several examples using M*plus* 7.2 (Muthén & Muthén, 1998-2017), the lavaan (Rosseel, 2012) and semtTools (Jorgensen et al., 2018) packages in R (R Core Team, 2018).[7] For illustration purposes, we utilize four items from the bullying scale on 2011 TIMSS 4th grade (Mullis et al., 2009) for three arbitrarily chosen countries (31 = Azerbaijan; 40 = Austria; 246 = Finland). All items are measured on a 4-point Likert-type scale, ranging from 0 (never) to 3 (at least once a week). Items on this scale ask students how often during the year has any of the following happened to them at school; for example: *I was made fun of or called names*. Sample sizes for Azerbaijan, Austria, and Finland are 3808, 4457, and 4520, respectively.

As some research has suggested, assuming incorrect data distributions and fitting incorrect models can lead to inappropriate inferences (Rutkowski & Svetina, 2017). Thus, in our tutorial, we guide researchers in conducting analyses for ordinal data, such as that typical of questionnaires with Likert-type responses. We present abridged input files used in M*plus* followed by the R code and relevant functions implemented in the lavaan and semTools packages. Additionally, all input, output, and data files are available by contacting the first author or visiting https://figshare.com/s/3f1d195da6c78195dd70.

Mplus: identifying the baseline model and testing for configural invariance

In the text, we identify various parts of the input commands, a few selected output results[8] and make connections between model identification (and later on parameter equality constraints) and M*plus* code. As with any M*plus* input files, the path to the data file needs to be specified, unless, as it was the case here, the input file and datafile are located in the same folder, which then only necessitates the data file name (by default the output file will be saved in the same folder where input file is located).

[7] We note that large number of resources, including discussion sites and groups, as well as supplemental documentation, are available for M*plus* (http://www.statmodel.com/) and lavaan package (http://lavaan.ugent.be/; https://groups.google.com/forum/#!forum/lavaan).

[8] In Appendix, we provide selected annotated output.

```
TITLE: Configural invariance for four items on bullying scale.

DATA:
  FILE IS 'BULLY.dat';   ! Data with four indicator
variables (see names below) and ID for each group.
```

The variable command requires that variable names are given: note that in the data file (BULLY.dat) the first row contains the data for the first examinee–no header or variable names are included. Next, variables that are used in the analyses are listed after USEVARIABLES. These variable names are expected to appear in the MODEL part. The grouping variable IDCNTRY contains values associated with the country identification code – in this case, the grouping variable contains three groups: 31 (Azerbaijan), 40 (Australia), and 24 (Finland).

```
VARIABLE:
   NAMES ARE IDCNTRY R09A R09B R09C R09D;
   USEVARIABLES ARE IDCNTRY R09A R09B R09C R09D;
   CATEGORICAL ARE R09A R09B R09C R09D;

    GROUPING IS IDCNTRY (31=AZE 40=AUT 246=FIN); !
Three groups are considered.
```

Given that the observed variables are categorical, we use the default– the mean and variance adjusted diagonally weighted least squares (WLSMV) estimator.

```
ANALYSIS:
ESTIMATOR = wlsmv; ! Estimation for ordinal variables/
default in Mplus
H1ITERATIONS = 3000;
```

The model comments below outline the model identification for a single factor model with four categorical indicators. The first four lines under MODEL command indicate the phantom variable to specify the latent variate y* (see above section on *Defining and establishing ME/I*), the loading of which is fixed to 1 for all items. Additionally, factor mean [`F1@0`] and variance `F1@1` in all groups are fixed to 0 and 1, respectively. Several additional identifications (per equation 7) are required. Factor scale {`R09A@1 R09B@1 R09C@1 R09D@1`} is fixed to 1 in all groups, while intercept means [`y1-y4@0`] and residual variances `y1-y4@0` are fixed to 0 in all groups.

Note that since this is a baseline model, we expect thresholds and loadings to be estimated freely across the groups. This is evident by using the * in the commends `F1 BY y1-y4*` and `[R09A$1-R09A$3*]…[R09D$1-R09D$3*]`; respectively.

```
MODEL AUT:
```

```
MODEL:
     y1 BY R09A@1;          ! Must be fixed to 1 for
                              identification
     y2 BY R09B@1;
     y3 BY R09C@1;
     y4 BY R09D@1;
     F1 BY y1-y4*;          ! Loadings here estimated in
                              all groups
     F1@1;                  ! Factor variance fixed to 1
                              in all groups
     [F1@0];                ! Factor means fixed to 0 in
                              all groups
     {R09A@1 R09B@1         ! Factor scale fixed to 1 in
     R09C@1 R09D@1};          all groups
     [y1-y4@0];            ! Intercept means fixed to 0 in
                              all groups
     y1-y4@0;              ! This is a kind of phantom
                              variable, res. Var. is 0
     [R09A$1-R09A$3*]; ! Thresholds here estimated
                              in all groups
     [R09B$1-R09B$3*];
     [R09C$1-R09C$3*];
     [R09D$1-R09D$3*];


      F1 BY y1-y4*;
      F1@1;
    [F1@0];
    {R09A@1 R09B@1 R09C@1 R09D@1};
    [y1-y4@0];
    y1-y4@0;
    [R09A$1-R09A$3*] ;
    [R09B$1-R09B$3*] ;
    [R09C$1-R09C$3*] ;
    [R09D$1-R09D$3*] ;
  MODEL FIN:
      F1 BY y1-y4*;
      F1@1;
    [F1@0];
    {R09A@1 R09B@1 R09C@1 R09D@1};
    [y1-y4@0];
    y1-y4@0;
    [R09A$1-R09A$3*] ;
    [R09B$1-R09B$3*] ;
    [R09C$1-R09C$3*] ;
    [R09D$1-R09D$3*] ;
  OUTPUT:
        tech1 tech4;

  SAVEDATA:
        DIFFTEST = prop4.dif;     ! Save data for chi-
square difference test
```

The `SAVEDATA: DIFFTEST` command produces the necessary information for the chi-square difference test that we will read in the `ANALYSIS` command in the next step of the analyses. That is, when we constrain equal thresholds as in Proposition 4 (Wu & Estabrook, 2016), we will input the file `prop4.dif`.[9]

Mplus: identifying the model with (multiple thresholds) and thresholds invariance

In order to test for threshold invariance and take into account the nested models, we employ the `DIFFTEST` command.

```
  ANALYSIS:
  ESTIMATOR = wlsmv;
  DIFFTEST = prop4.dif; ! File to be considered from pre-
vious step
  H1ITERATIONS = 3000;
```

In terms of model identification restrictions (according to Proposition 4), the intercept mean is fixed to 0 in group 1 [`y1-y4@0`] and estimated in all remaining groups [`y1-y4*`], the variance of Y* is fixed to 1, and for all groups: factor means fixed to zero $\boldsymbol{\kappa}^{(g)} = \mathbf{0}$ [`F1@0`] and factor variance `F1@1` fixed to 1 (diag(($\boldsymbol{\Phi}^{(g)}) = \mathbf{I}$). We put constraints on thresholds via (T1-T3)…(T10-T12).

```
SAVEDATA:
  DIFFTEST = prop7.dif;          ! For chi-square dif-
                                   ference test


MODEL:
   y1 BY R09A@1;                 ! Must always be fixed
                                   to 1
   y2 BY R09B@1;
   y3 BY R09C@1;
   y4 BY R09D@1;
   F1 BYy1-y4*;
   F1@1;                         ! Fixed to 1 in all
                                   groups
  [F1@0];                        ! Factor means fixed
                                   to 0 in all groups
  {R09A@1 R09B@1 R09C@1          ! Fixed to 1 in G1,
  R09D@1};                         estimated in others
  [y1-y4@0];                     ! Fixed to 0 in G1,
                                   estimated in others
  y1-y4@0;
  [R09A$1-R09A$3*] (T1-T3);      ! Thresholds
                                   constrained to
                                   equality
  [R09B$1-R09B$3*] (T4-T6);      ! across all groups
                                   via (T1-T3); etc.
```

[9] We note that the previously used standard chi-square difference test is not appropriate for categorical MG-CFA.

```
  [R09C$1-R09C$3*]              ! There are three
  (T7-T9);                      thresholds for 4
  [R09D$1-R09D$3*]              ! category items
  (T10-T12);

MODEL AUT:
    F1 BYy1-y4*;
    F1@1;
  [F1@0];
  {R09A* R09B* R09C* R09D*};    ! Estimated in all
                                  groups but G1
  [y1-y4*];                     ! Estimated in all
                                  groups but G1
  y1-y4@0;
  [R09A$1-R09A$3*] (T1-T3);     ! Thresholds
                                  constrained to
                                  equality
  [R09B$1-R09B$3*] (T4-T6);     ! across all groups
                                  via (T1-T3); etc.
  [R09C$1-R09C$3*] (T7-T9);     ! There are three
                                  thresholds for 4
  [R09D$1-R09D$3*] (T10-T12);   ! category items

MODEL FIN:
    F1 BYy1-y4*;
    F1@1;
  [F1@0];
  {R09A* R09B* R09C* R09D*};    ! Estimated in all
                                groups but G1
  [y1-y4*];                     ! Estimated in all
                                groups but G1
  y1-y4@0;
  [R09A$1-R09A$3*] (T1-T3);     ! Thresholds
                                constrained to
                                equality
  [R09B$1-R09B$3*] (T4-T6);     ! across all groups
                                via (T1-T3); etc.
  [R09C$1-R09C$3*] (T7-T9);     ! There are three
                                thresholds for 4
  [R09D$1-R09D$3*] (T10-        ! category items
  T12);
```

Here again, we use the SAVEDATA: DIFFTEST command which produces the necessary information for the chi-square difference test that we will read in the ANALYSIS command in the last step of the analysis. That is, when we constrain equal thresholds and loadings as in Proposition 7 (Wu & Estabrook, 2016), we input the file prop7.dif.

Mplus: identifying the model with thresholds and loading invariance with three (or more) thresholds

Model identification restrictions, following Proposition 7 in Wu and Estabrook, states that the intercept mean is fixed to 1 in group 1 [y1-y4@1] and estimated in all remaining groups [y1-y4*], the variance of Y* is fixed to 1, the factor variance F1@1 is fixed to 1 in group 1 only (diag(($\boldsymbol{\Phi}^{(1)}$) = 1), and estimated in all remaining groups F1*; and for all groups factor means are fixed to zero $\boldsymbol{\kappa}^{(g)} = \mathbf{0}$ [F1@0]. The following abridged code identifies the model and places thresholds and loadings equality restrictions.

Once the model is identified, in order to test for threshold and loading invariance and take into account the nested models, we employ the DIFFTEST command, and indicate the file (from previous step) by calling DIFFTEST = prop7.dif. Further, we put equality constraints on loadings via

```
 MODEL:
…
     F1 BY y1-y4* (L1-L4);       ! Constrain
                                   loadings across
                                   groups
     F1@1;                       ! Fixed to 1 in G1
   [F1@0];                       ! Fixed to 0 in all
                                   groups
   {R09A@1 R09B@1 R09C@1         ! Fixed to 1 in G1,
   R09D@1};                        estimated in
                                   others
   [y1-y4@1];                    ! Fixed to 1 in G1,
                                   estimated in
                                   others
   y1-y4@0;
   [R09A$1-R09A$3*] (T1-T3);     ! Thresholds
                                   constrained to
                                   equality
   [R09B$1-R09B$3*] (T4-T6);     ! across all groups
                                   via (T1-T3); etc.
   [R09C$1-R09C$3*] (T7-T9);     ! There are three
                                   thresholds for 4
   [R09D$1-R09D$3*] (T10-T12);   ! category items
```

(L1-L4) and thresholds (as illustrated in previous step) via (T1-T3)…(T10-T12) across all groups.

```
 MODEL AUT:
     F1 BY y1-y4* (L1-L4);       ! Constrain loadings
                                 across groups
     F1*;                        ! Estimated in all
                                 but G1
   [F1@0];
   {R09A* R09B* R09C* R09D*};    ! Estimated in all
                                 but G1
```

```
    [y1-y4*];                    ! Estimated in all
                                   but G1
    y1-y4@0;
    [R09A$1-R09A$3*] (T1-T3);    ! Thresholds
                                   constrained to
                                   equality
    [R09B$1-R09B$3*] (T4-T6);    ! across all groups
                                   via (T1-T3); etc.
    [R09C$1-R09C$3*] (T7-T9);    ! There are three
                                   thresholds for 4
    [R09D$1-R09D$3*] (T10-T12);  ! category items

MODEL FIN:
     F1 BY y1-y4* (L1-L4);       ! Constrain
                                   loadings across
                                   groups;
     F1*;                        ! Estimated in all
                                   but G1
    [F1@0];
    {R09A* R09B* R09C* R09D*};   ! Estimated in all
                                   but G1
    [y1-y4*];                    ! Estimated in all
                                   but G1
    y1-y4@0;
    [R09A$1-R09A$3*] (T1-T3);    ! Thresholds
                                   constrained to
                                   equality
    [R09B$1-R09B$3*] (T4-T6);    ! across all groups
                                   via (T1-T3);
                                   etc.
    [R09C$1-R09C$3*] (T7-T9);    ! There are three
                                   thresholds for 4
    [R09D$1-R09D$3*] (T10-T12);  ! category items
```

lavaan and semTools in R: identifying the baseline model

Use of the lavaan (Rosseel, 2012) and semTools (Jorgensen et al., 2018) packages in R is a convenient way to conduct MG-CFA, even when the data are assumed categorical. A nice feature in the semTools package is that it allows for easy implementation of Wu and Estabrook's model identification and delta parameterization. In the following, we mimic the above M*plus* examples and identify and fit a baseline model, a threshold equality model (Proposition 4) and a threshold and loading equality model (Proposition 7). First, we load the necessary packages in R.

```
library("lavaan")
library("semTools")
```

R allows for several ways to read in the data; in our example, we use the `read.table` function and indicate the path and data file name. The remaining set of commands formats the data file and performs data checks.

```
# '#' is used to make comments in R ('!' In Mplus)
# Read in data file into a created object called dat
# Recall that our data did not have names of variables
dat<-read.table("BULLY.dat", header=FALSE)

# Give names of variables in the BULLY.dat file
# Use of variable IDs as provided in TIMSS 2011
names(dat) <- c("IDCNTRY", "R09A", "R09B", "R09C",
"R09D")

# Check the first few rows of the data
head(dat)
```

We will store results from the baseline, Proposition 4, and Proposition 7 analyses in an empty matrix called `all.results`. For illustration purposes, we will extract chi-square, df, *p*, RMSEA, CFI, and TLI for the three analyses.

```
# Empty matrix of 3 rows (one for baseline, proposition
4, and proposition 7) # and 6 columns (six elements/fit
indices).
# It will be filled in later once results are obtained.

all.results<-matrix(NA, nrow = 3, ncol = 6)
```

In our example, we fit a single factor model with four observed variables. The following code specifies the model which will be used in subsequent analysis of measurement equivalence syntax.

```
# Specifying the baseline model with four items
mod.cat <- 'F1 =~ R09A + R09B + R09C + R09D'

# Baseline model: no constraints across groups or
repeated measures
baseline <- measEq.syntax(configural.model = mod.cat,
                data = dat,
                ordered = c("R09A", "R09B", "R09C",
                "R09D"),
                parameterization = "delta",
                ID.fac = "std.lv",
                ID.cat = "Wu.Estabrook.2016",
                group = "IDCNTRY",
                group.equal = "configural")
```

The `measEq.syntax` is a function within the `semTools` package which automatically generates `lavaan` model syntax for conducting confirmatory factor analysis. As can be seen in the baseline model specification, items are treated as ordered, delta parameterization and Wu and Estabrook's 2016 model identification are employed (other options are available, including Millsap & Yun-Tein, 2004 identification). The grouping variable is IDCNTRY. The next few lines of code are shown to gain additional information regarding model that is then fitted via cfa function in the lavaan package.

```
# For a little bit of orientation/instructions in what model looks like.
summary(baseline)

# To see all of the constraints in the model
cat(as.character(baseline))

# Have to specify as.character to submit to lavaan
model.baseline <- as.character(baseline)

# Fitting baseline model in lavaan via cfa function
fit.baseline <- cfa(model.baseline, data = dat, group = "IDCNTRY",
ordered = c("R09A", "R09B", "R09C", "R09D"))
```

Solutions from fitting the baseline model can be shown via summary function.

```
# Obtaining results from baseline model
summary(fit.baseline)

# Extracting fit indices into the first row of all.results matrix
all.results[1,]<-round(data.matrix(fitmeasures(fit.baseline,
fit.measures = c("chisq.scaled","df.scaled","pvalue.scaled", "rmsea.scaled", "cfi.scaled", "tli.scaled"))), digits=3)
```

LAVAAN AND SEMTOOLS IN R: IDENTIFYING THE MODEL WITH THRESHOLDS INVARIANCE

In order to test for threshold invariance, `group.equal` is changed to "thresholds." The remaining arguments remain the same.

```
# To remain consistent with Wu and Estabrook's (2016) notation, we call this # step as prop4 to indicate the alignment with Proposition 4 in Wu and
# Estabrook's article.
prop4 <- measEq.syntax(configural.model = mod.cat,
          data = dat,
          ordered = c("R09A", "R09B", "R09C", "R09D"),
          parameterization = "delta",
          ID.fac = "std.lv",
          ID.cat = "Wu.Estabrook.2016",
          group = "IDCNTRY",
          group.equal = c("thresholds"))

model.prop4 <- as.character(prop4)

# Fitting thresholds invariance model in lavaan via cfa function
fit.prop4 <- cfa(model.prop4, data = dat, group = "IDCNTRY", ordered = c("R09A", "R09B", "R09C", "R09D"))

# Obtaining results from thresholds invariance model
summary(fit.prop4)

# Extracting fit indices into the second row of all.results matrix
all.results[2,]<-round(data.matrix(fitmeasures(fit.prop4,
fit.measures = c("chisq.scaled","df.scaled","pvalue.scaled", "rmsea.scaled", "cfi.scaled", "tli.scaled"))), digits=3)
```

In order to examine relative model fit and compare the chi-square statistics between baseline model with the model where threshold equality constraints are employed, we use `lavTestLRT` function.

```
lavTestLRT(fit.baseline,fit.prop4)
```

lavaan and semTools in R: identifying the model with thresholds and loading invariance with three (or more) thresholds

While the majority of syntax in R remains the same as above, the main indication where parameter constraints are placed is again found in `group.equal =`, where now we indicate `c("thresholds", "loadings")` to be constrained to equality.

```
# Proposition 7 per Wu and Estabrook (2016)
prop7 <- measEq.syntax(configural.model = mod.cat,
            data = dat,
            ordered = c("R09A", "R09B", "R09C", "R09D"),
            parameterization = "delta",
            ID.fac = "std.lv",
            ID.cat = "Wu.Estabrook.2016",
            group = "IDCNTRY",
            group.equal = c("thresholds", "loadings"))

model.prop7 <- as.character(prop7)

fit.prop7 <- cfa(model.prop7, data = dat1, group = "IDCNTRY", ordered = c("R09A", "R09B", "R09C", "R09D"))

summary(fit.prop7)

# Extracting fit indices into the third row of all.results matrix
all.results[3,]<-round(data.matrix(fitmeasures(fit.prop7,
fit.measures = c("chisq.scaled","df.scaled","pvalue.scaled", "rmsea.scaled", "cfi.scaled", "tli.scaled"))), digits=3)
```

Examining the fit indices (all.results), we note that in general, model fit worsened as models were constrained by imposing the equality of thresholds (prop4) and thresholds and loadings (prop7).

```
   chisq.scaled df.scaled pvalue.scaled rmsea.scaled
cfi.scaled tli.scaled

baseline    50.944    6   0   0.042   0.997   0.991
prop4      111.985   14   0   0.041   0.993   0.992
prop7      210.644   20   0   0.047   0.987   0.989
```

We can conduct chi-square difference test between the models that put equality constraints of thresholds (proposition 4) and thresholds and loadings (proposition 7) to evaluate attainability of ME/I.

```
lavTestLRT(fit.prop4, fit.prop7)
```

In order to provide a comparison in comparability of the above-discussed analyses, we provide Tables 2 (model estimates) and 3 (model fit). As it can be noted, slight differences can be observed – for example, in Table 2, estimated parameter estimates are identical or nearly identical (few differences can be observed at the third decimal place). Similarly, only slightly larger difference can be observed in fit indices (see Table 3). Taken together, we consider the outputs from the two programs to be essentially equivalent. Additionally, in the Appendix, we annotate a sample output file from R and connect it to the M*plus* code. We discuss results with respect to relevant model parameter estimates and model fit.

PARTIAL INVARIANCE

The main purpose of this tutorial was to demonstrate various ways to conduct measurement invariance using an MG-CFA approach while taking into consideration categorical outcomes, such as those potentially found on ILSAs. In our presentation, we have restricted our demonstration to measurement invariance using the MG-CFA approach when testing for different levels of non-invariance. We purposefully chose to focus on this approach because such an approach is prevalent in operational testing and practice involving ILSAs and other surveys. However, when thresholds and loadings equality constraints yield inadequate model fit, researchers may resort to alternative methods to examine ME/I (see next two sections). One such way is what is known in the literature as partial invariance. We briefly describe this approach next and provide M*plus* and R code associated with the steps.

WHEN ME/I FAILS: AN ILLUSTRATION OF PARTIAL INVARIANCE

The partial invariance example is taken based on a revision to the model of threshold and loading invariance that is guided by modification indices. We begin with the most stringent model and request modification indices, focusing only on loadings, as we retained a model of equal thresholds.

The syntax in M*plus* to invoke the search for partial measurement invariance is given by modindices(3.84) in the OUTPUT command.

```
OUTPUT:
   tech1 tech4 modindices(3.84);
```

Reviewing the results, we obtain:

```
Group AZE

BY Statements


Y1    BY R09A    15.285     0.252     0.183     0.183
Y1    BY R09C    90.471    -0.629    -0.457    -0.457
Y1    BY R09D    17.423     0.262     0.190     0.190
Y2    BY R09A    15.285     0.259     0.183     0.183
Y2    BY R09C    90.471    -0.648    -0.457    -0.457
Y2    BY R09D    17.423     0.270     0.190     0.190
…
```

We see that freeing the loading for R09C produces the most improvement in fit. In M*plus* these values are MI_{AZE} = 90.471, $MI_{\text{AUT}} < 3.84$, and MI_{FIN} = 40.327.

Using lavaan in R, we utilize `modindices` function to seek where the modification in terms of freeing parameters ought to occur in order to improve model fit.

```
# Here, we fix threshold equality (maintaining prop 4)
and freely
# estimate a loading (for R09C) based on modification
indices from prop 7
fit.prop7 <- cfa(model.prop7, data = dat, group =
"IDCNTRY", ordered = c("R09A", "R09B", "R09C",
"R09D"))

summary(fit.prop7)

mi <- modindices(fit.prop7, free.remove = FALSE)
mi[mi$op == "=~",]

# The model of partial invariance
prop7.part <- measEq.syntax(configural.model = mod.cat,
        data = dat,
        ordered = c("R09A", "R09B", "R09C", "R09D"),
        parameterization = "delta",
```

TABLE 2
Estimated (Selected) Parameters across the Two Programs

Panel (a) Baseline Model								
Group AZE			*Group AUT*			*Group FIN*		
F1 BY	*M*plus	*R*		*M*plus	*R*		*M*plus	*R*
Y1	.746	.746	Y1	.765	.765	Y1	.773	.773
Y2	.714	.714	Y2	.736	.736	Y2	.701	.701
Y3	.720	.720	Y3	.792	.792	Y3	.784	.784
Y4	.753	.753	Y4	.529	.529	Y4	.519	.519
Thresholds								
R09A$1	.788	.788	R09A$1	-.200	-.200	R09A$1	-.167	-.167
R09A$2	1.023	1.023	R09A$2	.385	.385	R09A$2	.688	.688
R09A$3	1.236	1.236	R09A$3	.795	.795	R09A$3	1.274	1.274
R09B$1	.802	.802	R09B$1	.245	.245	R09B$1	.181	.181
R09B$2	1.021	1.021	R09B$2	.740	.740	R09B$2	1.023	1.023
R09B$3	1.309	1.309	R09B$3	1.167	1.167	R09B$3	1.591	1.591
R09C$1	.633	.633	R09C$1	-.035	-.035	R09C$1	.286	.286
R09C$2	1.029	1.029	R09C$2	.553	.553	R09C$2	1.043	1.043
R09C$3	1.358	1.358	R09C$3	.969	.969	R09C$3	1.680	1.680
R09D$1	1.021	1.021	R09D$1	.554	.554	R09D$1	.713	.713
R09D$2	1.347	1.347	R09D$2	1.148	1.148	R09D$2	1.551	1.551
R09D$3	1.628	1.628	R09D$3	1.423	1.423	R09D$3	1.971	1.971
Residual Variance*								
R09A	.444	.443	R09A	.415	.415	R09A	.403	.403
R09B	.490	.490	R09B	.459	.459	R09B	.509	.509
R09C	.482	.482	R09C	.372	.372	R09C	.386	.386
R09D	.433	.433	R09D	.720	.720	R09D	.731	.731
Panel (b) Equality of Thresholds (Proposition 4)								
Y1	.746	.746	Y1	.340	.339	Y1	.237	.236
Y2	.714	.714	Y2	.394	.395	Y2	.241	.241
Y3	.720	.720	Y3	.568	.568	Y3	.409	.409
Y4	.753	.753	Y4	.349	.349	Y4	.238	.237
Thresholds								
R09A$1	.781	.782	R09A$1	.781	.782	R09A$1	.781	.782
R09A$2	1.042	1.042	R09A$2	1.042	1.042	R09A$2	1.042	1.042
R09A$3	1.224	1.223	R09A$3	1.224	1.223	R09A$3	1.224	1.223
R09B$1	.786	.786	R09B$1	.786	.786	R09B$1	.786	.786
R09B$2	1.064	1.064	R09B$2	1.064	1.064	R09B$2	1.064	1.064
R09B$3	1.277	1.277	R09B$3	1.277	1.277	R09B$3	1.277	1.277
R09C$1	.631	.631	R09C$1	.631	.631	R09C$1	.631	.631
R09C$2	1.035	1.035	R09C$2	1.035	1.035	R09C$2	1.035	1.035
R09C$3	1.354	1.353	R09C$3	1.354	1.353	R09C$3	1.354	1.353
R09D$1	1.010	1.010	R09D$1	1.010	1.010	R09D$1	1.010	1.010
R09D$2	1.390	1.390	R09D$2	1.390	1.390	R09D$2	1.390	1.390
R09D$3	1.589	1.588	R09D$3	1.589	1.588	R09D$3	1.589	1.588

Residual Variance								
R09A	.443	.443	R09A	.082	.082	R09A	.038	.038
R09B	.490	.490	R09B	.132	.132	R09B	.060	.060
R09C	.482	.482	R09C	.191	.191	R09C	.105	.105
R09D	.433	.433	R09D	<u>.313</u>	<u>.312</u>	R09D	.153	.153
Panel (c) Equality of Thresholds and Loadings (Proposition 7)								
Y1	.726	.726	Y1	.726	.726	Y1	.726	.726
Y2	.706	.706	Y2	.706	.706	Y2	.706	.706
Y3	.764	.764	Y3	.764	.764	Y3	.764	.764
Y4	.731	.731	Y4	.731	.731	Y4	.731	.731
Thresholds								
R09A$1	.748	.748	R09A$1	.748	.748	R09A$1	.748	.748
R09A$2	1.050	1.050	R09A$2	1.050	1.050	R09A$2	1.050	1.050
R09A$3	1.261	1.261	R09A$3	1.261	1.261	R09A$3	1.261	1.261
R09B$1	.772	.772	R09B$1	.772	.772	R09B$1	.772	.772
R09B$2	1.067	1.067	R09B$2	1.067	1.067	R09B$2	1.067	1.067
R09B$3	1.292	1.292	R09B$3	1.292	1.292	R09B$3	1.292	1.292
R09C$1	.705	.705	R09C$1	.705	.705	R09C$1	.705	.705
R09C$2	1.012	1.012	R09C$2	1.012	1.012	R09C$2	1.012	1.012
R09C$3	1.256	1.256	R09C$3	1.256	1.256	R09C$3	1.256	1.256
R09D$1	.975	.975	R09D$1	.975	.975	R09D$1	.975	.975
R09D$2	1.406	1.406	R09D$2	1.406	1.406	R09D$2	1.406	1.406
R09D$3	1.633	1.633	R09D$3	1.633	1.633	R09D$3	1.633	1.633
Residual Variance								
R09A	.472	.472	R09A	.112	.112	R09A	.050	.050
R09B	.502	.502	R09B	.142	.142	R09B	.068	.068
R09C	.417	.417	R09C	.111	.111	R09C	.061	.061
R09D	.466	.466	R09D	.419	.419	R09D	.202	.202

Note: *as M*plus*, these values can be found under R-Square section of the output. Underlined values point to different estimates between the two programs.

TABLE 3
Fit Indices Results Based on Two Softwares/Programs for Ordinal Data across Baseline and Equality Constraints Models (Thresholds; Thresholds and Loadings)

	M*plus 7.2*			*lavaan in R*		
Fit Indices/Equality Constraints	*Baseline*	*Thresholds*	*Thresholds and Loadings*	*Baseline*	*Thresholds*	*Thresholds and Loadings*
χ^2	50.96	107.87	186.59	50.94	111.98	210.64
χ^2 *df*	6	14	20	6	14	20
χ^2 *p* value	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
RMSEA	.042	.040	.044	.042	.041	.047
CFI	.997	.994	. 989	.997	.993	.987
TLI	.991	.992	. 990	.991	.992	.989

```
        ID.fac = "std.lv",
        ID.cat = "Wu.Estabrook.2016",
        group = "IDCNTRY",
        group.equal = c("thresholds", "loadings"),
        group.partial = "F1 =~ R09C")

model.prop7.part <- as.character(prop7.part)

fit.prop7.part <- cfa(model.prop7.part, data = dat,
group = "IDCNTRY", ordered = c("R09A", "R09B",
"R09C", "R09D"))

summary(fit.prop7.part)

# Test of model fit between prop4 and prop 7 with one
loading freed
lavTestLRT(fit.prop7.part, fit.prop4)
```

Results in lavaan were consistent; however, the values of the modification indices were different: $MI_{\mathrm{AZE}} = 2.698$, $MI_{\mathrm{AUT}} = .008$, and $MI_{\mathrm{FIN}} = 1.164$. Freely estimating this loading in a subsequent model produced a non-significant chi-square difference test ($\chi^2_4 = 7.849$, $p = .10$; $\chi^2_4 = 6.813$, $p = .15$, in M*plus* and *lavaan*, respectively). The substantial discrepancy between the modification index values and the chi-square difference tests in M*plus* is due to the fact that the chi-square difference test is adjusted to account for the fact that the data are ordinal.

A BRIEF NOTE ON LATEST DEVELOPMENTS IN ME/I APPROACHES

In addition to a partial invariance approach, researchers proposed several methods to deal with failing to achieve ME/I. In what follows, we briefly elude to several studies that attempted to systematically examine ME/I (Asparouhov & Muthén, 2014; Finch & French, 2018; Pokropek, Davidov, & Schmidt, 2019; Raykov et al., 2012) that utilize/propose methods and approaches that show promise in the area of ME/I.

The alignment method (Asparouhov & Muthén, 2014), a relatively new approach to MG-CFA, offers an alternative to traditional methods of establishing measurement equivalence. The alignment method, counter to typical MG-CFA, does not assume measurement invariance. Rather, the method identifies an optimal solution that minimizes parameter invariance across groups. A feature of the alignment method is that the solution will exhibit identical fit to the configural or baseline solution while estimating all of the model parameters (factor means, factor variances, loadings, intercepts/thresholds, and residual variance). This is in contrast to the configural invariance model, which assumes that latent variable means and variances are fixed to values of 0 and 1, respectively. This sort of latent variable standardization implies that the latent variables are not on the same scale and, as a result, cannot be compared. Although the alignment method is a practical way to overcome problems associated with testing for parameter equality when the number of comparison groups is large, the method is primarily exploratory in nature, and as suggested in Pokropek et al. (2019) study, its performance varies across conditions.

Raykov et al. (2012) outlined a multiple testing procedure suitable for examining ME/I that uses Benjamini–Hochberg (BH) false discovery rate method and which controls the overall family-wise error rate at a preset significance level.[10] As the authors stated, the BH approach is flexible in that it allows for overall evaluation of the ME/I as well as it permits for testing different, more localized, levels of invariance (i.e., testing for loadings or intercepts invariance only).

Recently, Pokropek et al. (2019) investigated traditional and newer approaches to test for ME/I. In a large simulation study, the authors examined performance of five methods used to test for ME/I, including MG-CFA, partial MG-CFA, multigroup Bayesian structural equation modeling (SEM), partial multigroup Bayesian SEM, and MG-CFA

[10] Conceptually, this is the same idea as we would have in multiple groups mean comparisons and using a statistical omnibus test to control for multiple pair-wise comparisons.

with alignment optimization. Overall conclusions by Pokropek et al. can be summarized as following: partial measurement invariance may be a suitable (effective) method to recover path coefficients and latent means when many items are noninvariant; approximate measurement invariance models may be more appropriate to use in recovering latent means when many parameters are approximately (but not exactly) equal, while the alignment method might be appropriate for recovering latent means when only a few noninvariant parameters are present. As the authors suggested, future research is warranted in terms of better convergence and more efficient algorithms for some of these methods (in particular those that are more flexible).

Lastly, Finch and French (2018) examined the utility of the RMSEA equivalence testing approach described by Yuan and Chan (2016), Marcoulides and Yuan (2017), and Yuan, Chan, Marcoulides, and Bentler (2016), which showed promise in Yuan and Chan's study with a single dataset. The proposed equivalence testing approach came partly as a response to inadequacy of the chi-square difference test in controlling for Type I (and II) error(s). Thus, Finch and French systematically examined its performance via a simulation study, where both metric and scalar invariance was examined (i.e., noninvariance was modeled to be present in loadings and intercepts, respectively). The authors found support for using the equivalence approach for models that assume indicators to be normally distributed, and more importantly point to its ability to provide useful information regarding the degree to which invariance is present or lacking. This, as the authors state, is in contrast to more traditional approaches to ME/I, which typically arrive to conclusions of presence or absence of noninvariance.

CONCLUSION

In summary, the current tutorial aimed to be didactic for researchers who wish to make meaningful comparisons across groups, and who thus engage in establishing ME/I. We close by briefly mentioning several important issues that should be considered when an analyst conducts ME/I investigations. We direct a reader to Vandenberg and Lance (2000), who provided a comprehensive review of the literature and address some of these issues at greater length. One issue is controlling the Type I error rate and power. As Raykov et al. (2012) demonstrated, when conducting ME/I, it is important to control the overall significance error level associated with the multiple tests. Additionally, Finch, French, and Finch (2018) examined the performance of different estimators (maximum likelihood [ML], Bayesian, and generalized structured components analysis [GSCA]) in testing for metric invariance (noninvariance in loadings only) for small sample sizes and skewed latent trait distributions. The authors found that in very small sample sizes (n = 25), a trade-off between Type I error rate and power may need to be considered, and while in larger than 25 sample size, ML estimator performed reasonably well, convergence issues may make ML not as optimal option. Thus, Bayes (maintains Type I error rate but lower in power) or GSCA (higher power but inflated Type I errors) might need to be considered.

Secondly, an analyst should consider how a model is identified as often a typical choice is to use a reference variable. The choice in using a reference variable and fixing it to 1 for model identification purposes can produce problems when the item used as a referent indicator is an 'offending' item. Serious limitations of using a reference variable have been documented in the literature (e.g., Raykov, Marcoulides, & Li, 2012; Raykov et al., 2012; Vandenberg & Lance, 2000). For example, the arbitrary choice of a reference item may indeed yield poor partial invariance metric model fit if that item may indeed exhibit noninvariance but by default was fixed to equality. To deal with this, some researchers have put forward approaches – for example, previously mentioned Raykov and his colleagues (Raykov et al., 2012) and Pokropek et al. (2019), or as used in the current study, approach by Wu and Estabrook (2016) whose approach circumvents the issue of fixing an item's loading by identifying models for categorical data in different ways. Lastly, collapsing categories when data are treated ordinal may occur for substantive or methodological reasons. As Rutkowski, Svetina, and Liaw (2019) suggested, categorical MG-CFA approach to establishing ME/I typically requires that the number of categories is the same across groups. However, categories may be collapsed, for instance, due to low-frequency counts in some categories for some items or groups (e.g., strongly disagree and disagree are collapsed into one category). As Rutkowski et al. (2019) suggested, such collapsing can have meaningful impacts on model fit in terms of reduced scale reliability and what the authors termed as artifactual fit improvement. The above-mentioned issues are briefly addressed here to acknowledge that as researchers, we ought to consider a variety of issues when engaging in modeling that ultimately leads to cross-cultural comparisons.

FUNDING

This work was supported by the Norges Forskningsråd [255246].

REFERENCES

Asparouhov, T., & Muthén, B. O. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508. doi:10.1080/10705511.2014.919210

Bagozzi, R. P. (1977). Structural equation models in experimental research. *Journal of Marketing Research*, *14*(2), 209–226. doi:10.2307/3150471

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(2), 186–203. doi:10.1207/s15328007sem1302_2

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. Retrieved from 10.1037/0033-2909.107.2.238

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606. doi:10.1037/0033-2909.88.3.588

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. doi:10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. doi:10.1207/S15328007SEM0902_5

Deshon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science*, *46*(1), 137–149.

Finch, H. W., French, B. F., & Finch, M. E. (2018). Comparison of methods for factor invariance testing of a 1-factor model with small samples and skewed latent traits. *Frontiers in Psychology*, *9*, 332. doi:10.3389/fpsyg.2018.00332

Finch, W. H., & French, B. F. (2018). A simulation investigation of the performance of invariance assessment using equivalence testing procedures. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 673–686. doi:10.1080/10705511.2018.1431781

Fox, K. R., & Corbin, C. (1989). Physical self-perception profile: Development and preliminary validation. *Journal of Sport and Exercise Psychology*, *11*, 408–430.

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(3), 378–402. doi:10.1207/s15328007sem1303_3

Hagger, M. S., Biddle, S. J. H., Chow, E. W., Stambulova, N., & Kavussanu, M. (2003). Physical self perceptions in adolescence: Generalizability of a hierarchical multidimensional model across three cultures. *Journal of Cross-Cultural Psychology*, *34*, 611–628. doi:10.1177/0022022103255437

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT Area and Mantel-Haenszel methods. *Applied Measurement in Education*, *2*(4), 313–334. doi:10.1207/s15324818ame0204_4

Hancock, G. R. (1997). Structural equation modeling methods of hypothesis testing of latent variable means. *Measurement & Evaluation in Counseling & Development (American Counseling Association)*, *30*(2), 91–105.

Horn, J. L., McArdle, J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist*, *1*(4), 179–188.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117–144. doi:10.1080/03610739208253914

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–426. doi:10.1007/BF02291366

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2018). semTools: Useful tools for structural equation modeling. R package version 0.5-1. Retrieved from https://CRAN.R-project.org/package=semTools

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*, 524–544. doi:10.1080/10705511.2017.1304822

Little, T. D. (1997). Mean and Covariance Structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, *32*(1), 53–76. doi:10.1207/s15327906mbr3201_3

Lord, F. M. (1980). *Applications of item response to theory to practical testing problems*. New York, NY: Routledge.

Lubke, G. H., & Dolan, C. V. (2003). Can unequal residual variances across groups mask differences in residual means in the common factor model?. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(2), 175–192. doi:10.1207/S15328007SEM1002_1

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(4), 514–534. doi:10.1207/s15328007sem1104_2

Marcoulides, K. M., & Yuan, K.-H. (2017). New ways to evaluate goodness of fit: A note on using equivalence testing to assess structural equation models. *Structural Equation Modeling. A Multidisciplinary Journal*, *24*, 148–153.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*(3), 568–592. doi:10.1037/0021-9010.93.3.568

Megreya, A. M., Latzman, R. D., Al-Attiyah, A. A., & Alrashidi, M. (2016). The robustness of the nine-factor structure of the Cognitive Emotion Regulation Questionnaire across four Arabic-speaking Middle Eastern countries. *Journal of Cross-Cultural Psychology*, *47*(6), 875–890. doi:10.1177/0022022116644785

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, *115*, 300–307. doi:10.1037/0033-2909.115.2.300

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. doi:10.1007/BF02294825

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479–515. doi:10.1207/S15327906MBR3903_4

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. Boston, MA: TIMSS & PIRLS International Study Center.

Muthén, B., & 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*, 171–189. doi:10.1111/bmsp.1985.38.issue-2

Muthén, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus* (No. 4). Los Angeles: University of California, Los Angeles.

Muthén, B. O., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, *46*(4), 407–419. doi:10.1007/BF02293798

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide* (Eighth ed.). Los Angeles, CA: Muthén & Muthén.

OECD. (2014). *TALIS 2013 technical report*. Paris, France: Author. Retrieved from http://www.oecd.org/edu/school/TALIS-technical-report-2013.pdf

OECD. (2016). PISA 2015 results in focus. OECD Publishing. Retrieved from https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf

Pokropek, A., Davidov, E., & Schmidt, P. (2019). A Monte Carlo Simulation Study to Assess The Appropriateness of Traditional and Newer Approaches to Test for Measurement Invariance. Advanced online publication. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–21. doi:10.1080/10705511.2018.1561293

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Raykov, T., Marcoulides, G. A., & Li, C. (2012). Measurement invariance for latent constructs in multiple populations: A critical view and refocus. *Educational and Psychological Measurement*, *72*, 954–974. doi:10.1177/0013164412441607

Raykov, T., Marcoulides, G. A., & Millsap, R. E. (2012). Factorial invariance in multiple populations: A multiple testing procedure. *Educational and Psychological Measurement*, *73*, 713–727. doi:10.1177/0013164412451978

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. doi:10.18637/jss.v048.i02

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*(1), 31–57. doi:10.1177/0013164413498257

Rutkowski, L., & Svetina, D. (2017). Measurement invariance in international surveys: Categorical indicators & fit measure performance. *Applied Measurement in Education*, *30*(1), 39–51. doi:10.1080/08957347.2016.1243540

Rutkowski, L., Svetina, D., & Liaw, Y.-L. (2019). Collapsing Categorical Variables and Measurement Invariance. Online advanced publication. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–13. doi:10.1080/10705511.2018.1547640

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

Svetina, D., & Rutkowski, L. (2017). Multidimensional measurement invariance in an international context: Fit measure performance with many groups. *Journal of Cross-Cultural Psychology*, *48*(7), 991–1008. doi:10.1177/0022022117717028

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370. doi:10.1111/jedm.1990.27.issue-4

Thompson, M. S., & Green, S. B. (2006). Evaluating between-group differences in latent variable means. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling: A second course* (1st ed., pp. 119–169). New York, NY: IAP.

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*(1), 4–69. doi:10.1177/109442810031002

Wu, H., & Estabrook, R. (2016). Identification of Confirmatory Factor Analysis Models of Different Levels of Invariance for Ordered Categorical Outcomes. *Psychometrika*, *81*(4), 1014–1045. doi:10.1007/s11336-016-9506-0

Yuan, K.-H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square difference tests. *Psychological Methods*, *21*(3), 405–426. doi:10.1037/met0000080

Yuan, K.-H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indexes. *Structural Equation Modeling*, *23*(3), 319–330. doi:10.1080/10705511.2015.1065414

APPENDIX: ANNOTATED OUTPUT EXAMPLE

```
> summary(fit.baseline)
lavaan 0.6-3 ended normally after 16 iterations

  Optimization method                           NLMINB
  Number of free parameters                         48

  Number of observations per group
  31                                              3808
  40                                              4457
  246                                             4520

  Estimator                                       DWLS      Robust
  Model Fit Test Statistic                      26.941      50.944
  Degrees of freedom                                 6           6
  P-value (Chi-square)                           0.000       0.000
  Scaling correction factor                                  0.529
  Shift parameter for each group:
    31                                                       0.013
    40                                                       0.015
    246                                                      0.016
    for simple second-order correction (Mplus variant)

Chi-square for each group:

  31                                             8.247      15.593
  40                                            10.577      19.998
  246                                            8.118      15.353

Parameter Estimates:

  Information                                 Expected
  Information saturated (h1) model        Unstructured
  Standard Errors                           Robust.sem


Group 1 [31]:

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  F1 =~
    R09A    (l.1_)    0.746    0.018   41.589    0.000
    R09B    (l.2_)    0.714    0.019   38.354    0.000
    R09C    (l.3_)    0.720    0.017   41.671    0.000
    R09D    (l.4_)    0.753    0.020   37.765    0.000

Intercepts:
                   Estimate  Std.Err  z-value  P(>|z|)
   .R09A    (n.1.)    0.000
   .R09B    (n.2.)    0.000
   .R09C    (n.3.)    0.000
   .R09D    (n.4.)    0.000
    F1      (a.1.)    0.000

Thresholds:
                   Estimate  Std.Err  z-value  P(>|z|)
    R09A| (R09A.1)    0.788    0.023   34.594    0.000
    R09A| (R09A.2)    1.023    0.025   41.441    0.000
    R09A| (R09A.3)    1.236    0.027   45.627    0.000
    R09B| (R09B.1)    0.802    0.023   35.047    0.000
    R09B| (R09B.2)    1.021    0.025   41.387    0.000
    R09B| (R09B.3)    1.309    0.028   46.591    0.000
    R09C| (R09C.1)    0.633    0.022   28.950    0.000
    R09C| (R09C.2)    1.029    0.025   41.575    0.000
    R09C| (R09C.3)    1.358    0.029   47.115    0.000
    R09D| (R09D.1)    1.021    0.025   41.387    0.000
    R09D| (R09D.2)    1.347    0.029   47.003    0.000
    R09D| (R09D.3)    1.628    0.034   48.069    0.000

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
```

NOTES: Baseline Model in R
(()) double parenthesis connect to *Mplus* code and output. The first part shows estimates and Chi-square, df, and *p*.
In the second part, we show model fit statistics.

Robust Chi-square test statistic

Here we can see contribution of each group to the Chi-square. We note that each group contributed relatively similarly to Chi-square. Large deviations among the groups may indicate potential problems in some populations.

Model **Estimates** for Group 1 (ID 31, which in our example was Azerbaijan)

Estimates of Loadings for 4 items (item IDs R09A-R09D) for each group separately. In prop7, these are constrained to be equal.

Per Identification, we fixed latent intercepts to 0 in each group.
((Mplus [y1-y4@0];))

F1 mean also fixed to 0.
((Mplus F1@1;))

Thresholds are freely estimated in baseline model. In prop4 (and prop7), these are constrained to be equal across groups.
((Mplus [R09A$1-R09A$3*];
[R09B$1-R09B$3*];
[R09C$1-R09C$3*];
[R09D$1-R09D$3*];))

Estimates of Thresholds for 4 items and three categories (R09A.1.. R09D.3) for each group separately → these values differ from the ones in Group 1 and 3. (in baseline model, loadings and thresholds are freely estimated).

```
    F1       (p.1_)    1.000
   .R09A               0.443
   .R09B               0.490
   .R09C               0.482
   .R09D               0.433
```

Variance of F1 fixed to 1 in all groups. Residual variances estimated in each group. In Mplus output, these estimates are located under ((R-SQUARE under Observed Variable Residual Variance))

```
Scales y*:
                   Estimate  Std.Err  z-value  P(>|z|)
    R09A              1.000
    R09B              1.000
    R09C              1.000
    R09D              1.000
```

Fixed scale of phantom variable to 1 in all groups. In Mplus, under SCALES.

```
Group 2 [40]:

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  F1 =~
    R09A    (l.1_)    0.765    0.012   65.439    0.000
    R09B    (l.2_)    0.736    0.013   58.427    0.000
    R09C    (l.3_)    0.792    0.011   69.287    0.000
    R09D    (l.4_)    0.529    0.017   30.824    0.000
```

Estimates of Loadings for 4 items (item IDs R09A-R09D) for each group separately → these values differ from the ones in Group 1 and 3. (in baseline model, loadings and thresholds are freely estimated).

```
Intercepts:
                   Estimate  Std.Err  z-value  P(>|z|)
   .R09A    (n.1.)    0.000
   .R09B    (n.2.)    0.000
   .R09C    (n.3.)    0.000
   .R09D    (n.4.)    0.000
    F1      (a.1.)    0.000

Thresholds:
                   Estimate  Std.Err  z-value  P(>|z|)
    R09A| (R09A.1)   -0.200    0.019  -10.582    0.000
    R09A| (R09A.2)    0.385    0.019   19.973    0.000
    R09A| (R09A.3)    0.795    0.021   37.671    0.000
    R09B| (R09B.1)    0.245    0.019   12.913    0.000
    R09B| (R09B.2)    0.740    0.021   35.646    0.000
    R09B| (R09B.3)    1.167    0.024   48.128    0.000
    R09C| (R09C.1)   -0.035    0.019   -1.842    0.065
    R09C| (R09C.2)    0.553    0.020   27.853    0.000
    R09C| (R09C.3)    0.969    0.022   43.342    0.000
    R09D| (R09D.1)    0.554    0.020   27.882    0.000
    R09D| (R09D.2)    1.148    0.024   47.755    0.000
    R09D| (R09D.3)    1.423    0.028   51.531    0.000
```

Estimates of Thresholds for 4 items and three categories (R09A.1.. R09D.3) for each group separately → these values differ from the ones in Group 1 and 3. (in baseline model, loadings and thresholds are freely estimated).

```
Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
    F1      (p.1_)    1.000
   .R09A              0.415
   .R09B              0.459
   .R09C              0.372
   .R09D              0.720

Scales y*:
                   Estimate  Std.Err  z-value  P(>|z|)
    R09A              1.000
    R09B              1.000
    R09C              1.000
    R09D              1.000


Group 3 [246]:

Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)
  F1 =~
    R09A    (l.1_)    0.773    0.012   62.264    0.000
    R09B    (l.2_)    0.701    0.013   52.785    0.000
```

Estimates of Loadings for 4 items (item IDs R09A-R09D) for each group separately → these values

```
    R09C    (l.3_)    0.784    0.012   64.649    0.000
    R09D    (l.4_)    0.519    0.018   28.119    0.000

Intercepts:
                   Estimate  Std.Err  z-value  P(>|z|)
   .R09A    (n.1.)    0.000
   .R09B    (n.2.)    0.000
   .R09C    (n.3.)    0.000
   .R09D    (n.4.)    0.000
    F1      (a.1.)    0.000
Thresholds:
                   Estimate  Std.Err  z-value  P(>|z|)
    R09A| (R09A.1)   -0.167    0.019   -8.920    0.000
    R09A| (R09A.2)    0.688    0.020   33.845    0.000
    R09A| (R09A.3)    1.274    0.025   50.290    0.000
    R09B| (R09B.1)    0.181    0.019    9.632    0.000
    R09B| (R09B.2)    1.023    0.023   45.149    0.000
    R09B| (R09B.3)    1.591    0.030   52.429    0.000
    R09C| (R09C.1)    0.286    0.019   15.090    0.000
    R09C| (R09C.2)    1.043    0.023   45.666    0.000
    R09C| (R09C.3)    1.680    0.032   52.187    0.000
    R09D| (R09D.1)    0.713    0.020   34.838    0.000
    R09D| (R09D.2)    1.551    0.030   52.422    0.000
    R09D| (R09D.3)    1.971    0.040   49.144    0.000
Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
    F1      (p.1_)    1.000
   .R09A              0.403
   .R09B              0.509
   .R09C              0.386
   .R09D              0.731

Scales y*:
                   Estimate  Std.Err  z-value  P(>|z|)
    R09A              1.000
    R09B              1.000
    R09C              1.000
    R09D              1.000
```

differ from the ones in Group 1 and 3. (in baseline model, loadings and thresholds are freely estimated).

Estimates of Thresholds for 4 items and three categories (R09A.1.. R09D.3) for each group separately → these values differ from the ones in Group 1 and 3. (in baseline model, loadings and thresholds are freely estimated).

```
> all.results[1,]<-
round(data.matrix(fitmeasures(fit.baseline,fit.measures =
c("chisq.scaled","df.scaled","pvalue.scaled", "rmsea.scaled",
"cfi.scaled", "tli.scaled"))), digits=3)
> all.results
       [,1] [,2] [,3]  [,4]  [,5]  [,6]
[1,] 50.944    6    0 0.042 0.997 0.991
[2,]     NA   NA   NA    NA    NA    NA
[3,]     NA   NA   NA    NA    NA    NA

...
> all.results
         chisq.scaled df.scaled pvalue.scaled rmsea.scaled cfi.scaled tli.scaled
baseline       50.944         6             0        0.042      0.997      0.991
prop4         111.985        14             0        0.041      0.993      0.992
prop7         210.644        20             0        0.047      0.987      0.989
```

all.results is a matrix that we created to store selected model fit indices.
all.results[1,] – [1,] represents the first row to be populated, with fit measures for fit.baseline model.
50.944 is the same as the Robust chi-square value as noted above.

After all three models (baseline, imposing threshold equality, and imposing threshold and loading equality), we obtain the following results.
Here we see that overall, model fit worsens going from baseline to model with threshold equalities (albeit only slightly and for CFI only which went down from .997 to .993). TLI and RMSEA actually improved a little bit, going from .991 to .992 and .042 to .041, respectively. Imposing restrictions on thresholds and loadings (last row under prop7), we see that this model fit the worse, with largest chi-square and RMSEA values, alongside lowest CFI and TLI.
In order to evaluate tenability of ME/I, a researcher can compare these values to cuff-off values as presented in Table 1 in the main text.