NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Boris Sokolov

SENSITIVITY OF GOODNESS OF FIT INDICES TO LACK OF MEASUREMENT INVARIANCE WITH CATEGORICAL INDICATORS AND MANY GROUPS

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: SOCIOLOGY
WP BRP 86/SOC/2019

Boris Sokolov[1]

SENSITIVITY OF GOODNESS OF FIT INDICES TO LACK OF MEASUREMENT INVARIANCE WITH CATEGORICAL INDICATORS AND MANY GROUPS[2]

Using Monte Carlo simulation experiments, this paper examines the performance of popular SEM goodness-of-fit indices, namely CFI, TLI, RMSEA, and SRMR, with respect to a specific task of measurement invariance testing with categorical data and many groups (10-50 groups). Study factors include the number of groups, the level of non-invariance in the data, and the absence/presence of model misspecifications other than non-invariance. In sum, the study design yields a total of 81 conditions. All simulated data sets are analyzed using two popular SEM estimators, MLR and WLSMV. The main contribution of this paper to the methodological literature on cross-cultural survey research is that it produces revised guidelines for evaluating the goodness of fit of invariance MGCFA models with many groups.

JEL classification: C15

Keywords: SEM; MGCFA; measurement invariance; fit indices, simulations

[1] Research fellow, Laboratory for Comparative Social Research, National Research University Higher School of Economics. E-mail: bssokolov@gmail.com

[2]The study was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy by the Russian Academic Excellence Project '5-100'.

Introduction

Comparability of measurements, or measurement invariance/equivalence, across study units is an important concern in a wide range of social scientific disciplines where cross-national comparisons involving many countries or groups are of substantive interest, including, among others, comparative politics (Przeworski and Teune 1966; Stegmueller 2011), comparative sociology (Davidov et al. 2014), cross-cultural psychology (Poortinga 1989; Little 1997), educational assessment (Wicherts and Dolan 2010), consumer research (Steenkamp and Baumgartner 1998) and organizational studies (Vandenberg and Lance 2000).

Multi-group confirmatory factor analysis (MGCFA) is often used by researchers to test for measurement invariance, and standard structural equation modeling (SEM) goodness-of-fit measures, such as such as the Comparative Fit Index (henceforth CFI), the Tucker-Lewis Index (TLI), the Root Mean Squared Error of Approximation (RMSEA), or the Standardized Root Mean Square Residual (SRMR) are ordinarily used to decide whether comparability is a reasonable assumption in each particular case.

Most applied invariance tests reported in the recent social sciences literature employ the decision criteria for these fit indices, which were developed drawing on the results of few simulation studies of the two-group setting (e.g. Cheung and Rensvold 2002; Chen 2007; Meade, Johnson, and Braddy 2008). Do these criteria apply equally well to much larger, heterogeneous, and complex samples, typical for modern international survey projects, such as the European Social Survey or the World Values Survey, which may include as much as 50 or even 100 countries as their participants?

Though the issue of measurement invariance testing with many groups recently attracted some scholarly attention, we still know little with regard to how reliable the conventional procedures of invariance testing are in the complex context of large cross-national surveys. The efforts of most researchers in the field are aimed at development of novel methods for (a) unbiased estimation of substantive model parameters of interest (e.g., latent means or path coefficients) under varying amounts of non-invariance (e.g., van de Schoot et al. 2013; Kim et al. 2017; Muthén 2018) and (b) direct modelling of measurement non-invariance (e.g. Davidov et al. 2012).

The adequacy of existing guidelines for measurement invariance testing with respect to the situation when many groups are being compared has been studied to a much lesser extent. The only prominent exception is a recent series of simulation studies by Rutkowski and Svetina

(Rutkowski and Svetina 2014, 2017; Svetina and Rutkowski 2017). These authors explored, using the 10- and 20-group settings, how well popular SEM goodness-of-fit indices are able to detect measurement invariance in several scenarios typical for modern international surveys.

Using Monte Carlo simulation experiments, this paper complements Rutkowski and Svetina's findings and contributes to the literature on methodology of cross-national survey research by examining the performance of aforementioned goodness-of-fit indices with respect to a specific task of measurement invariance testing with categorical data and large samples (10-50 groups). Study factors include the number of groups, the level of non-invariance in the data, varying from full invariance to approximate invariance (van de Schoot et al. 2013) to large non-invariance, and the absence/presence of model misspecifications other than non-invariance. In sum, the study design yields a total of 81 conditions ($3 \times 9 \times 3$). All simulated data sets are analyzed using two popular SEM estimators, MLR and WLSMV.

The results of the simulation study suggest that the CFI (whatever estimation method is used) and the SRMR (only when MLR estimation is used), are generally able to distinguish between the fully invariant data and the highly non-invariant data, yet may sometimes fail to discriminate between the fully invariant data and the "weakly", or approximately non-invariant data. The TLI and the RMSEA on average perform poorly than the former two fit measures, especially when other misspecifications are present in the model, and therefore can serve only as auxiliary tools of invariance testing in cross-national contexts.

Importantly, it is found that different study factors exhibit non-trivial, and often non-linear and multiplicative, effects on the sensitivity of all studied fit measures to lack of measurement invariance, thereby making it difficult to formulate universally applicable decision criteria for equivalence testing with many groups. Thus, although the paper concludes with a set of suggestions regarding specific cut-off points for different fit measures and invariance levels, these recommendations should be used with great caution.

The paper proceeds as follows. The next section explains formally what measurement invariance is. The following section reviews what is known about how various goodness-of-fit indices react to lack of measurement invariance. The section that comes after it introduces the study design and discusses in which respects it differs from other similar studies (Rutkowski and Svetina 2014, 2017; Kim et al. 2017). The section that follows next reports the results of simulation experiments. The concluding section discusses the main findings and limitations of the study and outlines a set of recommendations for applied researchers regarding measurement invariance testing in the settings with many groups.

What is measurement invariance?

As Davidov et al. (2014, 58) define it, "Measurement invariance is a property of a measurement instrument (in the case of survey research: a questionnaire), implying that the instrument measures the same concept in the same way across various subgroups of respondents". Measurement invariance is an important prerequisite for making comparisons involving latent constructs across culturally distinct units, since it ensures that the latent construct of interest has the same scale and the same baseline in all units, and therefore latent individual and mean scores are comparable across units.

More formally, established measurement invariance ensures that individuals from different groups that have the same score on a latent scale will provide similar responses on observed indicators, and *vice versa*, that those who have different scores on a latent variable will give consistently different responses. Consider a standard MGCFA model for continuous data:

$$y_{ijg} = \upsilon_{jg} + \lambda_{jg}\eta_{ig} + \delta_{ijg} \quad (1)$$

where y_{ijg} represents the (continuous) response of the individual i from the group g on the item j, $\upsilon_{jg}$is the intercept for the item j in the group g, $\lambda_{jg}$is the factor loading for the item j in the group g, η_{ig} is the individual score on the latent variable η in the group g, and δ_{ijg} represents the residual for the individual i and the item j in the group g.

Three ordered levels of invariance are most frequently used in MGCFA. *Configural* invariance is the first and lowest level. It requires only that the loading patterns are the same across groups (that is, the same indicators have non-zero loadings on the same constructs in all groups). In short, configural invariance ensures that a proposed model measures the same construct in all groups. Lack of configural invariance, in its turn, implies that respondents from different countries understand the same construct in different ways (in other words, they define the same construct using structurally different sets of its attributes, corresponding to different sets of questionnaire items). As a consequence, comparing either individual or aggregated scores on that construct across groups is conceptually nonsensical (as nonsensical is comparing apples to oranges; see Stegmueller 2011). It is worth noting that the presence of configural invariance itself does not allow for a meaningful comparison of latent means or construct-related correlations across groups, though it is a necessary prerequisite of such comparisons.

The second level of invariance is called *metric* or *weak* invariance. It requires that factor loadings are equal across groups, that is $\lambda_{jg} = \lambda_{jg'}, g \neq g'$ for all j and g. Metric invariance

ensures the cross-group equality of the intervals of the scale on which the latent variable is measured. It implies that an increase of one unit on the measurement scale has the same meaning in all groups (Davidov et al. 2014, 63). Notice that the presence of metric invariance still does not permit a cross-group comparison of latent means though it is a sufficient condition for a cross-group comparison of covariances between the construct of interest and other variables (if those are too measured using invariant instruments).

Finally, the third level of measurement invariance – *scalar* or *strong* invariance—assumes that not only loadings, but also the indicator intercepts are equal across groups, that is $\lambda_{jg} = \lambda_{jg'}$ and $\nu_{jg} = \nu_{jg'},\ g \neq g'$ for all j and g (Steenkamp and Baumgartner 1998). Scalar invariance ensures additionally that the origins of the latent scales are the same in all groups or, to put it another way, that group differences in latent means consistently manifest themselves in group differences in the means of the observed items (Steenkamp and Baumgartner 1998, 80). While other types of invariance can be assumed and tested [e.g. invariance of residual variances $\sigma_{jg}(\delta_{ijg})$ across groups], it is generally considered that establishing joint metric-scalar invariance is sufficient to guarantee the reliability of latent means comparison across groups.

Measurement invariance with categorical data

It is a common situation in comparative social surveys when categorical items are used to measure latent constructs. In the factor model for ordered categorical data, the observed scores y_{ijg} are assumed to be determined by unobserved scores on the latent response variables $y*_{ijg}$ (Millsap and Yun-Tein 2004, 481-2). These latent response variables are continuous in scale, unlike the observed measures y_{ijg}. The observed measures can be viewed as discretized versions of the latent response variables, given that scores on the observed measures are determined through

$$y_{ijg} = c \text{ if } \tau_{c-1,jg} < y*_{ijg} < \tau_{cjg} \quad (2)$$

where $c = 1, 2, \ldots, C$ denote response categories of the item j, and τ_{cjg} are latent threshold parameters for the item j in the group g. Note that $\tau_{0jg} = -\infty$ and $\tau_{Cjg} = \infty$. The confirmatory factor model in this case is specified for the latent response variables $y*_{ijg}$ using the following equation:

$$y*_{ijg} = \upsilon_{jg} + \lambda_{jg}\eta_{ig} + \varepsilon_{ijg} \quad (3)$$

The three levels of invariance defined for the continuous case generally apply to the categorical case as well, with the single exception that the strong invariance assumption requires the cross-group equality of item thresholds, $\tau_{cjg} = \tau_{cjg'}$, not item intercepts.

Sensitivity of SEM fit indices to measurement non-invariance: Previous results

Two main approaches for assessing whether measurement invariance holds in the data have been advanced in the literature. Both approaches utilize the fact that a model assuming a stronger form of invariance is nested within a model assuming a weaker form of invariance. The first approach relies on a series of consecutive chi-square difference tests to determine do additional equality constraints required by the assumptions of metric and scalar invariance affect model fit negatively. Statistically significant chi-square differences suggest that a model imposing less equality constraints fits data better than a presumably invariant model, thus indicating lack of invariance.

Another approach is to use various alternative global fit indices, such as the CFI, the TLI, the RMSEA, or SRMR, to assess the *relative/incremental* goodness-of-fit of models assuming different levels of invariance (i.e. metric vs. configural and then scalar vs. metric), by looking at the differences in the absolute values of alternative fit indices between two competing model. If these differences do not exceed some pre-specified thresholds (see below), one can conclude that a more constrained (i.e. invariant at a given level) model fits no worse than a less restricted model and therefore can be preferred to the latter as a more parsimonious one.

Notice, however, that within both approaches configural invariance is tested by evaluating the *overall/absolute/global* fit of the model (that is, by looking at the absolute values of either (a) the chi-squared statistics or (b) the four alternative fit indices mentioned above). Also notice that it is sometimes recommended that the metric and the scalar model should be well-fitting also in terms of absolute fit, not only relative fit (e.g., Milfont and Fischer 2010).

The chi-square test is criticized by various authors (Cheung and Rensvold 2002; Davidov et al. 2014; Yuan and Chan 2016) because it tends to overestimate the discrepancy in goodness-of-fit between nested models in large samples, which are common in comparative survey research. For example, Rutkowski and Svetina (2014; 2017) find in their simulations that, with a large number of clusters (10 to 20), the chi-square test detects a lack of metric and scalar, and sometimes even configural, invariance in conditions where data are generated as fully invariant. So nowadays the

chi-square test is generally used as a complement to the alternative fit indices in measurement invariance checks, not as the primary decision criterion, as it was historically[3].

Regarding the overall model fit, the following criteria are typically used in cross-cultural survey research within the second approach. To claim that the given model is a good fit to the data, in absolute terms, one need to observe the CFI value and the TLI value both larger than 0.95 (or at least 0.90), the RMSEA value smaller than 0.05 (or at least 0.08), and the SRMR value smaller than 0.08 (Browne and Cudeck 1992; Hu and Bentler 1999).

As to relative fit, the most widely used guidelines are those proposed by Chen (2007). According to Chen's recommendations, if the within-group sample size is larger than 300 in each group and does not vary significantly across groups (which is typical for modern cross-national surveys), metric non-invariance is indicated by a change in the CFI value larger than $-.01$, when supplemented by a change in the RMSEA value larger than .015 and a change in the SRMR value larger than .03 compared with the configural equivalence model. With regard to scalar invariance, non-invariance is evidenced by a change in the CFI value larger than $-.01$ when supplemented by a change in the RMSEA value larger than .015 and a change in the SRMR value larger than .03 compared with the metric invariance model. Chen's cutoffs were obtained for the two-group setting.

Rutkowski and Svetina (2014) proposed updated cutoff criteria for the evaluation of the overall and relative goodness-of-fit of MGCFA models with categorical indicators (analyzed using maximum likelihood estimation), particularly suitable for settings where the number of second-level units is relatively large (10 or 20). In terms of the absolute fit measures, their results are as follows. For the CFI and TLI, they suggest that the standard cutoff of 0.95 performs well, but assuming a more stringent threshold of 0.97 may also be reasonable. For the RMSEA they recommend a cutoff of around 0.10 when there are at least 10 groups. Finally, they find that the SRMR is generally not a reliable indicator of the overall goodness-of-fit in large samples. In terms of the relative fit indices, Rutkowski and Svetina propose that metric non-invariance is indicated by a change in CFI larger than $-.02$, when supplemented by a change in the RMSEA larger than .03 compared with the configural equivalence model. With regard to scalar invariance, they conclude that non-invariance of intercepts is evidenced by a change in CFI larger than .02 when supplemented by a change in RMSEA larger than .01 compared with the metric invariance model.

[3] For a criticism of the use of relative goodness-of-fit measures in the context of invariance testing, please see Fan and Sivo (2009).

When the weighted least squares estimator is used to analyze categorical data with MGCFA, they suggest (Rutkowski and Svetina 2017) that the RMSEA value close to or lower than 0.55 is indicative of acceptable overall model fit. They also note that a liberal cutoff of 0.08, recommended by some authors, should not be used as an allowable threshold. As to the CFI and TLI, their conclusion is that neither of these measures should be used to assess the overall fit of categorical MGCFA models when the number of groups is relatively large. With regard to the incremental fit measures, they propose that non-invariance of loadings is indicated by a change in CFI larger than −.004, when supplemented by a change in RMSEA larger than .05 compared with the configural equivalence model. In its turn, non-invariance of thresholds can be claimed when a change in CFI is larger than −.004 and a change in RMSEA is larger than .01 compared with the metric equivalence model.

Study design

Simulation conditions

The design of the present simulation study attempts to reproduce real-life conditions that are often encountered by researchers dealing with large-scale international survey data, such as the European Social Survey (ESS), the European Value Study (EVS), the World Values Survey (WVS), or various Barometer studies (e.g. the Arab Barometer or the Eurobarometer). Study factors include the *number of groups* (also referred to as the second/group level sample size; three possible values), the *level of non-invariance* in the data (also referred to as *invariance condition*; nine possible values), and the absence/presence of *model misspecifications* other than non-invariance (three possible values). In sum, the study design yields a total of 81 conditions: $3 \times 9 \times 3$.

It is worth briefly discussing in which respects the set of study factors used in the present simulation experiment differs from that used by Rutkowski and Svetina (2014, 2017) and Svetina and Rutkowski (2017). In their simulations, these authors manipulated such factors as the scale length, the number of non-invariant items, the number of factors in the model, the proportion of non-invariant items, and the source of non-invariance. Their articles provide a very important piece of evidence regarding the performance of conventional SEM fit measures in invariance tests conducted with the data from large cross-national surveys but the further examination of the topic is obviously necessary. First of all, Rutkowski and Svetina considered only scenarios with no more than 20 groups, while many modern international surveys involve much more participating countries.

Furthermore, when modelling non-invariance they assumed that some items were fully invariant and some were fully non-invariant (the exact proportion of non-invariant items varied across conditions). In addition, they defined non-invariant parameters just by adding or subtracting some fixed number, always the same for each non-invariant group, from a baseline (invariant) parameter value. This data-generating process is unlikely to be encountered in real-world applications. A bit more probable scenario is the one in which each parameter takes different values in every group, but for some parameters the cross-group differences, summarized as standard deviation or variance of the distribution of group-specific deviations from the sample average parameter value, are relatively small (this feature is known as approximate invariance), while for other parameters these differences are relatively large (that is, they are fully non-invariant).

Importantly, some recent results suggest that substantive parameters of interest in MGCFA and MGSEM models can be reliably estimated even when full metric or scalar (or both) invariance is absent but approximate metric/scalar invariance holds for all or at least some parameters (van de Schoot et al. 2013; Muthén 2018; Pokropek, Davidov and Schmidt 2019). The concept of approximate invariance has been developed in the Bayesian context and, as of now, the respective assumption cannot be formally tested using frequentist tools. It is nevertheless important to understand how sensitive standard SEM fit indices are to minor deviations from full invariance which do no lead to biased substantive conclusions.

Finally, Rutkowski and Svetina focused mostly on the effects of various aspects of model complexity (e.g., the number of items per factor or the number of factors), but effectively ignored the potential impact of local model misspecifications, assuming that their models are correctly specified in all other respects than non-invariance. The latter feature of their study design is again quite unlikely to be representative of typical cross-national survey data.

The model under study is presented in a simplified form at Figure 1. It assumes a single latent construct which is measured using four observed categorical indicators, with each indicator having four ordered response categories. In other words, each indicator represents a four-category scale which is very widely used in sociological and political science surveys to measure attitudes, values, and opinions. Such data are often analyzed using the robust version of the maximum likelihood estimator (henceforth, MLR), an approach which some scholars consider to be theoretically incorrect (e.g. Lubke and Muthén 2004). Another popular estimator choice in this context is the weighted least square mean and variance adjusted estimator (henceforth, WLSMV), which is considered as the best option for categorical SEM in the methodological

literature. In this paper I study how different fit measures perform with measurement invariance tests using both approaches.

Figure 1 about here

Number of groups: This factor can take three different values: (1) 10 – (2) 30 – (3) 50. These values were selected in order to represent group-level sample sizes that are typical in modern comparative survey research. For example, 10 countries or so participate in such cross-national surveys as the Arab barometer or the Latinobarometro; nearly 30 countries participated in latest rounds of the ESS or the ISSP (International Social Survey Programme); finally, 50 and even more countries participated in latest rounds of the WVS. Rutkowski and Svetina (2014, 2017) considered only settings with a moderate second-level sample size (10 and 20 groups). Kim et al. (2017) considered larger second-level sample sizes (25 and 50 groups) but their focus was on comparing the performance of different methods of invariance testing, not different fit measures in the context of a single method, as in the present study.

Degree of invariance: This factor can take nine different values: (1) Full: full invariance – (2) Scalar 1: full metric invariance + approximate scalar invariance (small variation in item thresholds across groups) – (3) Scalar 2: full metric invariance + scalar non-invariance (large variation in item thresholds across groups) – (4) Metric 1: approximate metric invariance + approximate scalar non-invariance – (5) Metric 2: moderate metric non-invariance + approximate scalar non-invariance – (6) Metric 3: large amount of metric non-invariance + approximate scalar non-invariance – (7) Metric 4: approximate metric invariance + scalar non-invariance – (8) Metric 5: moderate metric non-invariance + scalar non-invariance – (6) Metric 6: large amount of metric non-invariance + scalar non-invariance.

Notice that configural invariance is assumed to hold in all conditions. For loading and threshold values corresponding to different levels of (non-)invariance please see the subsection *Parameter values* below.

Other model misspecifications: This factor can take three different values: (1) no residual covariances in the data-generating model – (2) one non-zero residual covariance (size varies across groups) added to the data-generating model – (3) two non-zero residual covariances (sizes vary across groups) added to the data-generating model.

Previous studies of the performance of SEM fit indices with respect to invariance testing with many groups did not consider this factor. Still, given the overall complexity of cross-cultural data, it is a highly realistic scenario that a MGCFA model being tested for invariance may

simultaneously be non-trivially misspecified in other respects, e.g. include non-zero residual covariances or cross-loadings. Furthermore, such misspecifications are likely to vary in size across compared groups. The presence of these misspecifications generally affects the absolute model fit negatively, which may potentially result in incorrect rejection of the invariance (particularly, configural invariance) assumption. The effect of other model misspecifications on the relative model fit is unclear, but definitely worth studying. Disentangling the impact of non-invariance and other fit-deteriorating factors on fit measure values is thus an important concern in the context of measurement invariance tests.

Parameter values

Factor loadings: Four different factor loading sets, corresponding to various levels of metric (non-)invariance, are used in simulations below.

(a) The first set corresponds to the strict/full metric invariance condition, i.e. in this set, all loadings are assumed to be invariant across groups. Loading sizes in this set are $\{0.75, 0.75, 0.6, 0.6\}$ for the first, second, third, and fourth indicator respectively.

(b) The second set corresponds to the approximate metric invariance condition. In the first group, factor loadings for four items are set to the values used in the first set, that is$\{0.75, 0.75, 0.6, 0.6\}$. For other groups, group-specific loadings are generated as draws from the truncated normal distribution. For the first two items, the following mean, standard deviation, lower and upper bounds are used: $TN(0.75, 0.05, 0.6, 0.9)$[4]. The respective values for the third and fourth item are $TN(0.6, 0.05, 0.45, 0.75)$.

(c) The third set of factor loadings corresponds to the condition with a moderate amount of metric non-invariance: Three items are generated as approximately invariant and one item is generated as fully non-invariant. In the first group, factor loadings for four items are set to the values used in the first set. For other groups, group-specific loadings for the first and second item are generated as draws from the truncated normal distribution. For the first and second item, the following mean, standard deviation, lower and upper bounds are used: $TN(0.75, 0.05, 0.6, 0.9)$. For the third item the following values are used: $TN(0.6, 0.05, 0.45, 0.75)$. Finally, for the fourth

[4] Pokropek, Davidov, and Schmidt (2019) find that the cross-group variance in loading/intercept sizes as large as 0.001-0.005 does not lead to critical biases in the latent means estimates. The standard deviation value that is used in this study to represent the scenarios of "approximate" loading/threshold invariance corresponds to a cross-group variance of 0.0025, which may be considered as a reasonable trade-off between the situation which is almost indistinguishable from full invariance, on one hand, and the situation in which the amount of non-invariance becomes non-ignorable, on the other hand,.

item group-specific loadings are generated as draws from the uniform distribution $U(\sqrt{0.1};\ 0.75)$[5].

(d) The fourth set of factor loadings corresponds to the condition with a relatively large amount of metric non-invariance: only two items are generated as approximately invariant, and the remaining two items are generated as fully non-invariant. In the first group, factor loadings for four items are set to the values used in the first set. For other groups, group-specific loadings for the first and third item are generated as draws from the truncated normal distribution. For the first item, the following mean, standard deviation, lower and upper bounds are used: $TN(0.75, 0.05, 0.6, 0.9)$. For the third item the following values are used: $TN(0.6, 0.05, 0.45, 0.75)$. For the second and fourth item group-specific loadings are generated as draws from the uniform distribution, $U(\sqrt{0.1};\ 0.9)$ and $U(\sqrt{0.1};\ 0.75)$ respectively.

Item thresholds: Three different item threshold sets, corresponding to various levels of scalar (non-) invariance, are used in simulations below.

(a) The first set corresponds to the full scalar invariance condition. In this set, all item thresholds are assumed to be invariant across groups. Item thresholds in this set are $\{-0.8, 0, 0.8\}$ for the first and second item and $\{-0.6, 0, 0.6\}$ for the third and fourth item.

(b) The second set corresponds to the approximate scalar invariance condition. In the first group, thresholds for four items are set to the values used in the first set. For other groups, group-specific thresholds for each item are generated as draws from the truncated normal distribution with the following mean, standard deviation, lower and upper bounds: $TN(\tau_{jc,Cond\ 1},\ 0.05,\ \tau_{jc,Cond\ 1} - 0.2,\ \tau_{jc,Cond\ 1} + 0.2)$, where $\tau_{jc,Cond\ 1}$ is the threshold value for the j-th item and the c-th response category in the first set.

(c) The third set corresponds to the condition with a large amount of scalar non-invariance. In the first group, thresholds for four items are set to the values used in the first set. For other groups, group-specific thresholds for each item are generated as draws from the truncated normal distribution with the following mean, standard deviation, lower and upper bounds: $TN(\tau_{jc,Cond\ 1},\ 0.2,\ \tau_{jc,Cond\ 1} - 0.35,\ \tau_{jc,Cond\ 1} + 0.35)$, where $\tau_{jc,Cond\ 1}$ is the threshold value for the j-th item and the c-th response category in the first set.

[5] In this, and other non-invariant conditions, the lower bound for non-invariant loadings is set to $\sqrt{0.1} \approx 0.32$ since this value is often referred to as a minimum reasonable factor loading size (e.g. Brown 2015).

Total and residual item variances: In the first group, all total item variances are set to 1, for all other groups item variances are generated from $U(0.8;\ 1.2)$. Then residual variances were calculated for each item by simply subtracting the respective squared factor loadings (invariant set) from the respective item variances for each item and group. Notice that, although residual variances sizes vary across groups, they do not change across simulation conditions, that is, for each item, its residual variance in a given group remains constant in all invariance/misspecification conditions and does not depend on its loading size in a given condition – only on its loading size and total variance in the fully invariant condition.

Latent means and variances: In the first group, the mean of the latent construct is set to zero and the variance of the latent construct is set to one. For the other groups, latent means are sampled form $N(0;\ 1)$ and latent variances are sampled form $U(0.6;\ 1.4)$[6].

Notice that, although all model parameters vary to some extent across groups, they do not change across simulation conditions

Data generation and analysis

For each condition, I generate 500 data sets. Within-group sample sizes are 1000 for the first 30% of groups, 1500 for the next 40% of groups, and 2000 for the remaining 30% of groups. Such within-group sample sizes are typical for most modern cross-national surveys, e.g. ESS or EVS/WVS. In each group 10% of observations have missing values on at least one indicator. Missing values are generated using the Missing Completely at Random (MCAR) mechanism. All model parameters (except those in the full invariance conditions) are sampled using a self-written R program, and then data for each replication are generated using the R package *simsem*. To each data set, I fit the configural invariance model, the metric invariance model, and the scalar invariance model. For estimation, I use the software package Mplus 7.11 (Muthén and Muthén 2015), calling it from the statistical computing environment R 3.4.1 using the package *MPlusAutomation* (Hallquist and Wiley 2018). Model identification is achieved using MPLUS defaults for both continuous and categorical CFA models[7]. A full information maximum likelihood (FIML) approach was used to handle missing values with the MLR estimator and a pairwise present approach was used to handle missing values with the WLSMV estimator.

[6] These values were chosen empirically to avoid convergence problems when running a data-simulating R script.

[7] Notice that the identification approach for categorical CFA models with different levels of invariance implemented in the MPLUS program generally follows the model specifications proposed by Millsap and Yun-Tein (2004). Wu and Estabrook (2016) recently showed that this identification approach may be suboptimal. I discuss the results of the analysis of the simulated data using WLSMV estimation and Wu and Estabrook's identification approach in another paper, currently under preparation. These results, however, do not differ much from what is reported here.

Goodness-of-fit measures under examination

This study focuses on four standard SEM goodness-of-fit measures: RMSEA, CFI, TLI, and SRMR. These four measures are by default reported by most SEM software packages (e.g. MPLUS or *Lavaan*) and represent by far the most often used tools of goodness-of-fit assessment in SEM, except maybe the chi-squared statistics. The latter however is not considered here, because previous studies (see references above), both in the two-group and many-group settings show that it is oversensitive to even minor violations of various types of measurement invariance and therefore is not especially useful for the purpose of equivalence assessment. For sake of brevity, I do not provide formal definitions of these fit indices but they can be found in popular SEM textbooks, such as Brown (2015, Chapter 3) or Kline (2015, Chapter 8)

Notice that MPLUS 7.11 cannot compute the SRMR when categorical indicators are used and thresholds are included in the model. Therefore, it is not considered in the second part of this study, where the simulated data are analyzed using the WLSMV estimator. I nevertheless studied how the categorical counterpart of SRMR, which is known as the Weighted Root Mean Square Residual, or simply WRMR, responds to lack of measurement invariance, but found that this measure performed very poorly in terms of both absolute and relative fit (see Figures A1 and A2 in Appendix). The respective results are therefore not included, in order to shorten the presentation of main findings.

Results

I first present the results of the MLR analysis and then the results of the WLSMV analysis. For each estimation method, the following discussion is focused on two main questions. First, how well each particular fit index performs with regard to differentiating between invariant and non-invariant models in terms of *absolute fit*. Second, how well the same task is performed by each fit measure under study, when the differences in the overall goodness-of-fit between models with nested levels of invariance (*relative fit*) are considered instead of the absolute values. In order to preserve space and make the presentation of results more efficient and understandable to the reader, all findings are presented in a graphical form. On each figure below, dots represent the average values of the respective fit statistics across the 500[8] replications within each respective condition and invariance level, and error bars show the corresponding 95% confidence intervals.

[8] With the WLSMV estimator, in some conditions the actual number of replications used in the analysis is smaller due to non-convergence, see discussion below.

MLR analysis

Convergence checks: Convergence was assessed using MPLUS defaults (max. number of iterations = 1000; convergence criterion = 0.0005). All models converged successfully in all conditions.

Absolute fit - CFI: As shown at Figure 3, when the configural model is fitted to the simulated data, CFI has a perfect or nearly perfect value of 1.000 in all invariance conditions if there are no other types of misspecification in the true model. When misspecifications are present, the CFI value deteriorates, but only slightly: across all conditions with misspecifications, there are only two where the average CFI value of the configural model is lower than 0.978 (Two misspecifications - Metric 2 - 10 Groups and Two misspecifications - Metric 5 - 10 Groups). Noticeably, when the data-generating model includes two misspecified residual covariances, this fit index's sensitivity to misspecifications decreases with sample size: on average, across 50-group conditions it has a higher value, than across 30-group conditions. This is not the case for one-misspecification conditions, where the CFI value decreases when the second-level sample size becomes larger, which is to be expected.

Figure 3 about here

As the level of assumed invariance increases from configural to metric and from metric to scalar invariance, the CFI predictably yields lower values. However, in conditions where the data-generating model is invariant or approximately invariant (e.g. Full, Scalar 1, Metric 1), the respective drop in the CFI is not especially large, except the 10-group two-misspecifications domain, where, for example, the scalar model fitted to the fully invariant data has the CFI value of only 0.95. In general, the CFI correctly detects metric non-invariant models, reacting to increasing amount of loading non-invariance negatively, There is one prominent exception from this rule: in all three 10-group conditions the CFI for the metric model in the two worst metric non-invariant conditions, Metric 3 and Metric 6, is by 0.01-0.03 higher than in conditions with the same level of scalar non-invariance but lower amounts of loadings non-invariance (Metric 1 - 2 and 4 - 5 respectively).

As to the case when the scalar model is fitted to the data, the CFI is able to detect correctly deviances from the assumption of intercept invariance in all studied conditions. In all conditions where Thresholds Set 2 were used for simulations, the CFI value is lower than 0.939 and in most such conditions it is even lower than a liberal cutoff value of 0.9. However, the CFI's ability to identify scalar non-invariance seems to degrade slightly in the 30-group and 50-group

conditions, compare to the 10-group conditions. Finally, it is worth noting that the CFI has the worst values in conditions that assume a significant amount of between-group variance in both loadings and intercepts (Metric 3-6). This finding suggests that when these two types of non-invariance are simultaneously present in data, they may affect the CFI's performance in a multiplicative way.

Figure 4 about here

Absolute fit - TLI: In conditions where there are no misspecifications other than non-invariance, the TLI behaves well and similarly to the CFI. For the configural model, it has a perfect value of 1.00 or very close to it in all no-misspecification conditions (see Figure 4). As to the metric and scalar invariance testing, it is generally able to discriminate between invariant or approximately invariant models and non-invariant models (though, along with the CFI, it fails in 10-group Metric 3 and Metric 6 conditions). However, the TLI's performance becomes miserable when this measure is applied to misspecified models. Both in one- and two-misspecification conditions the CFI fails to identify metric non-invariance, except few most non-invariant conditions. Moreover, when the data-generating model is fully or approximately invariant, it tends to indicate that the configural model fits even worse than the metric and scalar models. For example, in the 30-group full-invariance one-misspecification condition the configural model has the TLI value of only 0.938, while the metric model for the same condition has the TLI equal to 0.968 and the scalar model has the TLI equal to 0.976. In the one- and two-misspecification settings, the average TLI value for the configural model does not exceed 0.974 in any condition, and often is below 0.950. The TLI, however, is reactive to a simultaneous lack of metric and scalar non-invariance (which is indicated by the TLI values for the scalar model lower than 0.95), though, again alike the CFI, its responsiveness decreases with the group-level sample size.

Absolute fit - RMSEA: When the configural model is fitted to no-misspecification conditions, its average RMSEA value is in the range 0.005-0.010 and is in general not affected by (a) the amount of metric and scalar non-invariance present in the respective data set and (b) the second-level sample size (see Figure 5). When metric invariance is tested, the RMSEA still performs well, though it, for the metric model, may react not only to metric, but also to scalar non-invariance, but to a smaller extent. In general, if there are no misspecified residual covariances, the RMSEA value for the metric model higher than 0.025 suggests a considerable level of cross-group variability in sizes of factor loadings. As to scalar non-invariance, it is indicated by the RMSEA values larger than 0.07, when it comes alone (condition Scalar 2), and 0.08, when it is accompanied by loading non-invariance (conditions Metric 4-6).

Figure 5 about here

In conditions with one or two misspecified covariances, however, using the RMSEA to determine the absolute goodness-of-fit of invariance model under test seems problematic. The configural model has the RMSEA near or above the standard threshold of 0.05 in all conditions with misspecifications. The same is true for the metric and scalar models. Even if the true model is fully invariant but includes misspecifications not related to the invariance assumption, the RMSEA for either the loading-invariant or intercept-invariant model has a value of at least 0.038 (scalar model; condition 10 Groups - One Misspecification - Full) and is typically higher than 0.05. Moreover, as with the CFI, in invariant conditions with misspecifications, the RMSEA of the metric or scalar model may sometimes exceed that of the configural model, sometimes by a rate as high as 0.037 (configural vs. scalar model; condition 50 Groups - One Misspecification - Full).

Another problem with the RMSEA is that it faces difficulties with identifying loading non-invariance when the number of groups is small (10) and there are other misspecifications. When the group-level sample size increases to 30 or 50 units, the RMSEA correctly rank models with consecutively increasing levels of metric non-invariance, but more liberal absolute fit cutoffs should be imposed in this case, compare to the no-misspecification case: at least 0.06 when there is only one misspecified residual covariance, and 0.05 when there are two misspecified residual covariances. It nonetheless must be noted that even in the presence of other model misspecifications, the RMSEA is still responsive to a lack of scalar invariance, and the same cutoffs can be applied as in the no-misspecification setting: 0.07 if loadings are fully or approximately invariant and 0.08 if scalar non-invariance is complemented by metric non-invariance.

Absolute fit - SRMR: When configural invariance is tested, the SRMR is less sensitive to the presence of other model misspecifications than the TLI and the RMSEA but to some extent more sensitive than the CFI (see Figure 6). If there are no such misspecifications, the average SRMR value for the configural model across all invariance and group-level sample size conditions is 0.008 (with very little variation around it). If such misspecifications are present, the average SRMR value for the configural model is in the range 0.014-0.024, closer to the lower bound of that interval in 10-group conditions and closer to the upper bound in 30- and 50-group conditions when there is one misspecified residual covariance and the other way around when there are two misspecified residual covariances.

Figure 6 about here

With regard to metric invariance, the SRMR correctly retains loading-invariant models in 30- and 50-group conditions, irrespectively of the number of misspecified residual covariances, with SRMR values larger than 0.045 indicating non-trivial variation in loading sizes across groups. The same cannot be said about the 10-group setting, where the SRMR, as all measures considered above, fails in the two conditions with the largest amount of loading non-invariance (Metric 3 and 6), estimating the absolute fit of the respective metric models to be at the very same level as that of the metric model for the fully invariant condition.

Finally, as for all previously considered fit indices, a multiplicative effect of metric and scalar non-invariance on the SRMR value is observed: when data are generated with a significant amount of loading non-invariance, the scalar model for conditions with a moderate level of scalar non-invariance (Metric 1-3) fits as poor of even poorer as the scalar model for the condition with considerable scalar non-invariance, but metric invariant (Scalar 2). However, for a given level of metric (non-) equivalence, the SRMR can correctly order models with different degrees of scalar non-invariance. Overall, SRMR values larger than 0.05 indicate a relatively large intercept (or joint loading-intercept) cross-group heterogeneity.

Relative fit: As the results shown at Figure 7 suggest, when the assumption of metric invariance is tested, the CFI and the SRMR perform on average better than the TLI and the RMSEA. In the no-misspecification setting all fit measures demonstrate a comparable level of sensitivity to a lack of loading invariance, but in conditions with one or two misspecified residual covariances the latter fit indices often, and especially in invariant or approximately invariant conditions, indicate that the metric model fits *better* than the configural model. The most problematic in this respect is the 10-group setup. For example, only in one (Metric 5) out of nine 10-groups one-misspecification conditions (Figure 7, top central panel), the TLI and the RMSEA suggest that the configural model is superior to the metric model. In all other conditions, even those where the amount of loading non-invariance is huge (Metric 3 and 6), these measures fail to reject the metric invariance assumption. This deficiency is to some extent less pronounced in 30- and 50-group conditions, but even in those settings the TLI and the RMSEA respond only to the most severe violations of the assumption of equal loading size.

The CFI and the SRMR also face some difficulties in identifying severely metric non-invariant models in the 10-group setting, but otherwise they perform relatively well. It nonetheless must be noted that with a small number of groups, the differences between the values of these two measures for the configural and the metric model may sometimes suggest rejection of the metric invariance assumption when the data-generation model assumes strictly invariant loadings but

highly non-invariant intercepts (condition Scalar 2). Overall, the metric invariance hypothesis is supported by CFI differences between the configural and the metric model no larger than – 0.004 in the no-misspecification setting, – 0.006 in the one-misspecification setting, and – 0.008 in the two-misspecification setting. The SRMR, in terms of the relative fit, is not particularly sensitive to the presence of other model misspecifications. In almost all sample size and misspecification conditions, SRMR differences not larger than 0.02 provide evidence in favor of the metric invariance assumption.

Figures 7 and 8 about here

The relative fit results for scalar invariance tests (shown at Figure 8) indicate slightly better performance of all fit measures, compare to the metric equivalence assessment case. For the CFI, in no- or one-misspecification conditions, differences in this measure between the metric and the scalar model smaller than – 0.01 are indicative of full or at least approximate intercept equivalence. In the two-misspecification setup, simulations suggest a more liberal cutoff that falls in the range [– 0.015; – 0.025], depending on the group-level sample size (with stricter values corresponding to larger sample sizes). ΔCFI values larger than – 0.06 typically indicate the presence of significant amount of scalar non-invariance. They nevertheless may also arise in situations when approximate scalar invariance holds but approximate metric invariance does not (e.g. conditions Metric 2 and Metric 3), though CFI differences of [– 0.03; – 0.05] are more typical in the latter scenario.

The results for the TLI are generally similar to those for the CFI, though using this fit index in intercept invariance tests poses more problems in situations characterized by approximate scalar invariance accompanied by considerable metric non-invariance. TLI differences between the metric and the scalar lower than [– 0.005; – 0.010] suggest that either strict or approximate scalar invariance holds. The TLI actually fails to discriminate between these two types of invariance, since this fit index yields essentially the same metric-scalar differences in the respective simulation conditions (Full and Scalar 1).

As to the RMSEA, the difference in the value of this fit index between the metric and scalar model is lower than 0.002 in all but one simulation conditions assuming strict scalar invariance (the only exception is 10 Groups - One Misspecification - Full, where it takes a value of 0.005). When strict or almost strict metric invariance is established, the RMSEA differences as high as [0.010; 0.015] may indicate a non-critical level of intercept non-invariance. The SRMR is similar to the TLI in that it does not discriminate between the fully scalar invariant data and approximately scalar invariant data. Both situations are indicated by the SRMR differences

between the metric and the scalar model close to or smaller than 0.010 (except the 10-group one-misspecification setting). Values larger than 0.020 can be safely interpreted as evidence of a non-trivial amount of non-invariance in the data. Finally, it is worth attention that the sensitivity of all fit indices to the lack of strong invariance decreases with sample size, which suggests that more conservative cutoffs should be adopted when the number of groups is significantly larger than ten.

WLSMV analysis

Convergence checks: Convergence was assessed using MPLUS defaults (max. number of iterations = 1000; convergence criterion = 0.0005). In contrast to the MLR estimator, non-convergence is a common problem when WLSMV estimation is used, especially when the configural model is fitted to the data. Configural models experienced problems with convergence in 40 out of 81 conditions (49.4%). In 23 conditions (28.9%), the non-convergence rate (NCR) was greater than 10%, and in 13 conditions it was greater than 20%. The most problematic conditions were the following: 50 Groups - One Misspecification – Full (NCR = 62.6%), 50 Groups - One Misspecification – Full (NCR = 66.6%), and 50 Groups - One Misspecification – Full (NCR = 60.4%). These conditions were checked manually, and the source of non-convergence was the non positive-definite latent variance matrix in Group 39 (perhaps the reason for that was an extremely high population value of the latent variance in that group). Another frequent source of non-convergence was the presence of empty response categories for some items. For metric and scalar models, non-convergence rates were smaller but still substantial. Nonetheless, since even in the worst case the results from more than 150 replications are available and also due to time-saving considerations, I decided to proceed without re-estimation of the problematic conditions with different population values set for problematic parameters.

Absolute fit - CFI: Similar to the results for the MLR estimator, in all no-misspecification conditions the CFI suggests that the configural model has a perfect fit (CFI $\approx$ 1.00). In conditions with misspecifications, the CFI value becomes smaller, but not that much: it is not lower than 0.991 in any such condition (Figure 9). When the assumption of metric invariance is tested, in general CFI values smaller than 0.990 are indicative of a sizeable variation in loading sizes across groups, though in conditions where both approximate metric invariance and approximate scalar invariance hold (Metric 1) the CFI can sometimes take values larger than the aforementioned threshold.

Figure 9 about here

Yet, it is worth mentioning that when the number of second-level units is small, the absolute fit of the metric model, according to the CFI, may deteriorate not only due to loading non-invariance but also due to intercept non-invariance. For example, in the condition 10 Groups - One Misspecification - Scalar 2 (no metric non-invariance, high level of scalar non-invariance), the CFI of the metric model is 0.978 while in the condition 10 Groups - One Misspecification - Metric 1, which assumes a lower level of variation in threshold sizes but higher level of variation in loading sizes across groups, it is as high as 0.997. As to strong invariance, the CFI of the threshold-invariant model lower than 0.985 (0.990 in the no-misspecification setting) indicates than only approximate strong invariance holds, and the CFI of the threshold-invariant model lower than 0.97 suggests the presence of significant cross-group threshold heterogeneity.

Figure 10 about here

Absolute fit - TLI: As the results presented at Figure 10 show, the TLI performs well with configural invariance testing in the absence of misspecified model parameters, having a value of 1.00 or very close to it in all such conditions. In the one- and two-misspecification setups, the TLI of the configural model does not exceed 0.990 and is sometimes as low as 0.972. As to metric invariance testing, full or at least approximate metric invariance is supported by the CFI values larger than 0.990 in the no-misspecification setting, 0.985 in the one-misspecification setting, and 0.980 in the one-misspecification setting. As well as the CFI, when factor loadings are assumed to be equal in all groups, the TLI seems to be overreacting to the lack of strong invariance: It suggests rejecting the hypothesis of loading invariance in the Scalar 2 condition in all misspecification and number of groups settings, in spite of the fact that loadings are fully invariant in this condition. In contrast, this measure demonstrates good properties when it is used for threshold invariance testing. The TLI value for the threshold-invariant model lower than 0.975 can be safely interpreted as an indication of non-ignorable scalar non-invariance, and this cutoff applies to all group-level sample size and misspecification conditions considered in the study.

Figure 11 about here

Absolute fit - RMSEA: The RMSEA of the configural model is particularly sensitive to the presence of model misspecifications (see Figure 11). While in the no-misspecification setting it takes quite small values (0.004–0.006), in one- and two-misspecification conditions it is generally higher than the conventional threshold of 0.05 and sometimes even as high as 0.07. As with the CFI and TLI, the RMSEA of the metric invariance model is highly sensitive to severe violations of the scalar invariance assumption. When the amount of threshold variability in the

data is small to moderate and there are no misspecified residual covariances, RMSEA values smaller than 0.025 support the hypothesis of metric equivalence. In conditions with one or two misspecifications the cutoff for loading-invariant models raises to 0.05. With regard to scalar invariance, RMSEA values close to or lower than 0.06 support the respective assumption (in conditions with two misspecifications, a somewhat more liberal threshold might be reasonable).

Relative fit: When the metric model is compared to the configural model, the CFI, on average, demonstrates better performance than the TLI and the RMSEA (Figure 12). In the presence of model misspecifications, the latter two measures face problems with identifying metric non-invariant models, which feature is especially pronounced in the one-misspecification setting (though it must be noted that this deficiency becomes less problematic when the number of groups is large). Overall, CFI differences between the metric and the configural model smaller than – 0.005 may suggest that at least approximate metric invariance holds in the data. The respective cutoffs for the ΔTLI and the ΔRMSEA cannot be proposed since the results for these two measures are highly inconsistent across studied conditions in the context of metric invariance testing. Finally, as with the absolute fit evaluations, when the CFI, the TLI, and the RMSEA are used to assess the relative fit of the loading-invariant model, they are too much responsive to scalar non-invariance: the differences between the values of all three measures for the configural and for the metric model in the Scalar 2 condition (which assumes strict metric invariance) in all studied settings is larger than those in conditions Metric 1 to 3 (small to moderate to large amount of metric non-invariance).

Figures 12 and 13 about here

In contrast to the loading invariance case, these three measures are largely effective at identifying scalar non-invariant models (Figure 13). Specifically, the average CFI difference between the loading-equivalent and the threshold-invariant model in conditions that assume strict threshold invariance is about [– 0.004; – 0.002]. CFI differences of – 0.04 and smaller (10-group conditions) or – 0.02 and smaller (30- and 50-group conditions) are typical for conditions that assume only approximate threshold invariance (Scalar 1 and Metric 1-3). For the TLI, differences above zero are required to support full strong invariance and differences smaller than – 0.01 are required to support approximate strong invariance. The latter value applies to all group-level sample size and misspecification conditions. Finally, for the RMSEA the respective cutoffs are zero or lower for strict invariance and 0.01 and lower for approximate invariance.

Discussion

To preserve space and shorten the related discussion, the recommended absolute and relative fit cutoffs for the four fit measures considered in this study are presented in a tabular form. Table 1 reports suggestions regarding overall fit evaluations of models assuming configural, metric, and scalar invariance. Table 2 reports recommendations regarding assessment of the relative goodness-of-fit of the metric and scalar models. Critical values proposed in these tables are based on the average values of the 2.5^{th} (for CFI and TLI) or 97.5^{th} (for RMSEA and SRMR) percentiles of the respective fit indices across conditions in which full invariance of a given level holds.[9]

Tables 1 and 2 about here

It must nevertheless be underscored that all these recommendations reflect only the average performance of the four considered fit indices across studied conditions. None of the cutoffs shown in Tables 1 and 2 apply equally well to all studied conditions. In many conditions these criteria clearly fail and therefore should not be used.

First, the presence of other misspecifications, as expected, negatively affects both the absolute and the relative model fit for all fit indices and invariance levels. This effect is however not linear. For example, all considered fit indices have on average poorer absolute values in the one-misspecification setting, rather than in the two-misspecification setting. In terms of the relative fit indices, the impact of the presence of non-zero residual covariances on the performance of the RMSEA and the TLI is particularly miserable, especially when metric invariance is tested. The ΔCFI and the ΔSRMR perform on average better in this respect.

Second, the group-level sample size only slightly affects the absolute and incremental values of various fit indices. To a much larger extent, it affects their sampling variability: the 95% confidence intervals over simulated replications for all fit measures, either absolute or relative, are narrower in the 30-group and especially in the 50-group setting, compared to the 10-group setting. As a consequence, more liberal cut-off values may be reasonable with relatively small numbers of groups, like those proposed by Rutkowski and Svetina (2014, 2017).

Third, loading and intercept non-invariances generally have a multiplicative effect on model fit, whatever fit index is used to assess it, thus often leading to rejection of either the metric invariance assumption or the scalar invariance assumption due to the influence of an irrelevant

[9] All 81 conditions for configural invariance; Full, Scalar 1, and Scalar 2 for metric invariance (in sum, 27 conditions); Full for scalar invariance (in sum, 9 conditions).

type of non-invariance (loading non-invariance in the former case and intercept/threshold non-invariance in the latter case). In addition, a particular ability of a given index to detect non-invariance at a given level depends on what type of invariance is being tested. The TLI and the RMSEA are generally not able to detect metric non-invariance, except in the situations where its level is extreme or there are no local misspecifications in the model.

Fourth, the studied fit indices allow for distinguishing between highly non-invariant data and approximately invariant data, but often fail to discriminate between approximately invariant data and fully invariant data (still, more liberal cut-off values should be used for correctly detecting approximate invariance, compare to full invariance tests). There is, however, not so much evidence about whether approximate invariance is a sufficient condition for meaningful comparisons of latent means or path coefficients in structural regression models (although see Pokropek, Davidov, and Schmidt 2019 and Muthén 2018).

Fifth, though the differences in proposed cut-off values between different estimation methods are not dramatically large, for the WLSMV estimator the recommended relative fit thresholds are a little bit tougher than those for the MLR estimator.

Overall, my findings only partly overlap with those by Rutkowski and Svetina. With respect to MLR estimation, I obtain more or less similar results to those reported by Rutkowski and Svetina (2014) for the CFI, both in terms of absolute and relative fit, and, to less extent, for the TLI (they used this fit index only as an absolute fit measure). For the other two fit measures the results are less coherent across studies. For example, they recommend that "that the SRMR is not used in isolation, if it is used at all" (Ibid., 52), while my analysis suggests that the SRMR is the second-best measure of the absolute fit out of the four considered in this study, and it also performs well in terms of relative fit assessment (as with the TLI, they did not test the SRMR as a relative fit index). In its turn, for the RMSEA, they propose absolute and relative cut-off points somewhat similar to what could have been suggested using my results, but our conclusions regarding the overall usefulness of this fit index are rather different. While they consider the RSMEA to be a reliable measure of both absolute and relative model fit, my findings indicate that it may at best serve as an auxiliary tool when the scalar model is being compared to the metric model.

As to WLSMV estimation, Rutkowski and Svetina (2017) claim that the CFI should not be used as an absolute fit measure, which again contradicts my results. In contrast, their conclusions about the performance of the RMSEA are much more optimistic than mine. On the other hand, they suggest cut-off values for the ΔCFI which are very close to those proposed in Table 2 above. The findings of Svetina and Rutkowski (2017) are not directly comparable to those

reported here since in that paper the authors explored models with more than one latent construct.

All in all, it is worth mentioning that no one simulation study can embrace all possible scenarios that are encountered in cross-cultural research. The present study leaves several important variables out of its scope, such as the scale length, the presence of asymmetry in item thresholds, the number of latent constructs in the model, or the number of response categories of observed indicators. Though the effects of some missed factors were explored in other similar studies (Rutkowski and Svetina 2014, 2017; Svetina and Rutkowski 2017), there still may be non-trivial interactions between these factors and those studied in this paper (and those not yet covered by simulation research as well) in terms of their joint effects on model fit.

Therefore, I want to conclude by offering several general recommendations that can guide practical researchers conducting measurement invariance analysis in the context of large cross-cultural survey data. Ideally, in each application researchers should perform an *ad hoc* simulation study that attempts to mimic the most relevant features of the data at hand, in order to understand how specific characteristics of the particular study context affect the sensitivity of different fit measures to lack of measurement invariance. It can easily be done using the MPLUS software, R packages *simsem* and *lsasim* (Matta et al. 2018), or R scripts that can be found in the SI to the present study[10].

If conducting a simulation study is not feasible for some reasons, instead of following blindly the cutoffs proposed in this and other similar studies, applied researchers must judge smart and take into consideration both specific features of their data, extra-statistical information about sampling and data collection procedures provided by the team responsible for conduction of a given survey, and, last but not least, theoretical reasons. Another potentially helpful strategy is to complement invariance tests by the exploration of modification indices for the configural model, in order to understand how large is the amount of model misspecifications other than non-equivalence in the model under investigation, and then decide which fit indices, given their respective levels of sensitivity to those misspecifications, should have larger weights in the final decision about whether measurement invariance is a plausible assumption or not.

[10] Replication materials are available from the author upon request.

References

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford Publications.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological methods & research*, 21(2), 230-258.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological bulletin*, 105(3), 456.

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.

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

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., & Meuleman, B. (2012). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology*, 43(4), 558-575.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, 40, 55-75.

Fan, X., & Sivo, S. A. (2009). Using Δ-goodness-of-fit indexes in assessing mean structure invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(1), 54-69.

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. *Structural equation modeling: a multidisciplinary journal*, 25(4), 621-638.

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

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*, 24(4), 524-544.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. The Guilford Press.

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.

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*, 11(4), 514-534.

Matta, T. H., Rutkowski, L., Rutkowski, D., & Liaw, Y. L. (2018). lsasim: An R package for simulating large-scale assessment data. *Large-scale Assessments in Education*, 6(1), 15.

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.

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of psychological research*, 3(1), 111-130.

Muthén, B. (2018). Recent methods for the study of measurement invariance with many groups: alignment and random effects. *Sociological Methods & Research*, 47(4), 637-664.

Muthén, L. K., & Muthén, B. O. (2015). *Mplus user's guide (1998–2015)*. Muthén & Muthén: Los Angeles, CA.

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. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-21.

Poortinga, Y. H. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology*, 24(6), 737-756.

Przeworski, Adam, and Henry Teune. (1966). Equivalence in cross-national research. *Public Opinion Quarterly*, 30(4), 551–68

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.

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

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.

Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25(1), 78-90.

Stegmueller, D. (2011). Apples and oranges? The problem of equivalence in comparative research. *Political Analysis*, 19(4), 471-487.

Van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in psychology*, 4, 770.

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), 4-70.

Yuan, K. H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological methods*, 21(3), 405.

Figures and Tables

Table 1. Recommended Absolute Fit Cutoffs for Different Fit Measures and Invariance Levels

Measure	MLR analysis			WLSMV analysis		
	Configural	Metric	Scalar	Configural	Metric	Scalar
CFI	> 0.985	> 0.98	> 0.97	> 0.985	> 0.98	> 0.97
TLI	NR	NR	NR	NR	NR	> 0.975
RMSEA	NR	NR	NR	NR	NR	NR
SRMR	< 0.02	< 0.04	< 0.045	NA	NA	NA

Notes: Critical values proposed in these tables are based on the (approximate) average values of the 2.5th (CFI and TLI) or 97.5th (RMSEA and SRMR) percentiles of the respective fit indices averaged across all conditions in which full invariance of a given level holds. In some situations, more liberal cutoffs may be appropriate (see the Discussion section). NR = not recommended for use in a given setting. NA = not applicable.

Table 2. Recommended Relative Fit Cutoffs for Different Fit Measures and Invariance Levels

Measure	MLR analysis		WLSMV analysis	
	Metric	Scalar	Metric	Scalar
CFI	> – 0.01	> – 0.01	> – 0.005	> – 0.005
TLI	NR	> – 0.005	NR	0.000
RMSEA	NR	< 0.005	NR	0.000
SRMR	< 0.01	< 0.01	NA	NA

Notes: Critical values proposed in these tables are based on the (approximate) average values of the 2.5th (for CFI and TLI) or 97.5th (for RMSEA and SRMR) percentiles of the respective fit measure differences averaged across all conditions in which full invariance of a given level holds. In some situations, more liberal cutoffs may be appropriate (see the Discussion section). NR = not recommended for use in a given setting. NA = not applicable.

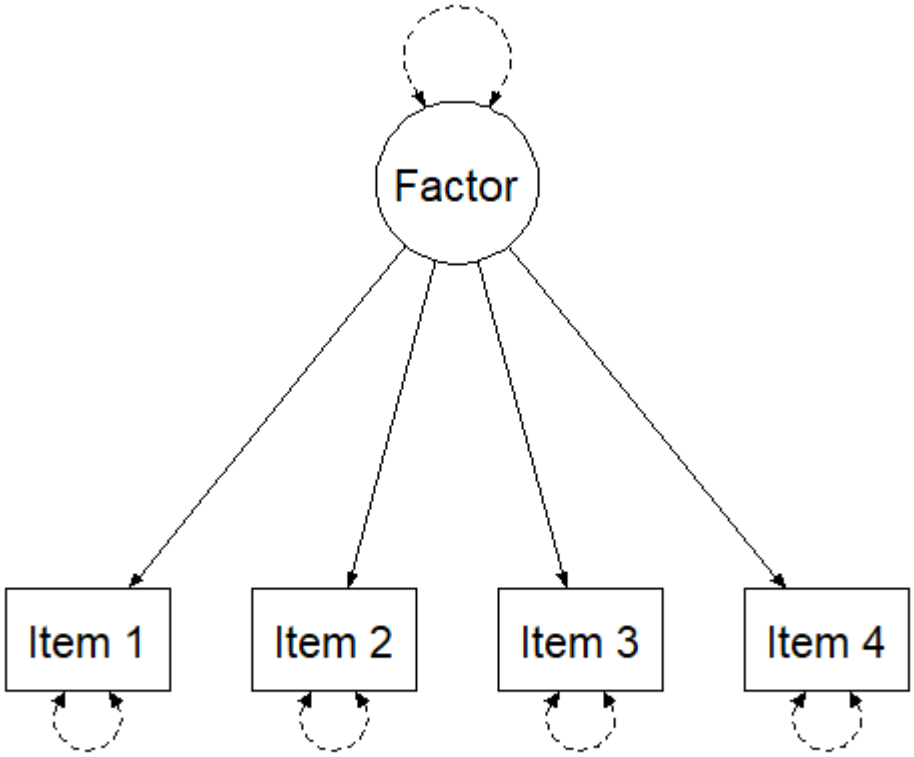

Figure 1 Model structure.

Note: Latent construct is represented using a circle. Rectangles correspond to observed indicators and arrows correspond to factor loadings. Residual variances and latent variance are represented using dashed arcs. Item thresholds are not shown.

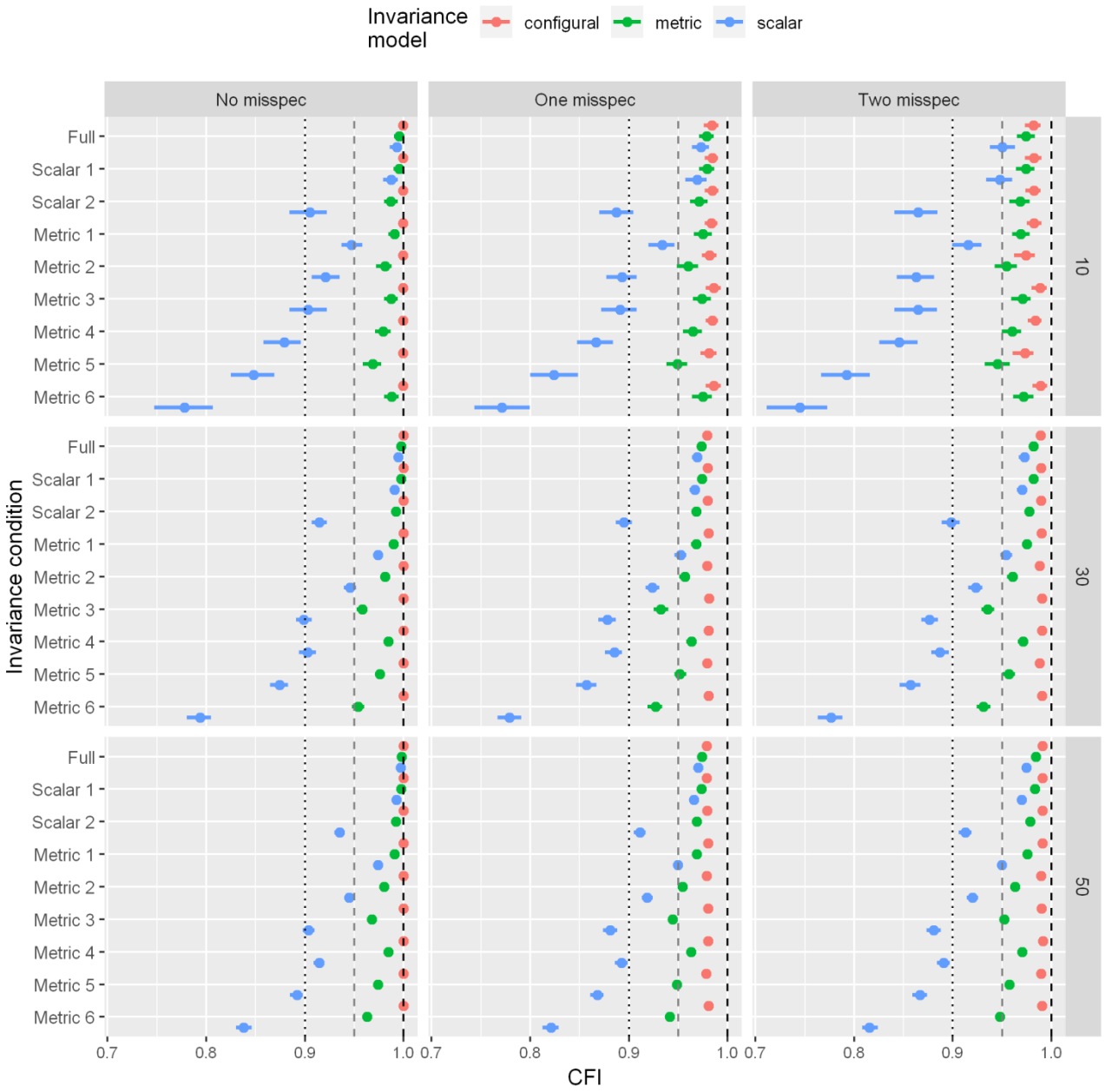

Figure 2 CFI values for configural, metric, and scalar models. MLR analysis

Note: Black dashed vertical line corresponds to CFI = 1.00. Dark grey dashed vertical line corresponds to CFI = 0.95. Dark grey dotted vertical line corresponds to CFI = 0.90. Dots show the average CFI values for each condition over 500 replications. Error bars show the 95% CIs.

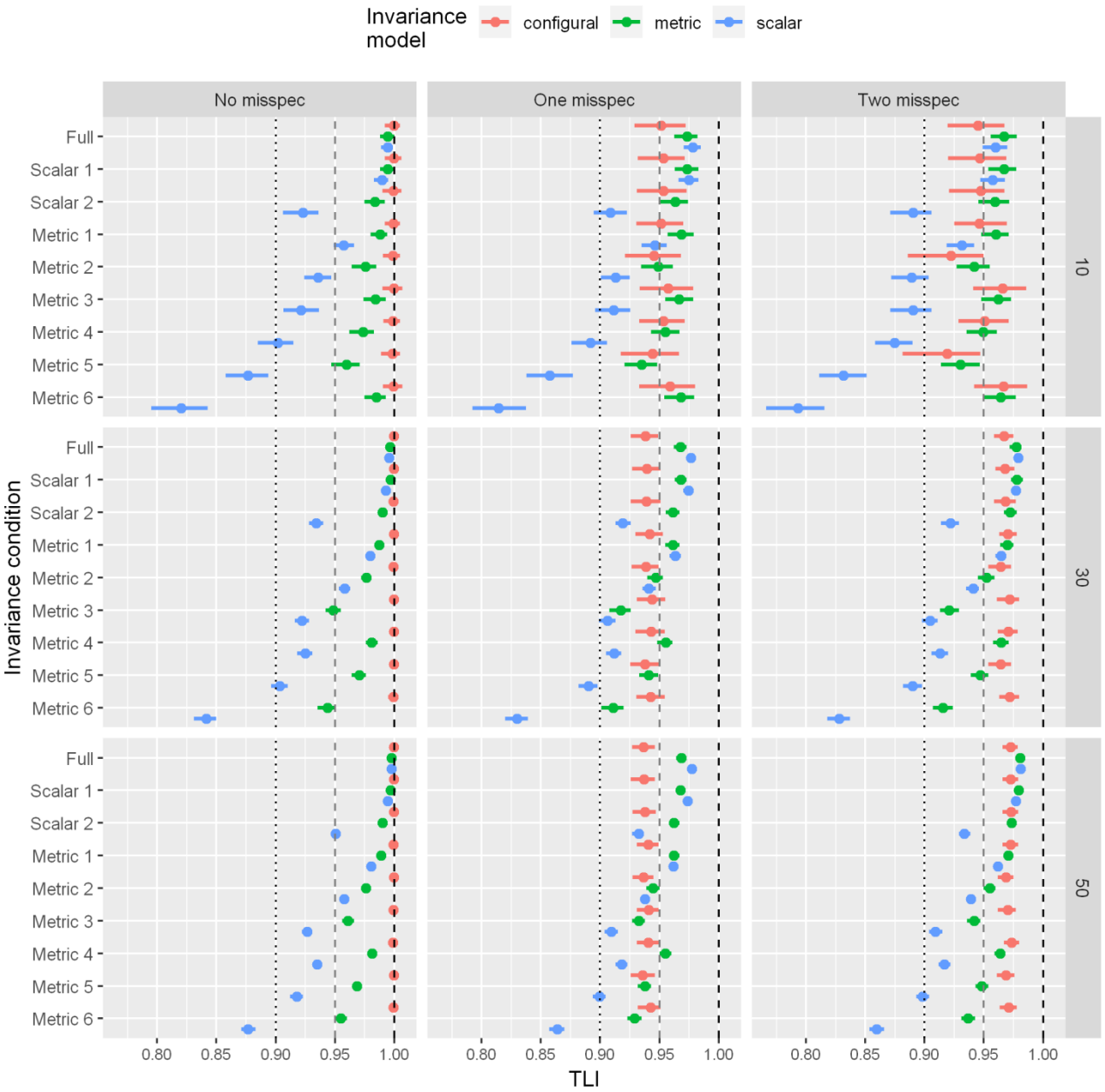

Figure 3 TLI values for configural, metric, and scalar models. MLR analysis

Note: Black dashed vertical line corresponds to TLI = 1.00. Dark grey dashed vertical line corresponds to TLI = 0.95. Dark grey dotted vertical line corresponds to TLI = 0.90. Dots show the average TLI values for each condition over 500 replications. Error bars show the 95% CIs.

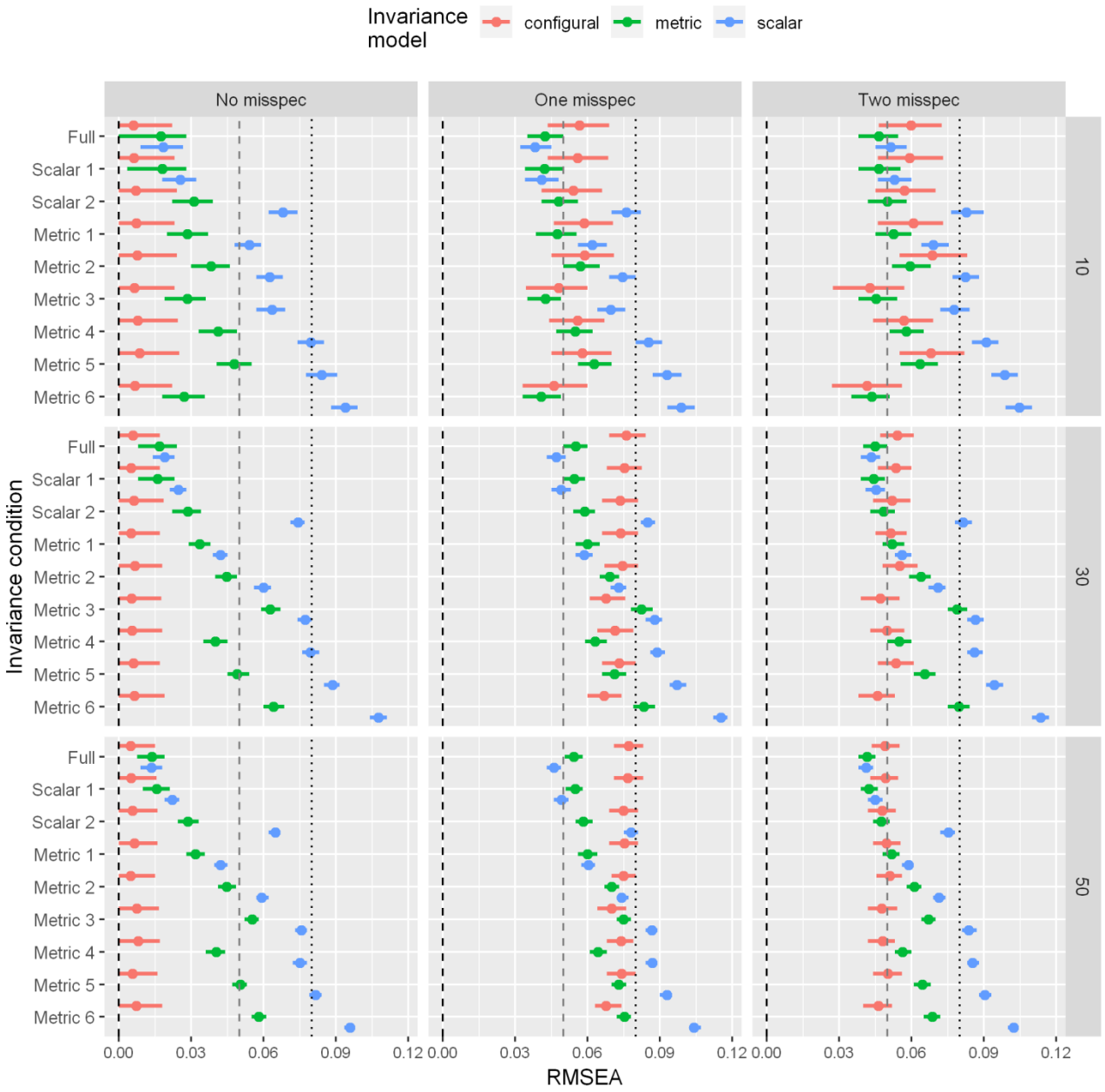

Figure 4 RMSEA values for configural, metric, and scalar models. MLR analysis

Note: Black dashed vertical line corresponds to RMSEA = 0.00. Dark grey dashed vertical line corresponds to RMSEA = 0.05. Dark grey dotted vertical line corresponds to RMSEA = 0.08. Dots show the average RMSEA values for each condition over 500 replications. Error bars show the 95% CIs.

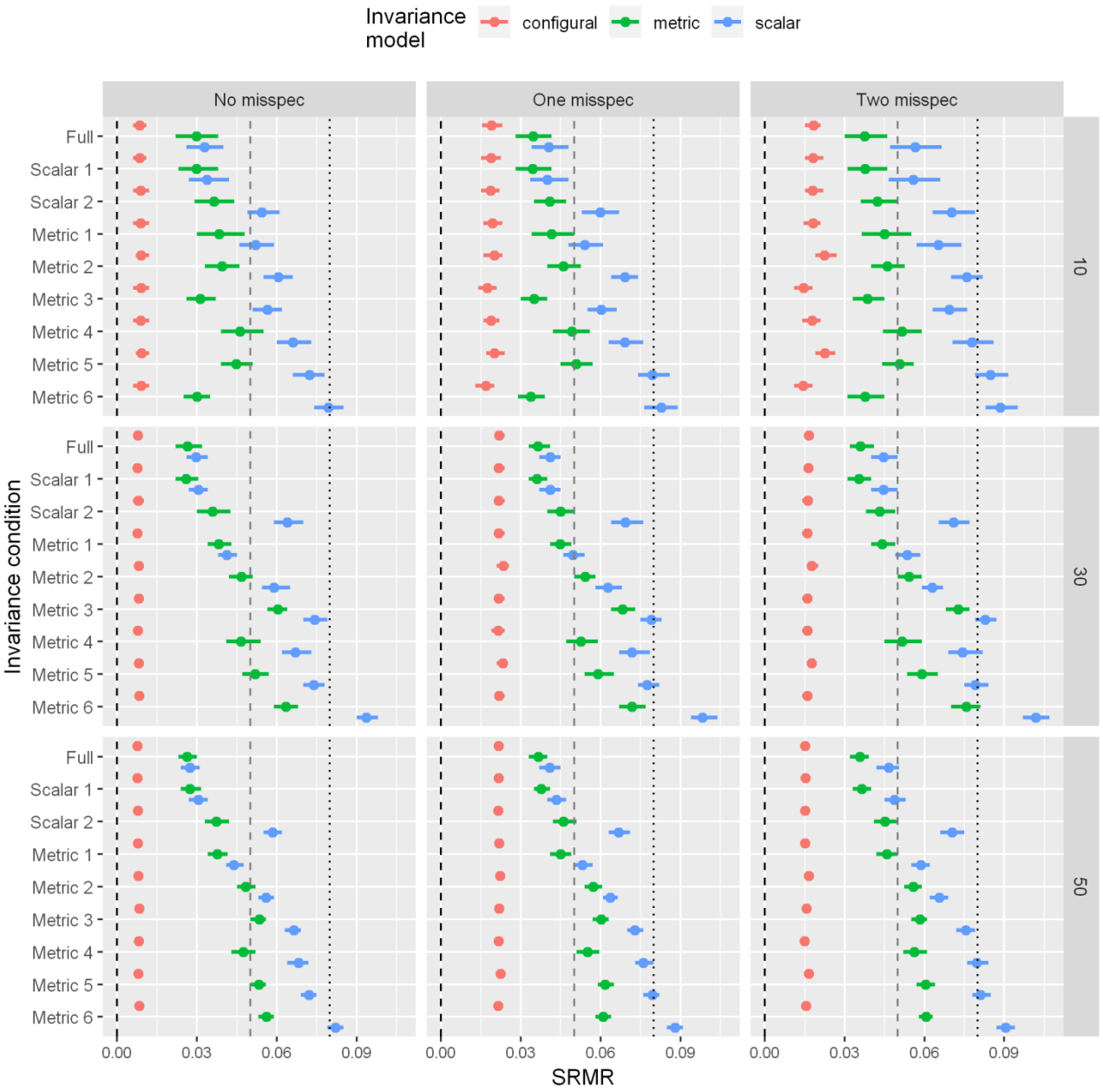

Figure 5 SRMR values for configural, metric, and scalar models. MLR analysis

Note: Black dashed vertical line corresponds to SRMR = 0.00. Dark grey dashed vertical line corresponds to SRMR = 0.05. Dark grey dotted vertical line corresponds to SRMR = 0.08. Dots show the average SRMR values for each condition over 500 replications. Error bars show the 95% CIs.

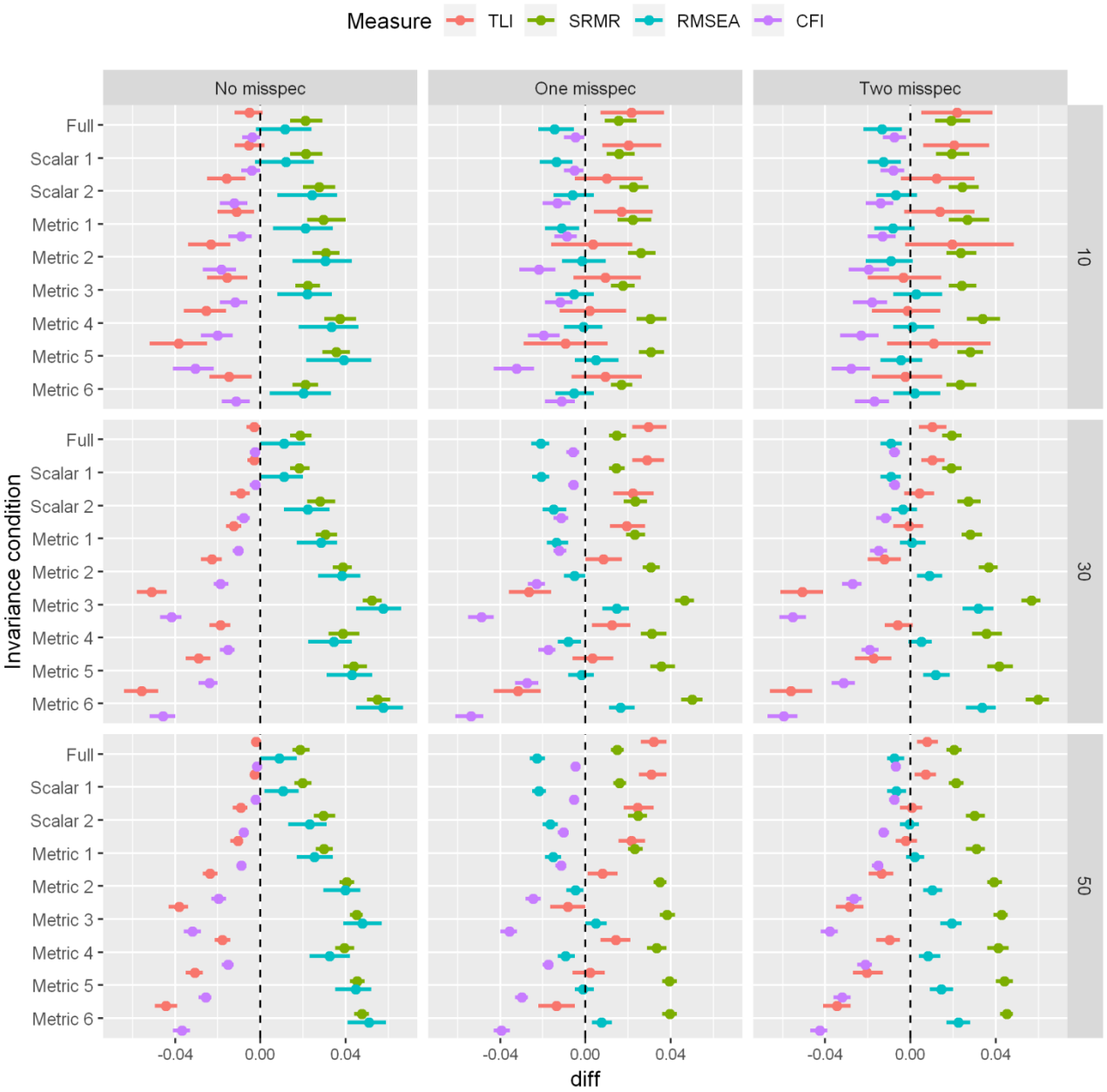

Figure 6 CFI, TLI, RMSEA, and SRMR differences. Configural vs. metric model. MLR analysis

Note: Black dashed vertical line corresponds to Δ (metric - configural) = 0.00. Dots show the average difference for each condition over 500 replications. Error bars show the 95% CIs.

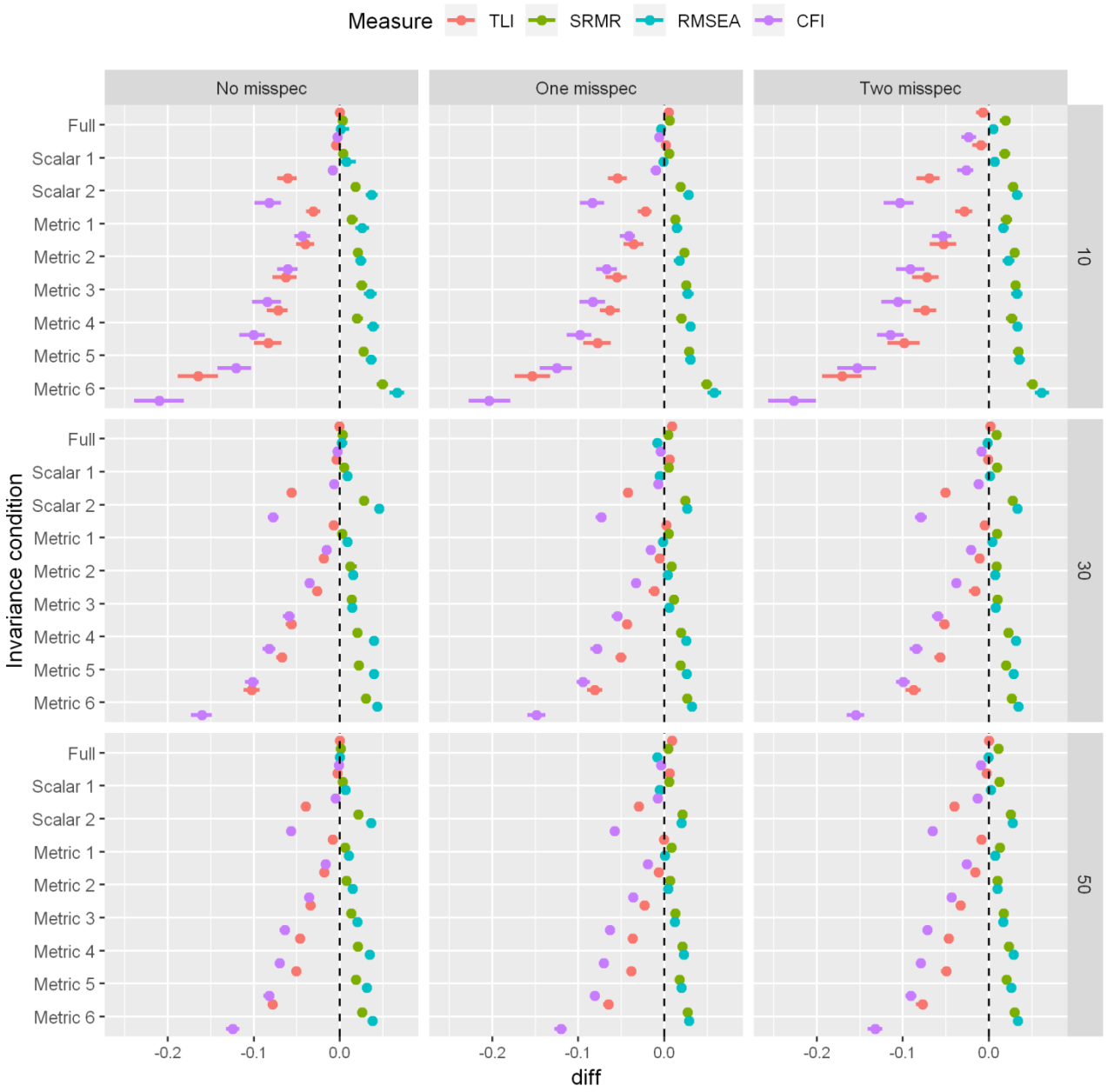

Figure 7 CFI, TLI, RMSEA, and SRMR differences. Metric vs. scalar model. MLR analysis

Note: Black dashed vertical line corresponds to Δ (scalar - metric)= 0.00. Dots show the average difference for each condition over 500 replications. Error bars show the 95% CIs.

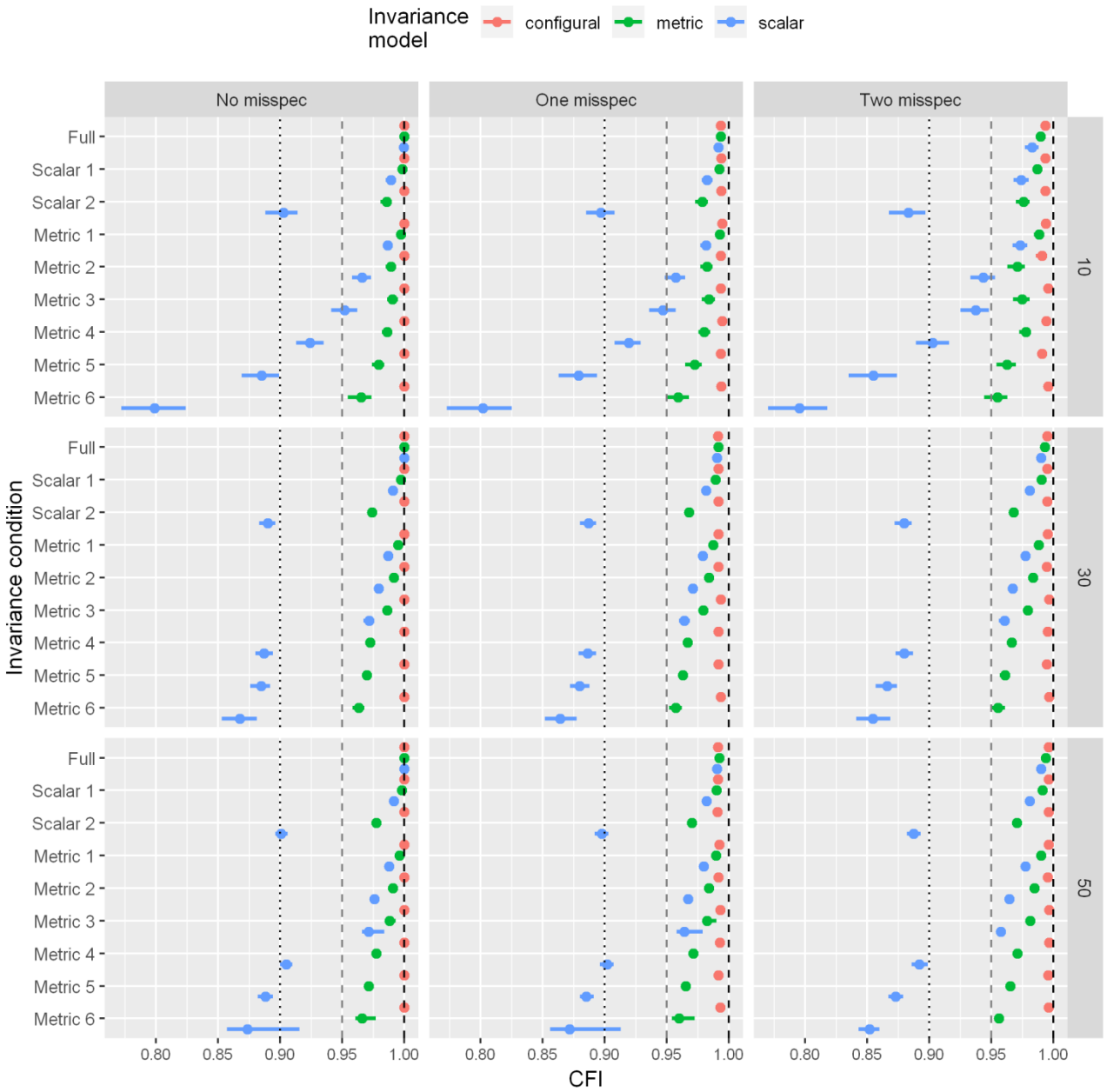

Figure 8 CFI values for configural, metric, and scalar models. WLSMV analysis

Note: Black dashed vertical line corresponds to CFI = 1.00. Dark grey dashed vertical line corresponds to CFI = 0.95. Dark grey dotted vertical line corresponds to CFI = 0.90. Dots show the average CFI values for each condition over all converged replications (out of 500). Error bars show the 95% CIs.

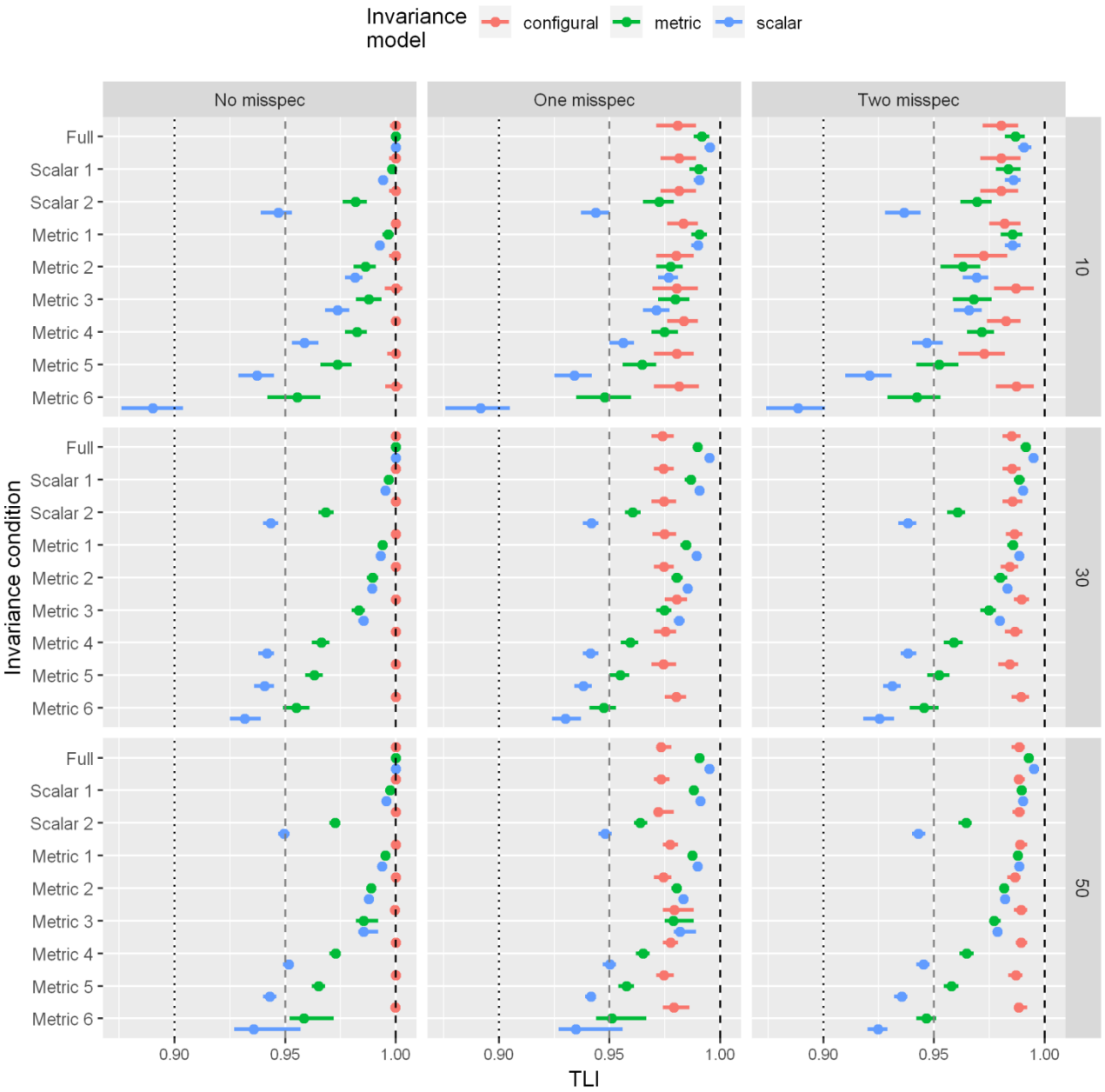

Figure 9 TLI values for configural, metric, and scalar models. WLSMV analysis

Note: Black dashed vertical line corresponds to TLI = 1.00. Dark grey dashed vertical line corresponds to TLI = 0.95. Dark grey dotted vertical line corresponds to TLI = 0.90. Dots show the average TLI values for each condition over all converged replications (out of 500). Error bars show the 95% CIs.

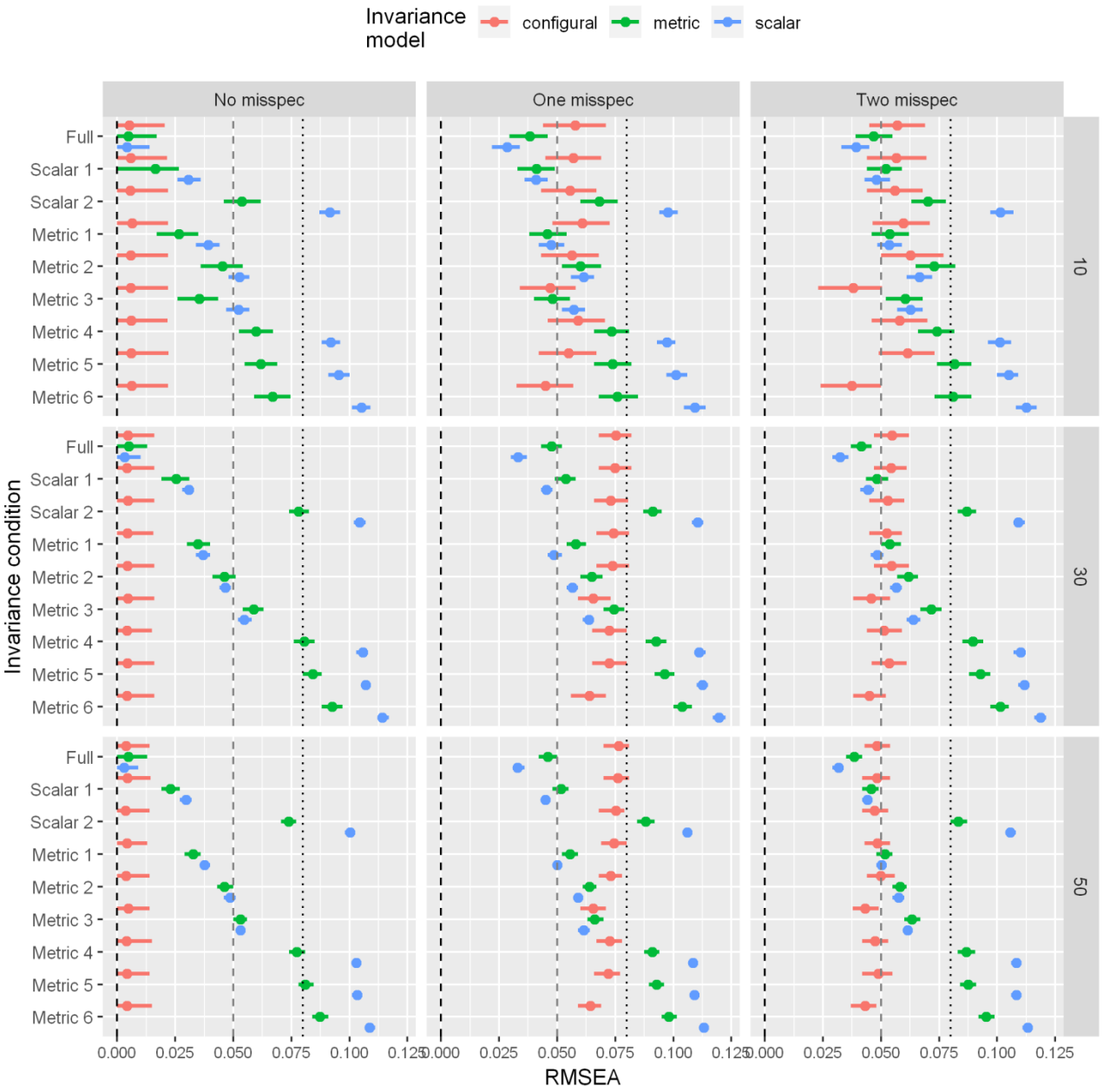

Figure 10 RMSEA values for configural, metric, and scalar models. WLSMV analysis

Note: Black dashed vertical line corresponds to RMSEA = 0.00. Dark grey dashed vertical line corresponds to RMSEA = 0.05. Dark grey dotted vertical line corresponds to RMSEA = 0.08. Dots show the average RMSEA values for each condition over all converged replications (out of 500). Error bars show the 95% CIs.

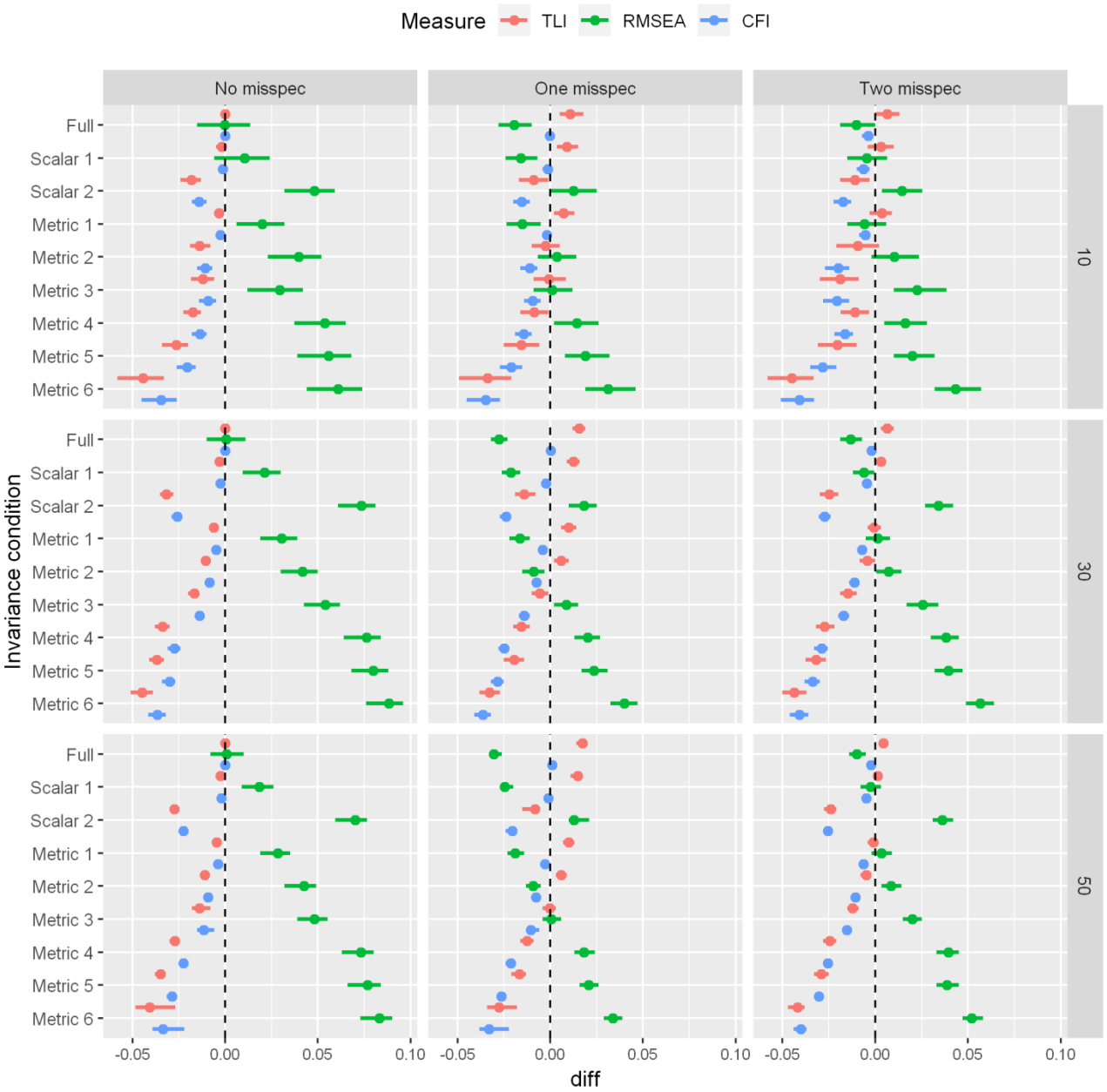

Figure 11 CFI, TLI, and RMSEA differences. Configural vs. metric model. WLSMV analysis

Note: Black dashed vertical line corresponds to Δ (metric - configural) = 0.00. Dots show the average difference for each condition over all converged replications (out of 500). Error bars show the 95% CIs.

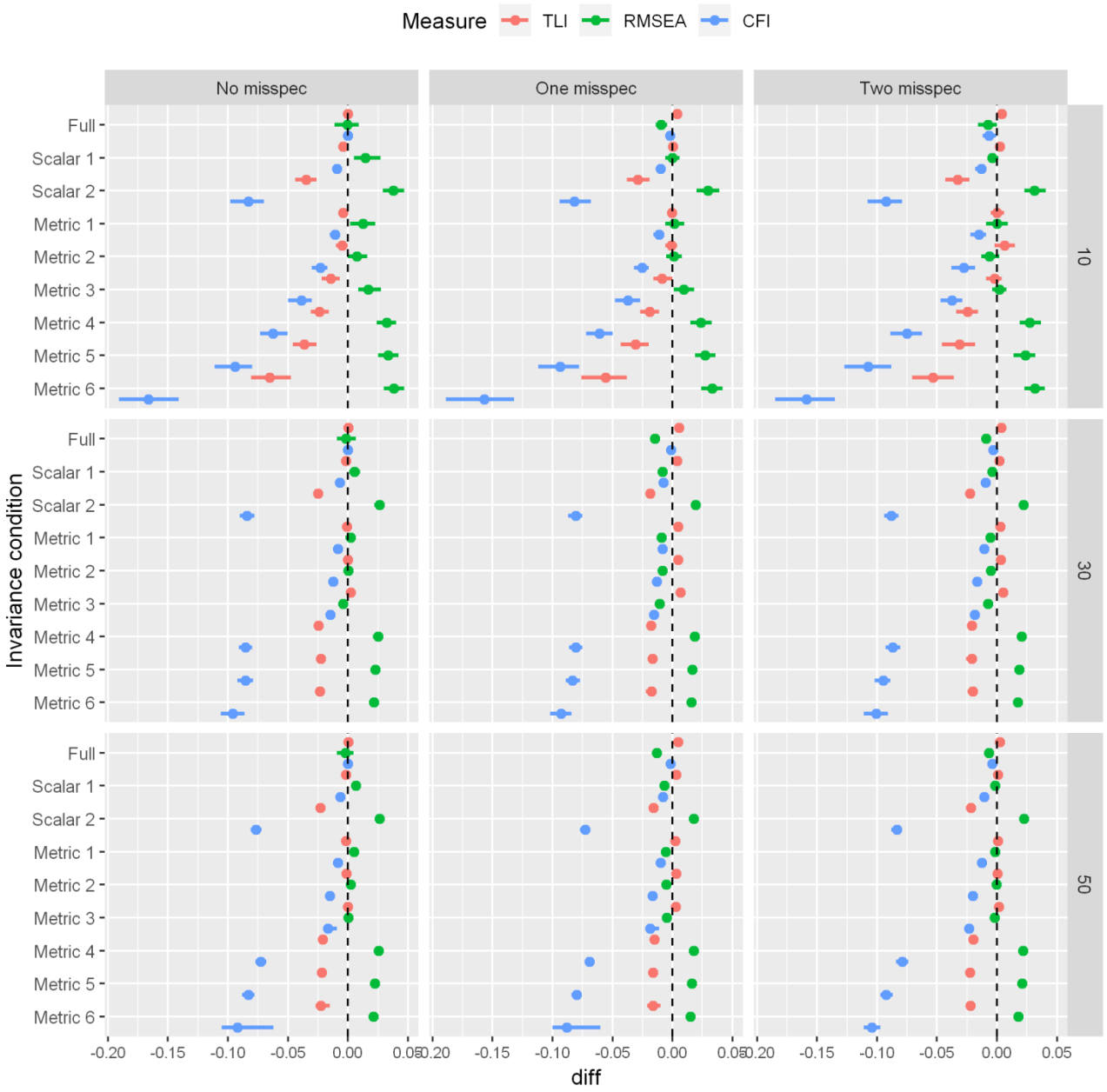

Figure 12 CFI, TLI, and RMSEA differences. Metric vs. scalar model. WLSMV analysis

Note: Black dashed vertical line corresponds to Δ (scalar - metric)= 0.00. Dots show the average difference for each condition over all converged replications (out of 500). Error bars show the 95% CIs.

Appendix

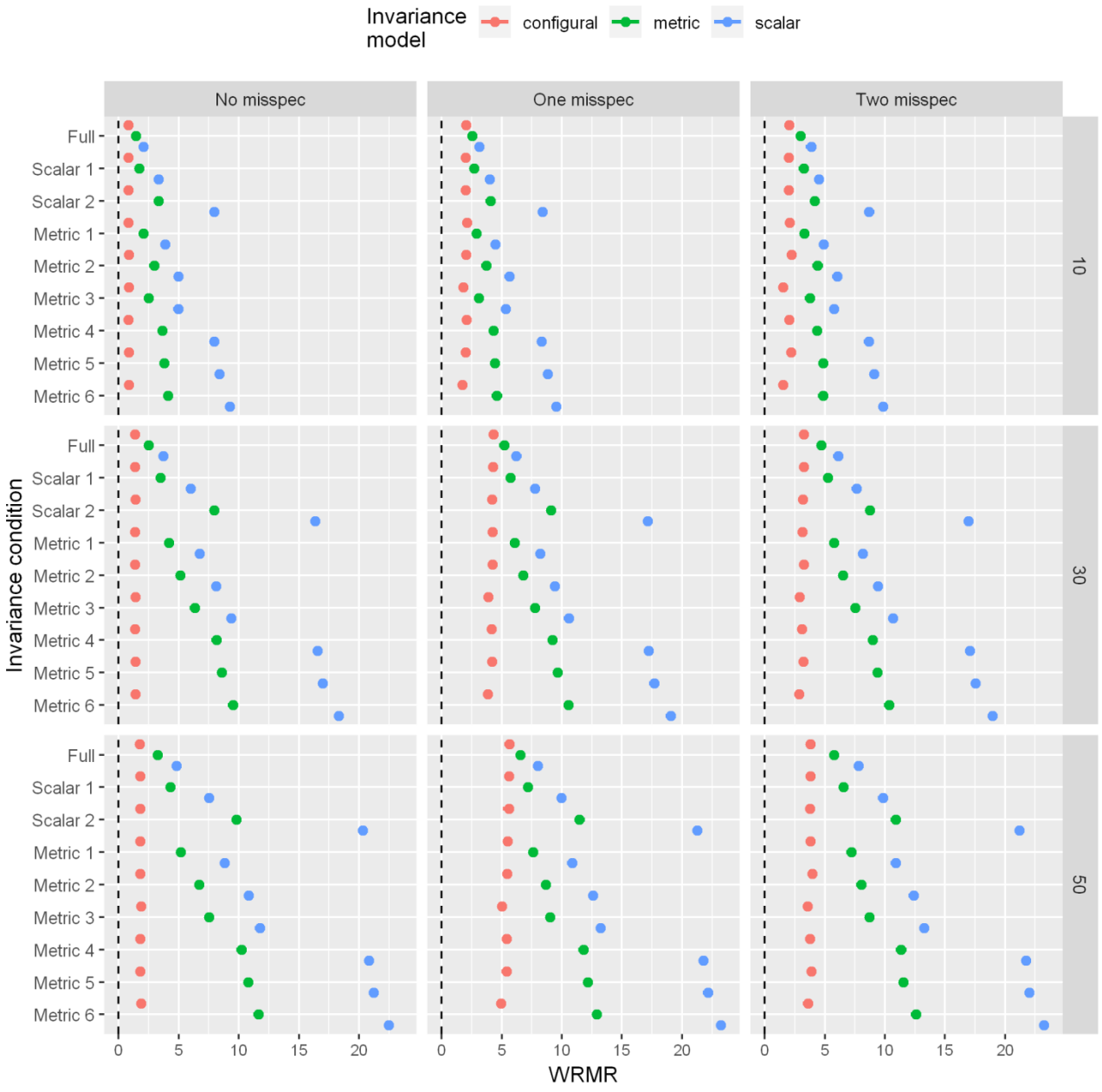

Figure A1 WRMR values for configural, metric, and scalar models. WLSMV analysis

Note: Black dashed vertical line corresponds to WRMR = 0.00. Dots show the average WRMR values for each condition over all converged replications (out of 500). Error bars show the 95% CIs.

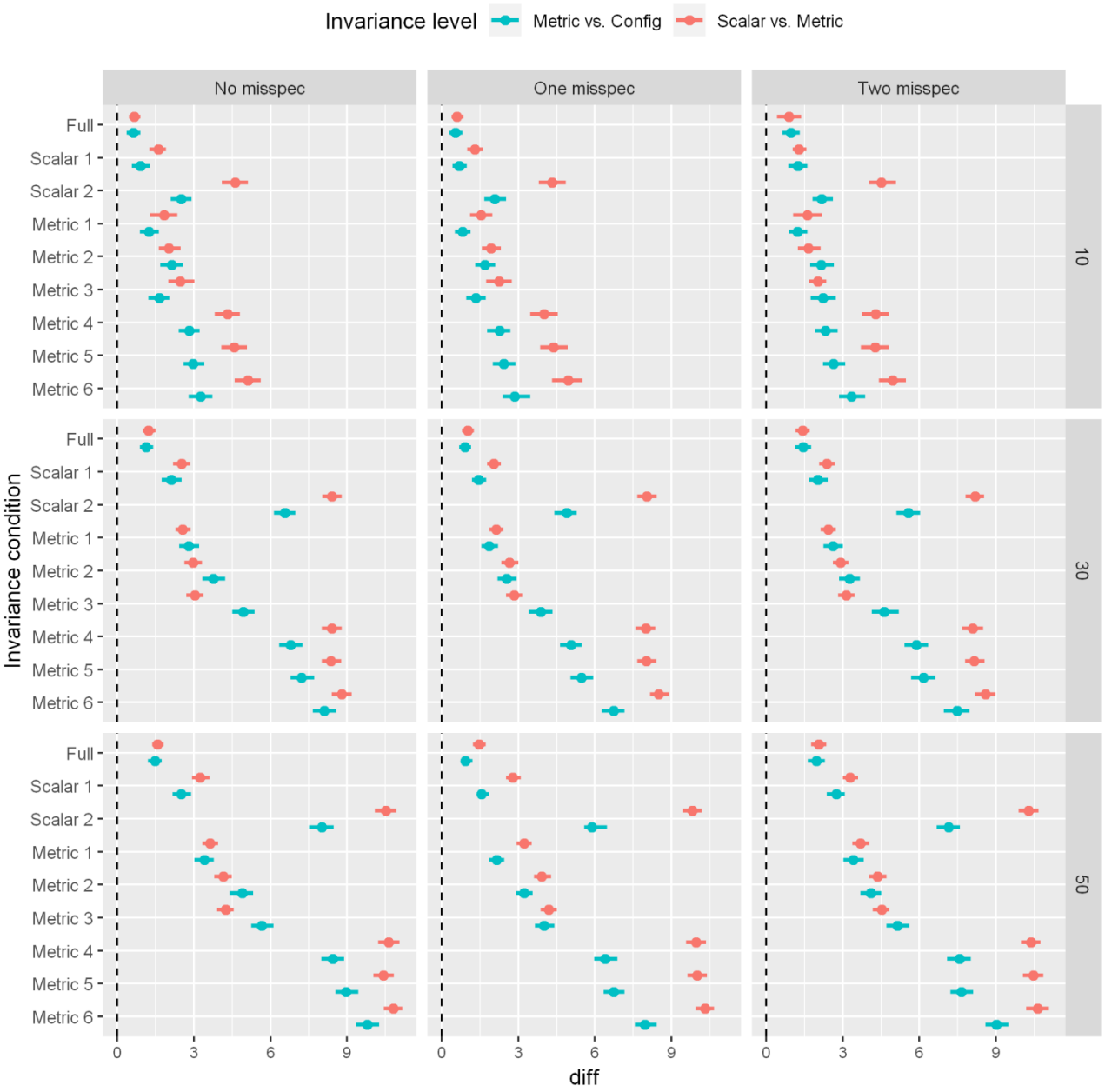

Figure 12 WRMR differences. WLSMV analysis

Note: Black dashed vertical line corresponds to ΔWRMR = 0.00. Dots show the average difference for each condition over all converged replications (out of 500). Error bars show the 95% CIs.

Boris Sokolov
National Research University Higher School of Economics
Laboratory for Comparative Social Research, Research Fellow
E-mail: bssokolov@gmail.com, bssokolov@hse.ru

Any opinions or claims contained in this Working Paper do not necessarily reflect the views of HSE.