Does acquiescence disagree with measurement invariance testing?

E. Damiano D'Urso[1], Jesper Tijmstra[1], Jeroen K. Vermunt[1], and Kim De Roover[1,2]

[1]Tilburg University, The Netherlands

[2]KU Leuven, Belgium

Abstract

Measurement invariance (MI) is required for validly comparing latent constructs measured by multiple ordinal self-report items. Non-invariances may occur when disregarding (group differences in) an acquiescence response style (ARS; an agreeing tendency regardless of item content). If non-invariance results solely from neglecting ARS, one should not worry about scale inequivalences but model the ARS instead. In a simulation study, we investigated the effect of ARS on MI testing, both when including ARS as a factor in the measurement model or not. For (semi-) balanced scales, disregarding a large ARS resulted in non-invariance already at the configural level. This was resolved by including an ARS factor for all groups. For unbalanced scales, disregarding ARS did not affect MI testing, and including an ARS factor often resulted in non-convergence. Implications and recommendations for applied research are discussed.

Keywords: measurement invariance (MI), multiple group categorical confirmatory factor analysis (MG-CCFA), acquiescence response style (ARS), psychometrics.

Does acquiescence disagree with measurement invariance testing?

1 Introduction

Social and behavioral scientists are often interested in assessing whether groups of individuals differ regarding latent constructs (e.g., extraversion). These unobservable constructs are often measured by self-report scales. Commonly, these scales consist of questionnaire items, where, for each item, respondents rate their level of agreement by selecting one of a few ordered response options on a Likert scale (e.g., "disagree ", "neutral ", "agree "). To validly draw conclusions about group differences on latent constructs, scales must function equivalently across groups. Measurement invariance (MI) testing evaluates the tenability of this hypothesis by assessing whether the measurement model (MM) of the psychological construct is equivalent across groups. As an example of an inequivalence, one may think of differences in item interpretations that may lead one group to systematically pick lower/higher response options for some items, which can result in under/overestimation of sum-scores (Jeong & Lee, 2019), item means (Jones & Gallo, 2002), and regression parameters in structural equation models (Guenole & Brown, 2014). Thus, testing for MI is an essential precursor to investigating group differences (Borsboom, 2006; Meredith & Teresi, 2006) to avoid building on latent construct differences that are purely due to measurement discrepancies and thus invalid or "biased".

Measurement invariance is often tested with a latent variable approach, which models the relationship between unobserved psychological constructs (i.e., latent variables) and observable behaviors (i.e., items). Within the latent variable modeling framework, multiple group categorical confirmatory factor analysis (MG-CCFA) and multiple group item response theory (MG-IRT) are the most popular approaches to evaluate MI for models with ordinal data (i.e., items with too few response categories to treat them as continuous). Note that while equivalences can be drawn between MG-CCFA and MG-IRT models (Chang, Hsu, & Tsai, 2017), some differences remain in the way MI is tested within each of these two approaches (D'Urso, De Roover, Vermunt, & Tijmstra, 2021). For instance, MG-CCFA primarily focuses on assessing MI at the scale level (i.e., for the complete set of items measuring a construct), whereas MG-IRT traditionally tests

MI for each item separately. In this paper, we focus on a MG-CCFA-based MI testing approach (i.e., scale level) since it is more commonly used in practice (Putnick & Bornstein, 2016). Specifically, in a MG-CCFA-based MI testing approach, different levels of MI are assessed in a step-wise procedure. For each step, increasingly restrictive models are estimated by imposing equality constraints on specific MM parameters. Then, the fit of the more constrained model to the data is compared to that of the less constrained one to evaluate whether the equality constraints worsen the model fit significantly, thus indicating non-invariance of (at least some of) the constrained MM parameters.

Apart from measurement model differences or "non-invariances" such as differential item interpretations, bias in latent variable comparisons may also arise when responses to self-report item rating scales are affected by response tendencies or response styles in some groups but not in others (Cheung & Rensvold, 2000). Acquiescence, or agreeing, response style (ARS) is a well-known one, which represents a tendency to agree with items regardless of their content (Paulhus, 1991). Interestingly, various studies have indicated that different groups of individuals may have a more or less pronounced ARS depending on their education (Meisenberg & Williams, 2008), age (Weijters, Geuens, & Schillewaert, 2010), gender (Austin, Deary, & Egan, 2006), length of employment (Johnson, Kulesa, Cho, & Shavitt, 2005) or culture (Bachman & O'Malley, 1984; Marin, Gamba, & Marin, 1992). ARS can inflate observed means (Van Vaerenbergh & Thomas, 2013) and affect the measurement model by introducing an additional factor (Billiet & McClendon, 2000; D'Urso, Tijmstra, Vermunt, & De Roover, 2022) or changing the strength of the relationships between items and factors (i.e., factor loadings; Ferrando & Lorenzo-Seva, 2010). To control for ARS, previous research indicated that including ARS as an additional factor in the MM (Billiet & McClendon, 2000) proved to effectively reduce bias in MM parameters recovery as well as estimated factor scores (Savalei & Falk, 2014).

Though it is confirmed that not taking ARS into account affects the MM in single group studies, extensive investigations about the impact of disregarding ARS on MI testing are currently lacking. Existing studies in the literature have either focused on assessing

the effect of other RSs on MI, such as extreme response style (ERS; Liu, Harbaugh, Harring, & Hancock, 2017) and non-effortful responding (NER; Arias, Garrido, Jenaro, Martínez-Molina, & Arias, 2020; Rios, 2021), or on evaluating the impact of including an additional ARS factor when assessing MI using empirical data (Welkenhuysen-Gybels, Billiet, & Cambré, 2003; Aichholzer et al., 2015). In case of empirical data, the "true" MM is unknown, however. Identifying in which conditions and to what extent ARS distorts measurement invariance conclusions may give us clues on how to correct for bias in latent variable score differences that is due to differential response tendencies across groups. For instance, when the influence of ARS is disregarded, researchers may wrongly conclude that there is non-invariance. Taking ARS into account when testing for MI likely facilitates distinguishing non-invariance of the scale itself from (amendable) non-invariance due to disregarding ARS. Indeed, if non-invariance results from not taking ARS into account, one needs to correct for the ARS instead of worrying about the scale being inequivalent across groups. Also, the lack of systematic investigations regarding the effect of ARS on MI testing under various conditions leaves the following questions unanswered. Firstly, ARS is often viewed as a weak factor (i.e., with all items having weak to moderate loadings on this factor; Ferrando, Condon, & Chico, 2004; Danner, Aichholzer, & Rammstedt, 2015), which may be insufficient to result in significant violations of MI (i.e., rejection of MI). Secondly, measurement non-invariance may be more easily detected in the case of balanced scales (i.e., scales where content factors are measured by an equal amount of positively and negatively worded items), since these allow distinguishing the ARS from the latent variables one wants to measure. The remainder of this paper proceeds as follows: In Section 2, we elaborate on MG-CCFA, MI testing, and how it may be affected by ARS. Then, in Section 3, we present a simulation study that evaluates the effect of ARS on MI both when (i) ARS is disregarded, and (ii) ARS is taken into account by including ARS as an additional factor in the MM. Finally, in Section 4, we discuss recommendations based on the simulation study results, limitations of our investigation, and potential future research directions.

2   Measurement invariance testing and the potential effects of ARS

In this section, we introduce MG-CCFA, describe the standard MI testing framework including the identification constraints proposed by Wu and Estabrook (2016) to assess MI with ordinal data, and discuss the potential effects of ARS on MI testing.

2.1   Multiple group categorical confirmatory factor analysis

Consider having data composed of J items for a group of N subjects, and that a grouping variable (e.g., nationality) exists to divide the N subjects into G groups. Then, let x_j be the polytomously scored response on item j that can take on C possible values with c = {0,1,2,...,C-1}. MG-CCFA assumes that each of the C possible observed responses is obtained from a discretization of a continuous unobserved response variable x_j^* through a set of threshold parameters $\tau_{j,c}^{(g)}$, which indicate the cut-off point for the response categories (e.g., division between scoring a 1 or a 2) for group g. Note that the first and last thresholds are defined as $\tau_{j,0}^{(g)} = -\infty$ and $\tau_{j,C}^{(g)} = +\infty$, respectively. Then, formally

$$x_j = c, \quad if \quad \tau_{j,c}^{(g)} < x_j^* < \tau_{j,c+1}^{(g)} \quad c = 0, 1, 2, ..., C - 1. \tag{1}$$

A factor-analytical model for a vector of latent response variables $\boldsymbol{x}^* = (\boldsymbol{x_1^*}, \boldsymbol{x_2^*}, ..., \boldsymbol{x_J^*})$ is obtained as:

$$\boldsymbol{x}^* = \boldsymbol{\nu}^{(g)} + \boldsymbol{\Lambda}^{(g)}\boldsymbol{\eta}^{(g)} + \boldsymbol{\epsilon}^{(g)}. \tag{2}$$

where $\boldsymbol{\nu}^{(g)}$ is a J-dimensional vector of latent intercepts (i.e., intercepts of the unobserved response variables in $\boldsymbol{x}^*$), $\boldsymbol{\Lambda}^{(g)}$ is a $J \times Q$ matrix of factor loadings, $\boldsymbol{\eta}^{(g)}$ is a Q-dimensional vector of scores on the Q factors, $\boldsymbol{\epsilon}^{(g)}$ is a J-dimensional vector of residuals. Note that the latent intercepts $\boldsymbol{\nu}_{\boldsymbol{j}}^{(g)}$, thresholds $\boldsymbol{\tau}_{\boldsymbol{j,c}}^{(g)}$ and loadings $\lambda_j^{(g)}$ in $\boldsymbol{\Lambda}^{(g)}$ are group specific, and that, within each group g, both the factors $\boldsymbol{\eta}$ and the item-specific residual components $\boldsymbol{\epsilon}$ are mutually independent and normally distributed, with:

$$\boldsymbol{\eta}^{(g)} \sim MVN(\boldsymbol{\kappa}^{(g)}, \boldsymbol{\Phi}^{(g)}), \quad and \quad \boldsymbol{\epsilon}^{(g)} \sim MVN(\boldsymbol{0}, \boldsymbol{\Psi}^{(g)}). \tag{3}$$

where $\boldsymbol{\kappa}^{(g)}$ are the group-specific factor means, $\boldsymbol{\Phi}^{(g)}$ the group-specific factors variance-covariance matrix, and $\boldsymbol{\Psi}^{(g)}$ is a diagonal matrix containing the group-specific unique variances of the items. Further, within each group, the model-implied mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is obtained as:

$$\boldsymbol{\mu}^{(g)} = \boldsymbol{\nu}^{(g)} + \boldsymbol{\Lambda}^{(g)}\boldsymbol{\kappa}^{(g)}, \quad \boldsymbol{\Sigma}^{(g)} = \boldsymbol{\Lambda}^{(g)}\boldsymbol{\Phi}^{(g)}\boldsymbol{\Lambda}^{(g)} + \boldsymbol{\Psi}^{(g)}. \tag{4}$$

2.2 Measurement invariance testing procedure

In MG-CCFA, MI is commonly evaluated by testing, for all items, equality of a set of MM parameters (e.g., loadings) across groups in a step-wise fashion. The starting point is to identify the MG-CCFA model, which requires: (i) setting the scale for the latent variable η, (ii) setting the scale for the unobserved response variable x_j^*, and (iii) aligning the scale of the latent variable η across groups. Note that the latter is necessary to make the groups comparable. Then, in addition to the identification constraints, additional equivalence constraints are imposed on MM parameters (e.g., thresholds) in a step-wise fashion to evaluate their invariance. Therefore, a new, more constrained model is estimated for each step, and its fit to the data is evaluated to conclude whether these new constraints significantly worsen the fit. Below, we first discuss the main steps and identification constraints to test MI for ordinal data following the recommendations by Wu and Estabrook (2016), summarized in Table 1, and then elaborate on standard goodness-of-fit criteria that are used to draw MI conclusions.

2.2.1 Configural invariance. Configural invariance is usually the first invariance level tested, where the goal is to test the equivalence of the number of factors and of the loadings pattern (i.e., which factors are measured by which items) across groups. In this step, following Wu and Estabrook (2016), the baseline model is identified by fixing, for all groups, the latent intercepts $\boldsymbol{\nu}$ to 0 and variances (i.e., diagonal elements of $\boldsymbol{\Sigma}$) to 1, which is commonly known as the delta parameterization (Muthén & Muthén, 2009). Similarly, the latent factor means $\boldsymbol{\kappa}^{(g)}$ and variances $\phi^{(g)}$ (i.e., diagonal elements of $\boldsymbol{\Phi}^{(g)}$) are also fixed to 0 and 1, respectively.

After specifying and estimating this factor model for all groups, conclusions on configural invariance are drawn following the examination of goodness-of-fit measures. If supported, configural equivalence indicates that the shape of the model (i.e., number of factors and pattern of zero and non-zero loadings) is the same across groups.

2.2.2 Thresholds invariance. If configural invariance holds, the invariance of thresholds is tested next. Here, the baseline model is identified by setting, for all groups, the latent content factor means $\boldsymbol{\kappa}^{(g)}$ and variances $\phi^{(g)}$ to 0 and 1, respectively. Additionally, for the reference group r, the vector of latent intercepts $\boldsymbol{\nu}^{(r)}$ and latent response variable variances in $\boldsymbol{\Sigma}^{(r)}$ are set to 0 and 1, respectively. On top of these identification constraints, thresholds $\tau_{j,c}$ are equated across groups and, after model estimation, the hypothesis of thresholds invariance is evaluated by evaluating the change in model fit between the configural model and the thresholds invariant model.

2.2.3 Loadings invariance. If thresholds invariance holds, invariance of loadings is assessed. To identify the baseline model, for all groups, the latent content factor means $\boldsymbol{\kappa}^{(g)}$ are set to 0, while, for the reference group r, the factor variances $\phi^{(r)}$ are set to 1, the latent intercepts $\nu^{(r)}$ to 0 and variances in $\Sigma^{(r)}$ to 1. In addition to these identification constraints, both thresholds $\tau_{j,c}$ and loadings $\boldsymbol{\Lambda}$ are constrained to be equal across groups. Again, the model is estimated and the hypothesis of loadings invariance is evaluated by assessing the change in model fit between the thresholds invariant model and the loadings invariant model. Note that, if the hypothesis of thresholds and loadings invariance holds, factor variances can be validly compared across groups.

2.2.4 Intercepts invariance. Finally, if loadings invariance holds, invariance of latent intercepts is assessed. To identify the baseline model, for the reference group, the latent content factor means $\boldsymbol{\kappa}^{(r)}$ and variances $\phi^{(r)}$ are set to 0 and 1, respectively. Additionally, building on the previous equality constraints on thresholds and loadings, the latent intercepts $\boldsymbol{\nu}$ are set to 0 and equated across groups. To assess the hypothesis of latent intercepts invariance, the model is estimated and its fit is compared to the loadings invariant model. Following non-rejection of latent intercepts invariance, the factor means can be validly compared across groups.

2.2.5 Criteria to assess model fit. Goodness-of-fit indices are commonly used as criteria to assess the tenability of MI hypotheses. This commonly entails evaluating the fit of the baseline model (i.e., configural model) and then the change in fit for the more restrictive models. To aid conclusions on whether the (change in) fit allows to conclude that a certain level of invariance (e.g., thresholds) holds, various criteria are inspected, each with its own proposed cut-off value determined via extensive simulation studies. Classically, only the chi-squared χ^2 test was used as a criterion to assess the significance of change for two nested models (Putnick & Bornstein, 2016) but multiple studies have shown that relying solely on this statistic is sub-optimal due to it being extremely sensitive to negligible MM differences in large samples (Bentler, 1990; French & Finch, 2006; French & Finch, 2008). Therefore, in practice, MI decisions are based on multiple criteria (Putnick & Bornstein, 2016), and, among them, two of the most commonly used are the root mean square error of approximation (RMSEA; Browne, Cudeck, et al., 1993) and the comparative fit index (CFI; Bentler, 1990). Configural invariance is concluded if RMSEA $\leq$ .06 and/or CFI is $\geq$ .95 (Brown, 2015). For the more restrictive models, the change in fit (e.g., ΔRMSEA) is assessed to conclude whether the additional constraints worsen the fit significantly. Cheung and Rensvold (2002) suggested to conclude non-invariance when ΔCFI $\leq$ -.01, while Chen (2007) recommended ΔRMSEA $\geq$ .01. Note that various criteria have been suggested for different fit measures, and we refer the reader to Svetina, Rutkowski, and Rutkowski (2020) for an overview. Also, while recent research has indicated that model-specific cut-off values may be generally preferred to evaluate model fit (McNeish & Wolf, 2021), there are no available guidelines for calculating these cut-offs for MI testing in ordered-categorical data.

2.3 From single group to multiple groups: the potential effects of (not) correcting for ARS.

In the literature, the bias resulting from disregarding ARS for single-group analyses is well-known but it is not yet clear to what extent it may generalize to multiple-group analyses, such as MG-CCFA. Response tendencies, such as ARS, represent sources of

systematic response bias that may appear as violations of measurement invariance (i.e., measurement non-invariance). When ARS affects individuals' responses in one of the groups, not taking into account this tendency towards acquiescence likely results in systematic differences in the responses across groups that are not purely due to the intended-to-be-measured (i.e., content) factors and, eventually, may lead to the rejection of measurement invariance. For instance, in single group studies, one well-known consequence of not accounting for ARS is that it may result in an additional factor (Billiet & McClendon, 2000; D'Urso et al., 2022). Therefore, it is reasonable to expect that researchers unaware of the (potential) influence of ARS would disregard this and reject configural invariance, which would lead them to conclude that the content factor(s) cannot be validly compared across groups since they (seem to) qualitatively differ. Additionally, single group studies showed that ARS can bias item (latent) intercepts (Cheung & Rensvold, 2000), and factor loadings (Ferrando & Lorenzo-Seva, 2010; Savalei & Falk, 2014; D'Urso et al., 2022). Again, neglecting this agreeing response tendency may result in non-equivalence (i.e., non-invariance) of intercepts and/or loadings, and lead researchers to conclude that the MM is non-invariant and, potentially, to allow some parameters to freely vary across groups (i.e., partial invariance) to reach an acceptable level of invariance before investigating differences in the content factor(s).

Another important aspect to consider is that the performance of psychometric approaches developed to correct for ARS has not been thoroughly investigated in the context of MI testing. Savalei and Falk (2014) have discussed some of the main factor-analytical approaches to correct for ARS, their underlying assumptions and compared their performance for single group analyses through a simulation study. The results have shown that the classical CFA-based approach (Billiet & McClendon, 2000), where ARS is specified as an additional factor orthogonal to the content factor(s) with all loadings set to 1 (i.e., the influence of ARS does not vary across items) outperformed the remaining ones[1],

---

[1]The other approaches discussed by Savalei and Falk (2014) are the Chan and Bentler (1993) approach and the EFA-based approach (Ferrando et al., 2004). In the former, data must be first mean-centered within person (i.e., ipsatized). Then, a residual structure must be specified by adding a linear combination

even when some of its main assumptions (e.g., equal ARS loadings) are violated. Thus, based on the authors' results, recommendations, and its straightforward implementation to MG-CCFA, we will mainly focus on this CFA-based approach in the remainder of this paper. Specifically, following this CFA-based approach, an additional ARS factor is added to the MM in all groups and all factor loadings on this additional factor are fixed to 1, which allows to freely estimate the ARS factor variance for all groups. Then, between-group differences in the amount of ARS are captured by differences in the ARS factor means, and within-group differences in the strength of ARS are captured by the ARS factor variances.

3   Simulation study

To assess the effect of ARS on MI testing, both when including ARS as an additional factor in the measurement model (MM) and when not including it, we conducted a simulation study where individual responses in one group were affected by an ARS. Our goal is to solely focus on whether the bias introduced by disregarding ARS results in measurement non-invariance, and whether this is rectified by including an additional ARS factor in the MM for all groups. Therefore, we did not simulate other sources of non-invariance (e.g., differences in factor loadings). Furthermore, a null scenario was simulated, where invariance holds and ARS is not at play for both groups, which only served as a comparison for evaluating the performance of MG-CCFA approaches. Note that we report these latter results in the Appendix (Tables A1 - A3).

The following 5 factors were manipulated:

- The number of subjects N within each group at 2 levels: 250, 1000;

- The type of scale at 3 levels: balanced, semi-balanced and unbalanced;

- The number of content factors Q at 2 levels: 1, 2;

———

of the original residual components for each of the ipsatized variables. In the latter, an additional factor is first extracted, and then a rotation is performed to a partially specified target, which allows to estimate both content and ARS factor loadings.

- The scale length (i.e., total number of items) at 2 levels: 12, 24;

- The strength of the ARS factor at 2 levels: medium and large.

For the minimum sample size within each group, we followed the recommendations from previous research, which indicated that for obtaining precise factor loading estimates a sample size of 250 is sufficient when item communalities are moderate (Fabrigar, Wegener, MacCallum, & Strahan, 1999; MacCallum, Widaman, Zhang, & Hong, 1999). Furthermore, we varied (a) the number of factors to simulate both unidimensional and multidimensional scales, (b) the total number of items to simulate scales that measure the psychological construct to a varying degree of accuracy, and (c) the type of scale (e.g., semi-balanced) to emulate scales that allow disentangling the content factor(s) from the ARS factor to a different extent (Savalei & Falk, 2014; de la Fuente & Abad, 2020).

In terms of the performance of MI testing, we hypothesize the following: violations of MI (i.e., non-invariance) will likely be detected when ARS is large and ignored (i.e., not included as an additional factor for both groups). Specifically, we expect that, for balanced and semi-balanced scales, disregarding a large ARS will result in non-invariance at all levels. For unbalanced scales, violations of MI may not be detected since ARS will likely affect structural rather than measurement parameters, like the covariance among content factors in case of a multidimensional scale (D'Urso et al., 2022), or factor variances, especially in the conditions with unidimensional scales (Ferrando & Lorenzo-Seva, 2010). In addition, we expect that including the additional ARS factor for both groups will allow for MI to be established in the case of balanced and semi-balanced scales. Finally, in the conditions with unbalanced scales, we hypothesize that including an additional ARS may result in model estimation issues since ARS cannot be easily disentangled from the content factor(s).

A full-factorial design was used with 2 (number of subjects) x 3 (type of scale) x 2 (number of content factors) x 2 (number of items) x 2 (strength of ARS) = 48 conditions. For each condition, 100 replications were generated, resulting in 4,800 data sets.

3.1 Methods

3.1.1 Data generation. Data were generated from a factor model with one or two factors and two groups, and the model parameters are displayed in Table 2. To simulate balanced scales, for the content factor(s), half of the loadings were positive (i.e., indicative items) and the other half were negative (i.e., contra-indicative items), whereas 33% and none of the loadings were negative for semi-balanced scales and unbalanced scales, respectively. Note that, for both groups, 0 and 1 were used as generating values for the content factor(s) means and variances, respectively. As displayed in Table 2, we simulated ordinal items with 5 categories and the distance between the first threshold of the easiest and the most difficult item was 2 standard deviations. To avoid estimation issues (e.g., non-convergence), we only retained data sets where each category for each item contains at least a single observation. In the rare cases where, for a specific item, a category was not observed among the generated scores, we repeated the data generation process until all response categories were observed.

We sampled the ARS factor scores from a right-censored normal distribution to match an agreeing tendency closely. Employing this distribution, we only simulated subjects who did or did not show an ARS (i.e., have a positive or zero factor score on the ARS dimension) without allowing for scores to represent a disagreeing tendency (i.e., a negative factor score). For simulating the effect of ARS on the item responses, we used loading values of .3 and .6 for the medium and large ARS scenario, respectively[2]. Note that, for the reference group, the ARS factor scores were simulated to be 0 for all subjects (i.e., ARS did not affect the item responses).

———

[2]The variance of a right-censored normal distribution is smaller than the identification restrictions imposed to set a scale for the variance of the ARS factor (i.e., fixing all its loadings to 1). In Table 2 we report the value of the original loadings on the ARS factor multiplied by the standard deviation of a right-censored normal distribution, which is $\approx$.583. This results in loadings on the ARS factor of .175 and .350 for medium and large ARS conditions, respectively. Note that these values match those used in previous studies to simulate ARS factor loadings that can be realistically expected in well-designed measures (Ferrando & Lorenzo-Seva, 2010; Danner et al., 2015).

3.1.2 Data Analysis. To simulate the effect of ignoring or including ARS when testing for MI, we considered two different MMs - and, thus, performed two different MG-CCFA analyses - for each replication, that is, with or without an additional ARS factor. For the latter, we used the standard CFA-based approach proposed by Billiet and McClendon (2000), where an additional ARS factor is specified with all loading on this factor fixed to 1. Note that no additional constraints are imposed on the content factor loadings nor on the variance of the ARS factor under this model. To identify the MG-CCFA models, we followed the Wu and Estabrook (2016) identification constraints for MI testing described in Section 2 for both the model with and without the ARS factor.

All MG-CCFA models were estimated using diagonally weighted least squares (DWLS), but the full weight matrix was used to compute the mean-and-variance-adjusted test statistics (default in *lavaan*; Rosseel, 2012). DWLS is a two-step estimation procedure, where the thresholds and polychoric correlation matrices for the groups are estimated in the first step, and, in the second step, the remaining parameters are estimated using the polychoric correlation matrices from the previous step.

3.1.3 Outcome measures. After fitting the models, we evaluated both the convergence rate (CR) and the performance of different model fit criteria. For the latter, we recorded the results obtained from the χ^2 test, the root mean square error of approximation (RMSEA; Browne et al., 1993) and the comparative fit index (CFI; Bentler, 1990) and we averaged across replications in a cell of the factorial design. In empirical practice, decisions about MI results are often dichotomous (i.e., invariant or not). Thus, we also calculated the false positive rate (FPR) for the different goodness-of-fit criteria, which is here defined as flagging the scale as non-invariant[3]. Specifically, configural non-invariance was concluded if: χ^2 test was significant ($\alpha = 0.05$), RMSEA > .06, CFI < .95. In addition, since common guidelines suggest to base invariance decisions on different goodness-of-fit indices, we created a combined criterion and concluded configural

———

[3]Note that, outside of the null condition, this is not formally a FPR. In fact, only when considering the ARS factor can we really say that the scale is invariant, whereas disregarding it results in between the scales.

invariance if both a significant χ^2-difference test and at least one between RMSEA < .06 and CFI > .95 was observed (Putnick & Bornstein, 2016). We compared the fit between the configural and the thresholds invariant models for thresholds invariance, between the threshold and loadings invariant models for loadings invariance, and between loadings and intercepts invariant models for intercepts invariance. For all these comparisons, non-invariance was concluded if: χ^2-difference test was significant ($\alpha = 0.05$), ΔRMSEA > .01, ΔCFI < -.01. Finally, for the combined criterion, non-invariance was concluded if we observed both a significant χ^2-difference test and at least one between ΔRMSEA > .01 and ΔCFI < -.01. Since we deem the results on the values of or differences in the fit indices to be more informative than these dichotomized results, we only display the latter in the Appendix (Tables A4 - A7).

3.1.4 Data simulation, softwares and packages. The data were simulated and analyzed using R (R Core Team, 2013). Specifically, for estimating MG-CCFA models and obtaining fit measures, we used the R package *lavaan* (Rosseel, 2012), while for specifying the MG-CCFA models we used the *semTools* package (Jorgensen et al., 2016).

3.2 Results

3.2.1 Without ARS factor.

3.2.1.1 Convergence. In the appendix, table **??** displays the convergence results when ARS is not included as an additional factor (i.e., disregarded). The convergence rate was always 100% across conditions for all MG-CCFA models and for both unidimensional and multidimensional scales. Therefore, disregarding the influence of ARS when assessing different levels of MI does not seem to affect model convergence.

3.2.1.2 MI testing. The average fit measures results obtained when evaluating MI for unidimensional and multidimensional scales are displayed in Table 3 to 6, respectively. The results indicate that the ARS strength and the type of scale were the most relevant design factors affecting the MI testing results. In fact, for both unidimensional and multidimensional scales, ignoring the influence of ARS deteriorated models fit at all MI levels, and especially in the conditions with large ARS and balanced or semi-balanced scales. In these conditions, the RMSEA was often > .10 and the CFI < .90, which, in

empirical practice, are commonly interpreted as "unacceptable" fit values. Note that, when the influence of ARS was small (i.e., $\lambda_{ARS} = .175$), model fit was often good (i.e., RMSEA $<.06$ and CFI $> .95$), which is line with previous research indicating that when loadings on the ARS factor are small (i.e., $\approx .1$) ignoring ARS does not seem to strongly affect the MM parameters recovery (Savalei & Falk, 2014). The fit measure values were good (i.e., RMSEA $< .06$ and CFI $> .95$) in the conditions with unbalanced scales regardless of the strength of the ARS factor and the MI level tested. Again, these results partially overlap with previous studies, indicating that the ARS factor gets absorbed by the content factors for unbalanced scales. Therefore, for unbalanced scales, one may conclude that MI holds even when one group has a strong agreeing tendency. Bear in mind that, for these scales, the bias introduced by an ARS does not seem to affect MI testing results but it may affect factor scores or factor covariances (e.g., see Savalei & Falk, 2014) for some groups, and thus (potentially) substantive conclusions. Concerning the dichotomized results (Tables A4 - A5 in the Appendix), almost all the considered criteria resulted in a close-to-one FPR when testing configural and intercepts invariance for balanced and semi-balanced scale, whereas for unbalanced scales the FPR was often close to 0.

3.2.2   With ARS factor.

3.2.2.1   Convergence.   Table 7 displays the model convergence results when including an additional ARS factor in the MM. Convergence was strongly affected by the type of scale. In fact, for unbalanced scales, the convergence rate was lower than in the conditions with (semi-) balanced scales and especially low when testing for configural invariance. Therefore, one may often fail to evaluate configural invariance when including ARS for an unbalanced scale in empirical practice. Note that this is likely caused by the fact that the ARS factor cannot be distinguished from the content factor(s), which is corroborated by previous research indicating that, in EFA, including an additional ARS factor for unbalanced scales may result in factor underextraction (Ferrando & Lorenzo-Seva, 2010; D'Urso et al., 2022).

3.2.2.2 MI testing. Tables 8 to 11 display the MI testing results when an additional ARS factor is included in the MM for unidimensional and multidimensional scales, respectively. The average fit measures results indicate that, for both unidimensional and multidimensional scales and for all MI levels tested, including the additional ARS factor yields good to perfect fit according to all fit measures regardless of the other design factors. For the dichotomized results (Tables A6 - A7 in the Appendix), almost all the considered criteria resulted in a close-to-zero FPR when testing MI at all levels.

4 Conclusions

The simulation study assessed the effect of disregarding and including an additional ARS factor on MI testing when responses in one group are affected by an ARS. The results showed that not taking a strong ARS into account resulted in wrongly concluding there is measurement non-invariance for balanced and semi-balanced scales. In fact, for these scales, model fit heavily deteriorated for all MI levels, and thus one may conclude that the content factor(s) MM differs across groups while this is purely due to a strong agreeing tendency in one group. Note that this result is significant for empirical practice, where researchers that follow the standard CFA-based MI testing approach may conclude that configural invariance does not hold and try to modify the MM to be able to compare the groups or, in the most extreme case, refrain from further analyses. In the balanced and semi-balanced scales conditions, this issue was solved by including, for all groups, an additional ARS factor with all its loadings fixed to 1, which resulted in concluding that MI held at all levels. For unbalanced scales, however, disregarding ARS (i.e., not including ARS as an additional factor) did not affect the MI testing results, which indicated that MI held at all levels. This latter result partially overlaps with previous research, showing that ARS gets absorbed by the content factor(s) in unbalanced scales (Ferrando & Lorenzo-Seva, 2010; D'Urso et al., 2022). Thus, for unbalanced scales, wrongly concluding that MI holds across groups, even when individual responses in one group are strongly affected by an ARS, is a realistic scenario that one should take into account. Including an additional factor to capture ARS is not a solution for unbalanced scales since this often led to

model non-convergence, especially when testing for configural invariance, likely due to the indistinguishability of the ARS factor from the content factor(s).

5 Discussion

In psychological science, self-report scales are widely used to compare targeted latent constructs (e.g., depression) across groups. To draw valid and unbiased conclusions concerning latent construct differences, one must ensure that the self-report scales used to measure these constructs function equivalently across groups. The latter is often assessed through measurement invariance (MI) testing, which evaluates the tenability of the hypothesis of measurement model (MM) equivalence. For scales composed of ordinal items, MI is often tested through multiple group categorical confirmatory factor analysis (MG-CCFA), which allows evaluating MM parameters' equivalence across groups in a step-wise fashion. In addition to the scale itself being inequivalent, non-invariances may emerge when disregarding the influence of an agreeing response style (ARS), which represents a tendency to agree with items regardless of their content (Paulhus, 1991). Though it is known that certain groups may be particularly prone to ARS (i.e., see Van Vaerenbergh & Thomas, 2013 for a review), and that such response tendency can bias MM parameters in single group studies (Ferrando & Lorenzo-Seva, 2010), this is the first paper thoroughly evaluating the effects of ARS on MI testing. Determining if disregarding ARS can appear as measurement non-invariance may help to ascertain how to correct for bias in latent construct differences due to differential response tendencies across groups. In fact, it is superfluous to look for scale-specific causes of non-invariance if these are entirely due to ARS. Instead, including an extra factor to model the ARS corrects for the bias. In this paper, we conducted a simulation study to evaluate in what conditions and to what extent an ARS affecting the individual responses in one of the groups is detected as measurement non-invariance, both when disregarding ARS and when including it as an additional factor in the measurement model (MM) with all its loadings fixed to 1.

One of the more significant findings from this study is that ignoring a large ARS resulted in measurement non-invariance at all levels for balanced and semi-balanced scales,

which was solved by including an additional factor capturing ARS. Therefore, when using (semi-) balanced scales, researchers should bear in mind that configural non-invariance may occur by disregarding a large ARS, and that including an additional ARS factor in the MM for all groups is an effective way to correct for this. In this way, researchers can ascertain that there is no need to look for or remedy inequivalences that pertain to the scale. For unbalanced scales, disregarding an ARS did not affect MI testing results, and including an additional ARS factor was not advantageous since it often led to model non-convergence. This is likely due to the fact that, for unbalanced scales, the intended-to-be-measured (i.e., content) factors cannot be easily distinguished from the ARS factor. Nevertheless, for these scales, one should not conclude that ignoring an ARS is harmless because content factor scores and/or correlations may be biased for groups with a large ARS (Savalei & Falk, 2014; de la Fuente & Abad, 2020; D'Urso et al., 2022), and thus affect substantive conclusions. Sadly, in practice, the bias due to neglecting an ARS (e.g., in factor covariances and factor scores) may not be detected when testing for MI using unbalanced scales, and the correction proposed by Billiet and McClendon (2000) is not a solution. Therefore, in settings where ARS could be present, using unbalanced scales is inherently problematic as they do not allow one to correct for this response tendency.

Taken all together, these results indicate that ARS is a serious threat to MI testing results and that (semi-)balanced scales should be preferred when suspecting that an ARS may be at play for specific groups. Balanced or semi-balanced scales may not always be constructed, however. For instance, negatively worded items often require higher reading levels or intellectual capacity that cannot be assumed for certain (e.g., clinical) populations (Chyung, Barkin, & Shamsy, 2018). In these cases, one may consider using specific "marker" items or scales tailored to measure ARS, but further research is needed to evaluate the feasibility of this approach in the context of multidimensional scales and multiple group models (Ferrando, Morales-Vives, & Lorenzo-Seva, 2016). Alternatively, model non-convergence may be solved by bounded estimation (De Jonckere & Rosseel, 2022), where data-driven upper and lower bounds for model parameters can be specified prior to estimating the model (e.g., setting the lower bound of the ARS factor variance

to a non-negative number), is a promising solution that has not yet been evaluated for MG-CCFA.

Our simulation study is subject to a few limitations that are worth noting. First, ARS was the only considered source of bias that, when disregarded, yielded us to conclude that there is non-invariance. However, in practice, it is reasonable to expect that other, scale-specific sources of non-invariance, such as differential item interpretation, may affect individual responses in some (or all) groups. In the future, it would be interesting to extend the current simulation study to evaluate whether non-invariance due to disregarding an ARS may be disentangled from other non-invariances such as specific factor loading differences for the content factors. Second, we assumed that ARS was equally affecting all items (i.e., loadings on the ARS factor were equal across items). This assumption may not always be realistic in practice, as certain items may be more affected by ARS than others (Ferrando, Lorenzo-Seva, & Chico, 2003). Savalei and Falk (2014) showed that the approach of Billiet and McClendon (2000) is robust even when loadings on the ARS factor differ across items. However, this CFA-based approach can still be limiting (i) with items that load on more than one factor at the same time (i.e., cross-loadings) and (ii) when researchers are interested in assessing (between-group differences in) the ARS factor loadings. To overcome these limitations, one may consider using multiple group exploratory factor analysis (MG-EFA; Jöreskog, 1970). MG-EFA does not impose an assumed structure on the factor loadings, and thus can account for cross-loadings. Furthermore, in MG-EFA, one does not need to assume that the influence of ARS is equal across items (i.e., the loadings on the ARS factor do not need to be constrained to be 1). In fact, by using a (semi-)specified target rotation, one may take the additional ARS factor into account and estimate its loadings by (fully or) partially specifying (part of) the rotation target according to a priori expectations on the MM and leave the ARS factor loadings unspecified (D'Urso et al., 2022). Third, we considered simulation scenarios with only two groups. MI testing has become increasingly relevant for cross-cultural and cross-national research, where large data sets with many groups are the norm (Rutkowski & Svetina, 2017). Hence, future research should evaluate the extent to which disregard-

ing ARS or not affects MI testing when this agreeing bias influences individual responses only for a subset of groups or when it gradually differs across all groups. Finally, we followed a scale-based MI testing framework, but alternative approaches, such as item-based analyses (e.g., multiple group item response theory; D'Urso et al., 2021), may also be of interest, especially when the bias caused by an ARS affects some items more than others or when trying to distinguish ARS from specific differences in the content factor loadings.

Nevertheless, the present study is the only thorough investigation of the effect of ARS on MI testing. We showed that correcting for agreeing bias when testing for MI allows to determine that the scale is invariant otherwise. We expect this outcome to be tremendously valuable in empirical practice as this avoids unnecessary worries about and investigations of scale non-equivalence (e.g., looking for non-invariant items).

6 References

Aichholzer, J., et al. (2015). Controlling acquiescence bias in measurement invariance tests. *Psihologija*, *48*(4), 409–429.

Arias, V. B., Garrido, L., Jenaro, C., Martínez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods*, *52*(6), 2489–2505.

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed rasch modelling of responses to neo-ffi items. *Personality and individual differences*, *40*(6), 1235–1245.

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly*, *48*(2), 491–509.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238.

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural equation modeling*, *7*(4), 608–628.

Borsboom, D. (2006). When does measurement invariance matter? *Medical care*, *44*(11), S176–S181.

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

Browne, M. W., Cudeck, R., et al. (1993). Alternative ways of assessing model fit. *Sage Focus Editions*, *154*, 136–136.

Chan, W., & Bentler, P. M. (1993). The covariance structure analysis of ipsative data. *Sociological Methods & Research*, *22*(2), 214–247.

Chang, Y.-W., Hsu, N.-J., & Tsai, R.-C. (2017). Unifying differential item functioning in factor analysis for categorical data under a discretization of a normal variant. *psychometrika*, *82*(2), 382–406.

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. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of cross-cultural psychology*, *31*(2), 187–212.

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

Chyung, S. Y., Barkin, J. R., & Shamsy, J. A. (2018). Evidence-based survey design: The use of negatively worded items in surveys. *Performance Improvement*, *57*(3), 16–25.

Danner, D., Aichholzer, J., & Rammstedt, B. (2015). Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *Journal of Research in Personality*, *57*, 119–130.

De Jonckere, J., & Rosseel, Y. (2022). Using bounded estimation to avoid nonconvergence in small sample structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(3), 412–427.

de la Fuente, J., & Abad, F. J. (2020). Comparing methods for modeling acquiescence in multidimensional partially balanced scales. *Psicothema*.

D'Urso, E. D., De Roover, K., Vermunt, J. K., & Tijmstra, J. (2021). Scale length does matter: recommendations for measurement invariance testing with categorical factor analysis and item response theory approaches. *Behavior Research Methods*, 1–32.

D'Urso, E. D., Tijmstra, J., Vermunt, J., & De Roover, K. (2022). Awareness is bliss: How acquiescence affects exploratory factor analysis. *Educational and Psychological Measurement*, 00131644221089857.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, *4*(3), 272.

Ferrando, P. J., Condon, L., & Chico, E. (2004). The convergent validity of acquiescence: An empirical study relating balanced scales and separate acquiescence scales. *Personality and individual differences*, *37*(7), 1331–1340.

Ferrando, P. J., & Lorenzo-Seva, U. (2010). Acquiescence as a source of bias and model and person misfit: A theoretical and empirical analysis. *British Journal of Mathematical and Statistical Psychology*, *63*(2), 427–448.

Ferrando, P. J., Lorenzo-Seva, U., & Chico, E. (2003). Unrestricted factor analytic procedures for assessing acquiescent responding in balanced, theoretically unidimensional personality scales. *Multivariate Behavioral Research*, *38*(3), 353–374.

Ferrando, P. J., Morales-Vives, F., & Lorenzo-Seva, U. (2016). Assessing and controlling acquiescent responding when acquiescence and content are related: A comprehensive factor-analytic approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 713–725.

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling*, *13*(3), 378–402.

French, B. F., & Finch, W. H. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(1), 96–113.

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in psychology*, *5*, 980.

Jeong, S., & Lee, Y. (2019). Consequences of not conducting measurement invariance tests in cross-cultural studies: A review of current research practices and recommendations. *Advances in Developing Human Resources*, *21*(4), 466–483.

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-cultural psychology*, *36*(2), 264–277.

Jones, R. N., & Gallo, J. J. (2002). Education and sex differences in the mini-mental state examination: effects of differential item functioning. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *57*(6), P548–P558.

Jöreskog, K. (1970). Simultaneous factor analysis in several populations. *ETS Research Bulletin Series*, *1970*(2), i–31.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., Rosseel, Y., Miller, P., Quick, C., . . . others (2016). *Package 'semtools'.*

Liu, M., Harbaugh, A. G., Harring, J. R., & Hancock, G. R. (2017). The effect of extreme response and non-extreme response styles on testing measurement invariance. *Frontiers in psychology*, *8*, 726.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological methods*, *4*(1), 84.

Marin, G., Gamba, R. J., & Marin, B. V. (1992). Extreme response style and acquiescence among hispanics: The role of acculturation and education. *Journal of cross-cultural psychology*, *23*(4), 498–509.

McNeish, D., & Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*.

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and individual differences*, *44*(7), 1539–1550.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical care*, S69–S77.

Muthén, B., & Muthén, B. O. (2009). *Statistical analysis with latent variables* (Vol. 123) (No. 6). Wiley New York.

Paulhus, D. L. (1991). Measurement and control of response bias.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental review*, *41*, 71–90.

Rios, J. A. (2021). Is differential noneffortful responding associated with type i error in measurement invariance testing? *Educational and Psychological Measurement*, *81*(5), 957–979.

Rosseel, Y. (2012). lavaan: An r package for structural equation modeling. *Journal of statistical software*, *48*, 1–36.

Rutkowski, L., & Svetina, D. (2017). Measurement invariance in international surveys:

Categorical indicators and fit measure performance. *Applied Measurement in Education*, *30*(1), 39–51.

Savalei, V., & Falk, C. F. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate behavioral research*, *49*(5), 407–424.

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). 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*, *27*(1), 111–130.

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217.

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological methods*, *15*(1), 96.

Welkenhuysen-Gybels, J., Billiet, J., & Cambré, B. (2003). Adjustment for acquiescence in the assessment of the construct equivalence of likert-type score items. *Journal of Cross-Cultural Psychology*, *34*(6), 702–722.

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.

Table 1

Identification and MI constraints Wu & Estabrook (2016) for MI testing with MG-CCFA.

	MI constraints for MG-CCFA with two groups									
	Identification constraints								MI constraints	Model comparison
	LV				LRV					
	κ		ϕ		ν		σ^2			
	G1	G2	G1	G2	G1	G2	G1	G2		
1 - Configural	0	0	1	1	0	0	1	1	$\boldsymbol{\mu}^{(g)} = \boldsymbol{\nu}^{(g)} + \boldsymbol{\Lambda}^{(g)}\boldsymbol{\kappa}^{(g)},$ $\boldsymbol{\Sigma}^{(g)} = \boldsymbol{\Lambda}^{(g)}\boldsymbol{\Phi}^{(g)}\boldsymbol{\Lambda}^{(g)} + \boldsymbol{\Psi}^{(g)},$ $\boldsymbol{T}^{(g)} = \boldsymbol{T}^{(g)}$	
2 - Thresholds	0	0	1	1	0	free	1	free	$\boldsymbol{\mu}^{(g)} = \boldsymbol{\nu}^{(g)} + \boldsymbol{\Lambda}^{(g)}\boldsymbol{\kappa}^{(g)},$ $\boldsymbol{\Sigma}^{(g)} = \boldsymbol{\Lambda}^{(g)}\boldsymbol{\Phi}^{(g)}\boldsymbol{\Lambda}^{(g)} + \boldsymbol{\Psi}^{(g)},$ $\boldsymbol{T}^{(g)} = \boldsymbol{T}$	**2 vs. 1**
3 - Loadings	0	0	1	free	0	free	1	free	$\boldsymbol{\mu}^{(g)} = \boldsymbol{\nu}^{(g)} + \boldsymbol{\Lambda}\boldsymbol{\kappa}^{(g)},$ $\boldsymbol{\Sigma}^{(g)} = \boldsymbol{\Lambda}\boldsymbol{\Phi}^{(g)}\boldsymbol{\Lambda} + \boldsymbol{\Psi}^{(g)},$ $\boldsymbol{T}^{(g)} = \boldsymbol{T}$	**3 vs. 2**
4 - Intercepts	0	free	1	free	0	0	1	free	$\boldsymbol{\mu}^{(g)} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\kappa}^{(g)},$ $\boldsymbol{\Sigma}^{(g)} = \boldsymbol{\Lambda}\boldsymbol{\Phi}^{(g)}\boldsymbol{\Lambda} + \boldsymbol{\Psi}^{(g)},$ $\boldsymbol{T}^{(g)} = \boldsymbol{T}$	**4 vs. 3**

Note. LV = Latent variable; LRV = Latent response variable.

Table 2

Population values for the simulation study.

	Content factor(s)			ARS factor	Thresholds			
	One factor	Two factors			τ			
	λ_1	λ_1	λ_2	λ_{ARS}	τ_1	τ_2	τ_3	τ_4
X_1	0.6	0.6	0	0.175/0.350	-0.250	-0.750	-1.250	-1.750
X_2	0.6	0	0.6	0.175/0.350	-0.068	-0.568	-1.068	-1.568
X_3	0.6	(-)0.6	0	0.175/0.350	0.114	-0.386	-0.886	-1.386
X_4	0.6	0	(-)0.6	0.175/0.350	0.295	-0.205	-0.705	-1.205
X_5	0.6	0.6	0	0.175/0.350	0.477	-0.023	-0.523	-1.023
X_6	0.6	0	0.6	0.175/0.350	0.659	0.159	-0.341	-0.841
X_7	(-)0.6	**(-)0.6**	0	0.175/0.350	0.841	0.341	-0.159	-0.659
X_8	(-)0.6	0	**(-)0.6**	0.175/0.350	1.023	0.523	0.023	-0.477
X_9	**(-)0.6**	0.6	0	0.175/0.350	1.205	0.705	0.205	-0.295
X_{10}	**(-)0.6**	0	0.6	0.175/0.350	1.386	0.886	0.386	-0.114
X_{11}	**(-)0.6**	**(-)0.6**	0	0.175/0.350	1.568	1.068	0.568	0.068
X_{12}	**(-)0.6**	0	**(-)0.6**	0.175/0.350	1.750	1.250	0.750	0.250

Note. Loadings in bold indicates those that were negative for semi-balanced scales. The original ARS factor loadings values were .3/.6 but here we display the rescaled ARS loadings obtained by multiplying the original ARS loadings values by the standard deviation of a right-censored normal distribution, which was used to generate the ARS factor scores.

Table 3

Average fit value for MI testing approaches when the ARS factor is not included in unidimensional scales in function of the simulated conditions

| | | | | Average fit value in MI testing for unidimensional scales without ARS factor | | | | | | | |
| | | | | Configural | | | | Thresholds | | | |
ARS	Scale	N	J	χ^2 (SD)	df	$RMSEA$ (SD)	CFI (SD)	χ^2_τ (SD)	df_τ	$RMSEA_\tau$ (SD)	CFI_τ (SD)
0.175	Bal	250	12	131.4 (16.0)	108	0.027 (0.012)	0.990 (0.007)	157.6 (16.9)	132	0.026 (0.010)	0.989 (0.007)
		250	24	587.0 (41.5)	504	0.025 (0.007)	0.988 (0.006)	636.6 (42.1)	552	0.024 (0.006)	0.988 (0.006)
		1000	12	202.7 (31.2)	108	0.029 (0.005)	0.990 (0.004)	234.3 (34.4)	132	0.027 (0.005)	0.989 (0.004)
		1000	24	861.3 (74.0)	504	0.026 (0.003)	0.987 (0.003)	920.8 (76.0)	552	0.026 (0.003)	0.986 (0.003)
	Semi	250	12	129.9 (15.8)	108	0.026 (0.011)	0.991 (0.006)	155.2 (17.4)	132	0.024 (0.011)	0.990 (0.007)
		250	24	581.0 (35.6)	504	0.024 (0.006)	0.989 (0.006)	631.0 (36.5)	552	0.023 (0.006)	0.988 (0.006)
		1000	12	186.2 (26.7)	108	0.027 (0.005)	0.992 (0.003)	216.5 (29.3)	132	0.025 (0.005)	0.991 (0.003)
		1000	24	852.5 (68.7)	504	0.026 (0.003)	0.987 (0.003)	911.6 (70.9)	552	0.025 (0.003)	0.987 (0.003)
	Unbal	250	12	109.7 (12.7)	108	0.010 (0.011)	0.998 (0.003)	133.6 (14.3)	132	0.010 (0.011)	0.998 (0.003)
		250	24	516.8 (22.6)	504	0.009 (0.007)	0.998 (0.002)	565.3 (23.4)	552	0.009 (0.007)	0.998 (0.002)
		1000	12	110.0 (14.8)	108	0.005 (0.006)	0.999 (0.001)	134.2 (16.1)	132	0.005 (0.006)	0.999 (0.001)
		1000	24	512.8 (28.1)	504	0.004 (0.004)	0.999 (0.001)	560.8 (29.7)	552	0.004 (0.004)	0.999 (0.001)
0.350	Bal	250	12	539.7 (80.1)	108	**0.126 (0.012)**	**0.895 (0.021)**	597.4 (87.2)	132	**0.118 (0.011)**	**0.886 (0.023)**
		250	24	1863.9 (243.8)	504	**0.104 (0.009)**	**0.875 (0.026)**	1952.9 (252.4)	552	**0.101 (0.009)**	**0.871 (0.027)**
		100	12	1927.0 (170.4)	108	**0.130 (0.006)**	**0.885 (0.013)**	2110.4 (186.8)	132	**0.122 (0.006)**	**0.875 (0.014)**
		100	24	6523.6 (564.5)	504	**0.109 (0.005)**	**0.861 (0.016)**	6806.0 (589.2)	552	**0.106 (0.005)**	**0.855 (0.016)**
	Semi	250	12	538.1 (107.0)	108	**0.126 (0.015)**	**0.898 (0.027)**	594.5 (115.8)	132	**0.118 (0.015)**	**0.890 (0.029)**
		250	24	1876.6 (232.0)	504	**0.104 (0.009)**	**0.877 (0.024)**	1966.3 (240.6)	552	**0.101 (0.009)**	**0.873 (0.024)**
		1000	12	1953.5 (211.9)	108	**0.131 (0.008)**	**0.885 (0.015)**	2136.3 (232.0)	132	**0.123 (0.007)**	**0.875 (0.016)**
		1000	24	6497.7 (473.9)	504	**0.109 (0.004)**	**0.860 (0.014)**	6779.1 (494.4)	552	**0.106 (0.004)**	**0.855 (0.015)**
	Unbal	250	12	113.4 (14.1)	108	0.012 (0.013)	0.999 (0.002)	137.3 (15.5)	132	0.012 (0.012)	0.999 (0.002)
		250	24	508.1 (20.0)	504	0.006 (0.007)	0.999 (0.001)	556.0 (20.5)	552	0.006 (0.006)	0.999 (0.001)
		1000	12	105.9 (13.0)	108	0.004 (0.005)	1.000 (0.000)	130.5 (14.5)	132	0.004 (0.005)	1.000 (0.000)
		1000	24	508.2 (24.4)	504	0.003 (0.004)	1.000 (0.000)	557.5 (26.2)	552	0.003 (0.004)	1.000 (0.000)

Note. ARS = ARS factor loadings; N = sample size within each group; Bal = Balanced scale; Semi = Semi-balanced scale; Unbal = Unbalanced scale; J = number of items. Values in bold indicate conditions where the average model fit results were below commonly accepted cut-off values for "good" fit.

Table 4

Average fit value for MI testing approaches when the ARS factor is not included in unidimensional scales in function of the simulated conditions

| | | | | Average fit value in MI testing for unidimensional scales without ARS factor | | | | | | | |
| | | | | Loadings | | | | Intercepts | | | |
ARS	Scale	N	J	χ^2_λ (SD)	df_λ	$RMSEA_\lambda$ (SD)	CFI_λ (SD)	χ^2_ν (SD)	df_ν	$RMSEA_\nu$ (SD)	CFI_ν (SD)
0.175	Bal	250	12	167.4 (18.3)	143	0.024 (0.011)	0.989 (0.007)	220.8 (32.2)	154	0.040 (0.011)	0.972 (0.014)
			24	656.0 (42.6)	575	0.023 (0.007)	0.988 (0.006)	749.4 (62.0)	598	0.031 (0.007)	0.978 (0.009)
		1000	12	237.7 (32.8)	143	0.025 (0.005)	0.990 (0.004)	449.4 (56.2)	154	0.044 (0.004)	0.969 (0.006)
			24	925.0 (75.0)	575	0.025 (0.003)	0.987 (0.003)	1305.4 (108.0)	598	0.034 (0.003)	0.974 (0.004)
	Semi	250	12	166.1 (18.5)	143	0.023 (0.011)	0.990 (0.007)	202.0 (27.1)	154	0.034 (0.011)	0.980 (0.011)
			24	651.2 (37.0)	575	0.022 (0.006)	0.989 (0.006)	735.3 (45.3)	598	0.030 (0.005)	0.980 (0.007)
		1000	12	221.6 (30.5)	143	0.023 (0.005)	0.992 (0.003)	374.8 (51.5)	154	0.038 (0.004)	0.977 (0.005)
			24	915.7 (71.9)	575	0.024 (0.003)	0.987 (0.003)	1294.8 (100.9)	598	0.034 (0.002)	0.974 (0.004)
	Unbal	250	12	143.3 (15.3)	143	0.009 (0.010)	0.998 (0.004)	154.8 (16.7)	154	0.009 (0.010)	0.997 (0.004)
			24	587.7 (25.0)	575	0.008 (0.008)	0.998 (0.003)	608.8 (24.9)	598	0.007 (0.007)	0.998 (0.003)
		1000	12	144.0 (18.3)	143	0.005 (0.006)	0.999 (0.001)	154.0 (19.0)	154	0.004 (0.005)	0.999 (0.001)
			24	579.4 (30.4)	575	0.004 (0.004)	0.999 (0.001)	597.5 (33)	598	0.003 (0.003)	1.000 (0.001)
0.350	Bal	250	12	585.4 (85.4)	143	**0.111 (0.011)**	**0.892 (0.022)**	711.4 (103.7)	154	**0.120 (0.011)**	**0.864 (0.026)**
			24	1939.6 (249.7)	575	**0.097 (0.009)**	**0.875 (0.027)**	2112.6 (264.2)	598	**0.100 (0.009)**	**0.861 (0.029)**
		100	12	2032.2 (184.4)	143	**0.115 (0.006)**	**0.881 (0.014)**	2600.3 (221.8)	154	**0.126 (0.006)**	**0.846 (0.016)**
			24	6729.6 (588.4)	575	**0.103 (0.005)**	**0.858 (0.016)**	7548.7 (594.0)	598	**0.108 (0.005)**	**0.839 (0.017)**
	Semi	250	12	581.3 (111.9)	143	**0.110 (0.014)**	**0.896 (0.028)**	664.2 (116.9)	154	**0.115 (0.013)**	**0.879 (0.029)**
			24	1955.3 (240.5)	575	**0.098 (0.009)**	**0.876 (0.024)**	2129.0 (250.8)	598	**0.101 (0.008)**	**0.863 (0.025)**
		1000	12	2046.6 (225.3)	143	**0.115 (0.007)**	**0.881 (0.016)**	2413.6 (248.6)	154	**0.121 (0.007)**	**0.859 (0.017)**
			24	6702.4 (492.9)	575	**0.103 (0.004)**	**0.857 (0.014)**	7517.4 (509.4)	598	**0.108 (0.004)**	**0.839 (0.015)**
	Unbal	250	12	149.3 (17.6)	143	0.012 (0.012)	0.998 (0.002)	159.0 (18.7)	154	0.011 (0.012)	0.998 (0.002)
			24	581.7 (21.9)	575	0.007 (0.007)	0.999 (0.001)	603.9 (22.4)	598	0.006 (0.006)	0.999 (0.001)
		1000	12	145.5 (18.1)	143	0.005 (0.006)	1.000 (0.001)	156.3 (18.9)	154	0.005 (0.006)	1.000 (0.001)
			24	589.1 (30.4)	575	0.004 (0.004)	1.000 (0.000)	611.7 (34.7)	598	0.004 (0.004)	1.000 (0.000)

Note. ARS = ARS factor loadings; N = sample size within each group; Bal = Balanced scale; Semi = Semi-balanced scale; Unbal = Unbalanced scale; J = number of items. Values in bold indicate conditions where the average model fit results were below commonly accepted cut-off values for "good" fit.

Table 5

Average fit value for MI testing approaches when the ARS factor is not included in multidimensional scales in function of the simulated conditions

| | | | | Average fit value in MI testing for multidimensional scales with ARS factor | | | | | | | |
| | | | | Configural | | | | Thresholds | | | |
ARS	Scale	N	J	χ^2 (SD)	df	$RMSEA$ (SD)	CFI (SD)	χ^2_ν (SD)	df_ν	$RMSEA_\nu$ (SD)	CFI_ν (SD)
0.175	Bal	250	12	123.4 (17.1)	106	0.023 (0.013)	0.986 (0.011)	148.8 (19.0)	130	0.021 (0.013)	0.985 (0.013)
			24	558.2 (32.2)	502	0.020 (0.007)	0.986 (0.008)	608.2 (33.4)	550	0.019 (0.007)	0.985 (0.008)
		1000	12	163.4 (22.1)	106	0.023 (0.004)	0.989 (0.004)	192.7 (24.4)	130	0.022 (0.004)	0.988 (0.005)
			24	705.0 (58.7)	502	0.020 (0.003)	0.988 (0.004)	761.1 (60.9)	550	0.019 (0.003)	0.987 (0.004)
	Semi	250	12	122.4 (14.4)	106	0.022 (0.012)	0.987 (0.010)	147.9 (15.9)	130	0.021 (0.012)	0.986 (0.011)
			24	553.3 (32.6)	502	0.019 (0.007)	0.987 (0.008)	602.9 (34.0)	550	0.018 (0.007)	0.987 (0.008)
		1000	12	151.2 (22.1)	106	0.020 (0.006)	0.991 (0.004)	179.7 (24.3)	130	0.019 (0.006)	0.991 (0.005)
			24	669.5 (35.6)	502	0.018 (0.002)	0.990 (0.002)	723.9 (37.6)	550	0.018 (0.002)	0.990 (0.002)
	Unbal	250	12	109.8 (16.2)	106	0.012 (0.013)	0.994 (0.008)	134.2 (17.8)	130	0.011 (0.012)	0.994 (0.009)
			24	520.7 (28.2)	502	0.011 (0.008)	0.995 (0.005)	569.3 (29.1)	550	0.010 (0.008)	0.995 (0.006)
		1000	12	106.3 (12.9)	106	0.004 (0.005)	0.999 (0.001)	130.7 (14.5)	130	0.004 (0.005)	0.999 (0.002)
			24	507.5 (32.0)	502	0.004 (0.004)	0.999 (0.001)	556.0 (33.6)	550	0.004 (0.004)	0.999 (0.001)
0.350	Bal	250	12	392.1 (65.8)	106	**0.103 (0.012)**	**0.866 (0.031)**	441.7 (71.9)	130	**0.097 (0.011)**	**0.854 (0.034)**
			24	1415.5 (175.1)	502	**0.085 (0.008)**	**0.862 (0.026)**	1494.6 (181.5)	550	**0.083 (0.008)**	**0.857 (0.027)**
		100	12	1315.6 (148.4)	106	**0.107 (0.007)**	**0.857 (0.019)**	1456.5 (162.8)	130	**0.101 (0.006)**	**0.843 (0.021)**
			24	4378.3 (357.6)	502	**0.088 (0.004)**	**0.853 (0.014)**	4586.4 (372.3)	550	**0.086 (0.004)**	**0.847 (0.014)**
	Semi	250	12	377.6 (66.3)	106	**0.101 (0.012)**	**0.879 (0.030)**	425.2 (71.9)	130	**0.095 (0.011)**	**0.868 (0.033)**
			24	1382.4 (176.7)	502	**0.083 (0.009)**	**0.872 (0.028)**	1459.6 (182.4)	550	**0.081 (0.008)**	**0.868 (0.028)**
		1000	12	1188.6 (111.9)	106	**0.101 (0.005)**	**0.877 (0.013)**	1316.3 (122.7)	130	**0.095 (0.005)**	**0.866 (0.014)**
			24	4192.1 (371.8)	502	**0.086 (0.004)**	**0.864 (0.014)**	4390.7 (387.3)	550	**0.083 (0.004)**	**0.858 (0.015)**
	Unbal	250	12	110.3 (15.4)	106	0.012 (0.013)	0.998 (0.003)	134.5 (16.7)	130	0.012 (0.012)	0.997 (0.004)
			24	515.6 (29.1)	502	0.009 (0.008)	0.998 (0.002)	564.4 (30.0)	550	0.009 (0.008)	0.998 (0.002)
		1000	12	114.8 (14.3)	106	0.008 (0.006)	0.999 (0.001)	139.9 (16.4)	130	0.007 (0.006)	0.999 (0.001)
			24	531.9 (37.8)	502	0.007 (0.005)	0.999 (0.001)	581.7 (39.9)	550	0.007 (0.005)	0.999 (0.001)

Note. ARS = ARS factor loadings; N = sample size within each group; Bal = Balanced scale; Semi = Semi-balanced scale; Unbal = Unbalanced scale; J = number of items. Values in bold indicate conditions where the average model fit results were below commonly accepted cut-off values for "good" fit

Table 6

Average fit value for MI testing approaches when the ARS factor is not included in multidimensional scales in function of the simulated conditions

				Average fit value in MI testing for multidimensional scales without ARS factor							
				Loadings				Intercepts			
ARS	**Scale**	**N**	**J**	χ^2_λ **(SD)**	df_λ	$RMSEA_\lambda$ **(SD)**	CFI_λ **(SD)**	χ^2_ν **(SD)**	df_ν	$RMSEA_\nu$ **(SD)**	CFI_ν **(SD)**
0.175	Bal	250	12	159.3 (18.3)	140	0.021 (0.012)	0.984 (0.013)	203.4 (25.6)	150	0.036 (0.011)	0.959 (0.020)
			24	628.3 (33.4)	572	0.019 (0.007)	0.986 (0.008)	697.4 (36.6)	594	0.026 (0.005)	0.974 (0.009)
		1000	12	201.4 (25.8)	140	0.020 (0.004)	0.988 (0.005)	353.2 (48.4)	150	0.037 (0.004)	0.961 (0.009)
			24	781.2 (60.6)	572	0.019 (0.003)	0.988 (0.004)	1054.7 (80.7)	594	0.028 (0.003)	0.973 (0.005)
	Semi	250	12	158.3 (16.1)	140	0.020 (0.012)	0.985 (0.011)	195.1 (22.7)	150	0.033 (0.010)	0.966 (0.017)
			24	624.8 (35.8)	572	0.018 (0.007)	0.987 (0.009)	686.9 (41.2)	594	0.024 (0.006)	0.977 (0.011)
		1000	12	191.7 (24.3)	140	0.019 (0.005)	0.990 (0.004)	323.8 (39.0)	150	0.034 (0.004)	0.967 (0.007)
			24	747.1 (37.5)	572	0.017 (0.002)	0.990 (0.002)	971.3 (60.7)	594	0.025 (0.002)	0.977 (0.004)
	Unbal	250	12	144.3 (18.1)	140	0.011 (0.012)	0.993 (0.009)	154.2 (19.1)	150	0.011 (0.012)	0.993 (0.009)
			24	592.1 (31.1)	572	0.010 (0.008)	0.994 (0.006)	612.8 (32.0)	594	0.010 (0.008)	0.994 (0.006)
		1000	12	141.1 (16.1)	140	0.005 (0.005)	0.999 (0.002)	150.9 (17.4)	150	0.004 (0.005)	0.999 (0.002)
			24	576.3 (34.6)	572	0.004 (0.004)	0.999 (0.001)	596.3 (34.2)	594	0.003 (0.004)	0.999 (0.001)
0.350	Bal	250	12	447.9 (71.7)	140	**0.093 (0.011)**	**0.855 (0.034)**	574.9 (81.9)	150	**0.106 (0.010)**	**0.800 (0.039)**
			24	1501.6 (179.5)	572	**0.080 (0.008)**	**0.860 (0.027)**	1686.1 (200.0)	594	**0.086 (0.008)**	**0.835 (0.030)**
		100	12	1461.1 (162.6)	140	**0.097 (0.006)**	**0.844 (0.021)**	1985.9 (204.7)	150	**0.111 (0.006)**	**0.783 (0.027)**
			24	4584.4 (370.9)	572	**0.084 (0.004)**	**0.848 (0.014)**	5392.3 (414.1)	594	**0.090 (0.004)**	**0.818 (0.016)**
	Semi	250	12	437.7 (71.3)	140	**0.092 (0.011)**	**0.867 (0.032)**	538.1 (87.1)	150	**0.101 (0.011)**	**0.827 (0.040)**
			24	1468.6 (180.1)	572	**0.079 (0.008)**	**0.870 (0.028)**	1616.4 (193.9)	594	**0.083 (0.008)**	**0.851 (0.030)**
		1000	12	1338.0 (122.4)	140	**0.092 (0.005)**	**0.864 (0.014)**	1771.5 (145.7)	150	**0.104 (0.005)**	**0.816 (0.017)**
			24	4402.5 (387.1)	572	**0.082 (0.004)**	**0.858 (0.014)**	5084.9 (425.6)	594	**0.087 (0.004)**	**0.834 (0.016)**
	Unbal	250	12	143.9 (18.2)	140	0.011 (0.012)	0.997 (0.004)	154.0 (18.4)	150	0.011 (0.011)	0.997 (0.004)
			24	587.7 (29.8)	572	0.009 (0.008)	0.998 (0.003)	608.3 (29.9)	594	0.009 (0.008)	0.998 (0.003)
		1000	12	152.6 (17.8)	140	0.008 (0.006)	0.999 (0.001)	161.4 (18.9)	150	0.007 (0.006)	0.999 (0.001)
			24	608.4 (40.7)	572	0.007 (0.005)	0.999 (0.001)	630.0 (40.6)	594	0.007 (0.005)	0.999 (0.001)

Note. ARS = ARS factor loadings; N = sample size within each group; Bal = Balanced scale; Semi = Semi-balanced scale; Unbal = Unbalanced scale; J = number of items. Values in bold indicate conditions where the average model fit results were below commonly accepted cut-off values for "good" fit

Table 7

Convergence Rate in the ARS scenario for MI testing in function of the simulated conditions when the ARS factor is included.

| | | | | Convergence Rate MI testing with ARS factor | | | | | | | |
| | | | | 1 Content Factor | | | | 2 Content Factors | | | |
ARS	Scale	N	J	Configural	Thresholds	Loadings	Intercepts	Configural	Thresholds	Loadings	Intercepts
		250	12	1	1	1	1	1	1	1	1
			24	1	1	1	1	1	1	1	1
	Bal	1000	12	1	1	1	1	1	1	1	1
			24	1	1	1	1	1	1	1	1
		250	12	1	1	1	1	1	1	1	1
			24	1	1	1	1	1	1	1	1
0.175	Semi	1000	12	1	1	1	1	1	1	1	1
			24	1	1	1	1	1	1	1	1
		250	12	0.410	0.690	0.990	0.990	0.430	0.760	0.970	0.960
			24	0.690	0.780	1	0.980	0.350	0.620	0.890	0.940
	Unbal	1000	12	0.480	0.690	0.970	0.990	0.430	0.790	1	0.990
			24	0.600	0.630	1	1	0.400	0.870	0.930	0.950
		250	12	1	1	1	1	1	1	1	1
			24	1	1	1	1	1	1	1	1
	Bal	100	12	1	1	1	1	1	1	1	1
			24	1	1	1	1	1	1	1	1
		250	12	1	1	1	1	1	1	1	1
			24	1	1	1	1	1	1	1	1
0.350	Semi	1000	12	1	1	1	1	1	1	1	1
			24	1	1	1	1	1	1	1	1
		250	12	0.410	0.820	0.990	0.990	0.290	0.730	0.860	0.920
			24	0.640	0.870	0.990	1	0.170	0.560	0.860	0.830
	Unbal	1000	12	0.420	0.790	1	0.990	0.250	0.780	0.920	0.930
			24	0.550	0.720	0.970	1	0.120	0.750	0.850	0.890

Note. ARS = ARS factor loadings; N = sample size within each group; Bal = Balanced scale; Semi = Semi-balanced scale; Unbal = Unbalanced scale; J = number of items.

Table 8

Average fit value for MI testing approaches when the ARS factor is included in unidimensional scales in function of the simulated conditions

| | | | | Average fit value in MI testing for unidimensional scales with ARS factor | | | | | | | |
| | | | | Configural | | | | Thresholds | | | |
ARS	Scale	N	J	χ^2 (SD)	df	$RMSEA$ (SD)	CFI (SD)	χ^2_ν (SD)	df_ν	$RMSEA_\nu$ (SD)	CFI_ν (SD)
0.175	Bal	250	12	109.9 (13.4)	106	0.011 (0.012)	0.997 (0.004)	134.7 (14.1)	130	0.011 (0.011)	0.997 (0.005)
			24	515.0 (22.1)	502	0.009 (0.007)	0.997 (0.002)	562.7 (22.4)	550	0.009 (0.007)	0.997 (0.002)
		1000	12	106.2 (13.4)	106	0.004 (0.006)	0.999 (0.001)	129.8 (14.8)	130	0.004 (0.005)	0.999 (0.001)
			24	507.1 (25.3)	502	0.004 (0.004)	1.000 (0.001)	555.4 (26.5)	550	0.004 (0.003)	1.000 (0.001)
	Semi	250	12	111.3 (12.4)	106	0.013 (0.012)	0.997 (0.004)	135.3 (13.6)	130	0.012 (0.011)	0.997 (0.004)
			24	514.4 (23.5)	502	0.009 (0.008)	0.998 (0.003)	563.0 (24.7)	550	0.009 (0.007)	0.997 (0.003)
		1000	12	106.6 (15.1)	106	0.005 (0.006)	0.999 (0.001)	130.4 (16.7)	130	0.005 (0.006)	0.999 (0.001)
			24	510.7 (24.6)	502	0.004 (0.004)	0.999 (0.001)	559.1 (25.0)	550	0.004 (0.003)	0.999 (0.001)
	Unbal	250	12	99.1 (12.3)	106	0.005 (0.008)	0.999 (0.002)	118.8 (13.3)	130	0.003 (0.007)	0.999 (0.002)
			24	510.5 (24.4)	502	0.008 (0.008)	0.998 (0.002)	548.8 (23.4)	550	0.005 (0.006)	0.999 (0.002)
		1000	12	98.8 (15.5)	106	0.003 (0.005)	1.000 (0.001)	121.1 (17.4)	130	0.003 (0.005)	1.000 (0.001)
			24	502.9 (29.2)	502	0.003 (0.004)	1.000 (0.001)	542.1 (31.9)	550	0.002 (0.003)	1.000 (0.000)
0.350	Bal	250	12	116.1 (14.8)	106	0.017 (0.013)	0.997 (0.003)	140.2 (16.4)	130	0.015 (0.012)	0.997 (0.003)
			24	568.6 (34.6)	502	0.022 (0.007)	0.994 (0.003)	584.2 (27.8)	550	0.014 (0.007)	0.997 (0.002)
		100	12	124.5 (18.3)	106	0.011 (0.007)	0.999 (0.001)	147.7 (21.3)	130	0.010 (0.007)	0.999 (0.001)
			24	778.6 (81.5)	502	0.023 (0.004)	0.994 (0.002)	696.4 (56.3)	550	0.016 (0.003)	0.997 (0.001)
	Semi	250	12	115.3 (15.1)	106	0.016 (0.013)	0.997 (0.003)	137.6 (16.4)	130	0.014 (0.012)	0.997 (0.003)
			24	572.8 (37.4)	502	0.023 (0.007)	0.994 (0.004)	584.8 (29.9)	550	0.015 (0.007)	0.997 (0.003)
		1000	12	132.7 (23.6)	106	0.014 (0.008)	0.998 (0.001)	154.1 (23.8)	130	0.012 (0.007)	0.998 (0.001)
			24	776.7 (66.4)	502	0.023 (0.003)	0.994 (0.002)	695.1 (49.7)	550	0.016 (0.003)	0.997 (0.001)
	Unbal	250	12	99.5 (13.6)	106	0.004 (0.009)	1.000 (0.001)	123.7 (15.8)	130	0.005 (0.010)	0.999 (0.001)
			24	501.1 (23.4)	502	0.005 (0.006)	0.999 (0.001)	540.8 (21.1)	550	0.003 (0.005)	1.000 (0.001)
		1000	12	95.8 (14.5)	106	0.002 (0.004)	1.000 (0.000)	117.8 (15.1)	130	0.002 (0.003)	1.000 (0.000)
			24	499.9 (27.6)	502	0.003 (0.003)	1.000 (0.000)	543.3 (25.0)	550	0.002 (0.003)	1.000 (0.000)

Note. ARS = ARS factor loadings ; N = sample size within each group; Bal = Balanced scale; Semi = Semi-balanced scale; Unbal = Unbalanced scale; J = number of items.

Table 9

Average fit value for MI testing approaches when the ARS factor is included in unidimensional scales in function of the simulated conditions

				Average fit value in MI testing for multidimensional scales without ARS factor							
				Loadings				Intercepts			
ARS	**Scale**	**N**	**J**	χ^2_λ **(SD)**	df_λ	$RMSEA_\lambda$ **(SD)**	CFI_λ **(SD)**	χ^2_ν **(SD)**	df_ν	$RMSEA_\nu$ **(SD)**	CFI_ν **(SD)**
0.175	Bal	250	12	146.2 (16.0)	141	0.012 (0.011)	0.996 (0.005)	155.8 (18.6)	151	0.011 (0.012)	0.996 (0.006)
			24	586.0 (23.9)	573	0.009 (0.007)	0.997 (0.003)	607.4 (26.4)	595	0.008 (0.007)	0.997 (0.003)
		1000	12	142.1 (17.4)	141	0.005 (0.006)	0.999 (0.001)	152.1 (20.4)	151	0.004 (0.006)	0.999 (0.002)
			24	578.9 (30.7)	573	0.004 (0.004)	0.999 (0.001)	597.3 (34.0)	595	0.003 (0.004)	0.999 (0.001)
	Semi	250	12	147.2 (14.6)	141	0.012 (0.011)	0.996 (0.004)	156.2 (16.2)	151	0.012 (0.011)	0.996 (0.005)
			24	586.5 (26.3)	573	0.009 (0.007)	0.997 (0.003)	606.5 (28.0)	595	0.008 (0.007)	0.997 (0.003)
		1000	12	141.2 (20.0)	141	0.005 (0.006)	0.999 (0.001)	153.2 (21.5)	151	0.005 (0.006)	0.999 (0.002)
			24	581.7 (30.8)	573	0.004 (0.004)	0.999 (0.001)	603.5 (36.8)	595	0.004 (0.004)	0.999 (0.001)
	Unbal	250	12	134.9 (14.0)	141	0.005 (0.008)	0.999 (0.003)	141.7 (13.4)	151	0.003 (0.007)	0.999 (0.002)
			24	580.2 (24.5)	573	0.007 (0.007)	0.998 (0.002)	594.7 (23.7)	595	0.005 (0.006)	0.999 (0.002)
		1000	12	133.1 (17.6)	141	0.003 (0.005)	1.000 (0.001)	140.0 (16.8)	151	0.002 (0.004)	1.000 (0.001)
			24	573.7 (30.7)	573	0.003 (0.004)	1.000 (0.001)	583.2 (31.0)	595	0.002 (0.003)	1.000 (0.000)
0.350	Bal	250	12	152.6 (18.2)	141	0.016 (0.012)	0.997 (0.004)	162.0 (18.0)	151	0.015 (0.012)	0.997 (0.003)
			24	639.0 (40.5)	573	0.020 (0.008)	0.994 (0.004)	658.8 (39.5)	595	0.020 (0.007)	0.994 (0.003)
		100	12	163.7 (22.5)	141	0.011 (0.007)	0.998 (0.001)	172.2 (24.5)	151	0.010 (0.007)	0.999 (0.001)
			24	844.6 (83.5)	573	0.021 (0.004)	0.994 (0.002)	852.1 (79.2)	595	0.021 (0.003)	0.994 (0.002)
	Semi	250	12	150.9 (17.2)	141	0.015 (0.012)	0.997 (0.003)	160.3 (17.7)	151	0.014 (0.011)	0.997 (0.003)
			24	645.9 (40.1)	573	0.022 (0.007)	0.993 (0.004)	663.9 (39.5)	595	0.020 (0.007)	0.994 (0.004)
		1000	12	174.3 (27.5)	141	0.014 (0.007)	0.998 (0.002)	182.6 (27.3)	151	0.013 (0.007)	0.998 (0.002)
			24	843.8 (69.4)	573	0.022 (0.003)	0.994 (0.002)	851.2 (67.6)	595	0.021 (0.003)	0.994 (0.002)
	Unbal	250	12	139.5 (16.4)	141	0.007 (0.011)	0.999 (0.002)	144.4 (16.3)	151	0.005 (0.009)	0.999 (0.001)
			24	573.4 (22.3)	573	0.005 (0.006)	0.999 (0.001)	591.3 (22.9)	595	0.004 (0.006)	1.000 (0.001)
		1000	12	134.1 (15.8)	141	0.003 (0.004)	1.000 (0.000)	140.6 (16.3)	151	0.002 (0.004)	1.000 (0.000)
			24	582.1 (28.1)	573	0.004 (0.004)	1.000 (0.000)	597.5 (30.0)	595	0.003 (0.004)	1.000 (0.000)

Note. ARS = ARS factor loadings ; N = sample size within each group; Bal = Balanced scale; Semi = Semi-balanced scale; Unbal = Unbalanced scale; J = number of items.

Table 10

Average fit value for MI testing approaches when the ARS factor is included in multidimensional scales in function of the simulated conditions

| | | | | Average fit value in MI testing for multidimensional scales with ARS factor | | | | | | | |
| | | | | Configural | | | | Thresholds | | | |
ARS	**Scale**	**N**	**J**	χ^2 **(SD)**	**df**	**RMSEA (SD)**	**CFI (SD)**	χ^2_ν **(SD)**	**df_ν**	**$RMSEA_\nu$ (SD)**	**CFI_ν (SD)**
0.175	Bal	250	12	108.2 (15.2)	104	0.013 (0.012)	0.994 (0.008)	132.5 (17.0)	128	0.012 (0.012)	0.993 (0.009)
			24	514.7 (26.8)	500	0.010 (0.008)	0.995 (0.005)	562.9 (27.9)	548	0.010 (0.007)	0.995 (0.005)
		1000	12	104.2 (14.4)	104	0.005 (0.006)	0.999 (0.002)	127.8 (16.0)	128	0.005 (0.005)	0.999 (0.002)
			24	507.0 (30.9)	500	0.004 (0.004)	0.999 (0.001)	555.1 (31.9)	548	0.004 (0.004)	0.999 (0.001)
	Semi	250	12	109.4 (13.4)	104	0.013 (0.012)	0.994 (0.008)	134.0 (15.1)	128	0.013 (0.012)	0.993 (0.008)
			24	512.1 (29.1)	500	0.009 (0.008)	0.996 (0.006)	560.5 (30.6)	548	0.008 (0.008)	0.995 (0.006)
		1000	12	103.9 (14.5)	104	0.005 (0.006)	0.999 (0.002)	128.0 (16.0)	128	0.005 (0.005)	0.999 (0.002)
			24	504.9 (27.2)	500	0.004 (0.004)	0.999 (0.001)	552.8 (29.1)	548	0.004 (0.004)	0.999 (0.001)
	Unbal	250	12	99.0 (13.0)	104	0.006 (0.009)	0.998 (0.005)	120.5 (12.0)	128	0.004 (0.007)	0.999 (0.003)
			24	494.5 (24.1)	500	0.004 (0.006)	0.998 (0.003)	550.3 (22.4)	548	0.006 (0.006)	0.998 (0.003)
		1000	12	95.8 (10.2)	104	0.002 (0.003)	1.000 (0.001)	117.9 (11.9)	128	0.001 (0.003)	1.000 (0.001)
			24	480.4 (30.3)	500	0.001 (0.003)	1.000 (0.001)	538.8 (31.5)	548	0.002 (0.003)	1.000 (0.001)
0.350	Bal	250	12	113.5 (14.2)	104	0.017 (0.013)	0.995 (0.006)	136.9 (15.0)	128	0.014 (0.012)	0.995 (0.006)
			24	539.8 (27.6)	500	0.017 (0.007)	0.994 (0.004)	577.0 (29.3)	548	0.013 (0.008)	0.995 (0.004)
		100	12	134.8 (18.5)	104	0.016 (0.006)	0.996 (0.002)	158.1 (20.9)	128	0.014 (0.006)	0.996 (0.002)
			24	620.7 (39.9)	500	0.015 (0.003)	0.995 (0.001)	623.9 (39.1)	548	0.011 (0.003)	0.997 (0.001)
	Semi	250	12	111.8 (15.7)	104	0.016 (0.013)	0.995 (0.006)	135.2 (17.6)	128	0.014 (0.013)	0.995 (0.006)
			24	531.6 (28.7)	500	0.014 (0.008)	0.995 (0.004)	570.6 (28.8)	548	0.011 (0.008)	0.996 (0.004)
		1000	12	128.5 (20.4)	104	0.014 (0.007)	0.997 (0.002)	152.7 (20.3)	128	0.013 (0.006)	0.997 (0.002)
			24	586.1 (36.3)	500	0.013 (0.003)	0.997 (0.001)	604.8 (34.9)	548	0.010 (0.004)	0.998 (0.001)
	Unbal	250	12	93.9 (11.9)	104	0.003 (0.007)	1.000 (0.001)	118.5 (14.0)	128	0.004 (0.007)	0.999 (0.001)
			24	494.5 (25.4)	500	0.004 (0.006)	0.999 (0.001)	541.7 (25.0)	548	0.004 (0.006)	0.999 (0.001)
		1000	12	98.0 (10.7)	104	0.002 (0.004)	1.000 (0.000)	122.6 (13.7)	128	0.003 (0.004)	1.000 (0.000)
			24	474.2 (18.7)	500	0.000 (0.000)	1.000 (0.000)	541.9 (24.6)	548	0.002 (0.003)	1.000 (0.000)

Note. ARS = ARS factor loadings; N = sample size within each group; Bal = Balanced scale; Semi = Semi-balanced scale; Unbal = Unbalanced scale; J = number of items.

Table 11

Average fit value for MI testing approaches when the ARS factor is included in multidimensional scales in function of the simulated conditions

				Average fit value in MI testing for multidimensional scales with ARS factor							
				Loadings				Intercepts			
ARS	Scale	N	J	χ^2_λ (SD)	df_λ	$RMSEA_\lambda$ (SD)	CFI_λ (SD)	χ^2_ν (SD)	df_ν	$RMSEA_\nu$ (SD)	CFI_ν (SD)
0.175	Bal	250	12	143.0 (16.4)	138	0.012 (0.011)	0.993 (0.009)	152.7 (16.8)	147	0.012 (0.011)	0.992 (0.009)
			24	584.2 (28.3)	570	0.010 (0.007)	0.995 (0.005)	603.8 (27.5)	591	0.009 (0.007)	0.995 (0.005)
		1000	12	137.0 (17.3)	138	0.004 (0.005)	0.999 (0.002)	145.6 (17.9)	147	0.004 (0.005)	0.999 (0.002)
			24	577.5 (31.2)	570	0.004 (0.004)	0.999 (0.001)	596.3 (32.9)	591	0.004 (0.004)	0.999 (0.001)
	Semi	250	12	143.9 (15.5)	138	0.012 (0.011)	0.993 (0.009)	152.2 (17.2)	147	0.012 (0.012)	0.993 (0.009)
			24	582.7 (31.7)	570	0.008 (0.008)	0.995 (0.006)	604.3 (33.8)	591	0.008 (0.008)	0.995 (0.007)
		1000	12	137.7 (15.7)	138	0.004 (0.005)	0.999 (0.002)	146.6 (16.7)	147	0.004 (0.005)	0.999 (0.002)
			24	574.7 (28.8)	570	0.003 (0.004)	0.999 (0.001)	595.0 (29.1)	591	0.003 (0.004)	0.999 (0.001)
	Unbal	250	12	135.2 (15.1)	138	0.007 (0.010)	0.997 (0.006)	141.4 (15.8)	147	0.005 (0.009)	0.997 (0.006)
			24	575.3 (29.9)	570	0.007 (0.007)	0.997 (0.005)	592.6 (29.0)	591	0.006 (0.007)	0.997 (0.004)
		1000	12	130.4 (13.6)	138	0.002 (0.004)	1.000 (0.001)	136.4 (13.8)	147	0.001 (0.003)	1.000 (0.001)
			24	558.3 (34.9)	570	0.002 (0.003)	0.999 (0.001)	574.5 (34.9)	591	0.002 (0.003)	1.000 (0.001)
0.350	Bal	250	12	150.1 (15.9)	138	0.016 (0.012)	0.993 (0.007)	158.5 (17.0)	147	0.016 (0.012)	0.993 (0.007)
			24	610.5 (31.7)	570	0.016 (0.007)	0.994 (0.004)	630.5 (33.6)	591	0.015 (0.008)	0.994 (0.004)
		100	12	178.0 (23.0)	138	0.016 (0.006)	0.995 (0.003)	187.7 (23.7)	147	0.016 (0.006)	0.995 (0.003)
			24	696.2 (43.7)	570	0.015 (0.003)	0.995 (0.002)	715.2 (44.8)	591	0.014 (0.003)	0.995 (0.002)
	Semi	250	12	147.7 (18.6)	138	0.015 (0.012)	0.994 (0.007)	156.0 (18.0)	147	0.015 (0.012)	0.994 (0.006)
			24	601.7 (29.4)	570	0.013 (0.007)	0.995 (0.004)	621.0 (29.5)	591	0.013 (0.007)	0.995 (0.004)
		1000	12	168.3 (22.6)	138	0.014 (0.006)	0.997 (0.003)	175.5 (23.3)	147	0.013 (0.006)	0.997 (0.002)
			24	656.2 (40.1)	570	0.012 (0.003)	0.997 (0.002)	676.6 (42.4)	591	0.012 (0.003)	0.997 (0.002)
	Unbal	250	12	130.8 (16.1)	138	0.005 (0.009)	0.999 (0.002)	138.6 (16.2)	147	0.004 (0.008)	0.999 (0.002)
			24	564.6 (24.7)	570	0.004 (0.006)	0.999 (0.001)	582.6 (24.0)	591	0.004 (0.005)	0.999 (0.001)
		1000	12	135.3 (15.8)	138	0.004 (0.005)	1.000 (0.001)	140.7 (16.9)	147	0.003 (0.004)	1.000 (0.001)
			24	563.8 (31.7)	570	0.002 (0.003)	1.000 (0.000)	583.2 (31.0)	591	0.002 (0.003)	1.000 (0.000)

Note. ARS = ARS factor loadings; N = sample size within each group; Bal = Balanced scale; Semi = Semi-balanced scale; Unbal = Unbalanced scale; J = number of items.