Social Science Research 110 (2023) 102805



Contents lists available at ScienceDirect

Social Science Research

journal homepage: www.elsevier.com/locate/ssresearch





Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives

Heinz Leitgöb [a,b,*], Daniel Seddig [c,d], Tihomir Asparouhov [e], Dorothée Behr [f], Eldad Davidov [c,g], Kim De Roover [h,i], Suzanne Jak [j], Katharina Meitinger [k], Natalja Menold [l], Bengt Muthén [e,m], Maksim Rudnev [n], Peter Schmidt [o,p], Rens van de Schoot [k]

[a] University of Leipzig, Germany
[b] University of Frankfurt, Germany
[c] University of Cologne, Germany
[d] University of Münster, Germany
[e] Mplus, USA
[f] GESIS – Leibniz Institute for the Social Sciences, Germany
[g] University of Zurich and URPP Social Networks, Switzerland
[h] Tilburg University, the Netherlands
[i] KU Leuven, Belgium
[j] University of Amsterdam, the Netherlands
[k] Utrecht University, the Netherlands
[l] Technische Universitaet Dresden, Germany
[m] University of California, USA
[n] University of Waterloo, Canada
[o] University of Giessen, Germany
[p] University of Mainz, Germany

ARTICLE INFO

Keywords:
Comparative research
Measurement invariance
Item bias
Noninvariance detection
Multiple group confirmatory factor analysis
Scale construction

ABSTRACT

This review summarizes the current state of the art of statistical and (survey) methodological research on measurement (non)invariance, which is considered a core challenge for the comparative social sciences. After outlining the historical roots, conceptual details, and standard procedures for measurement invariance testing, the paper focuses in particular on the statistical developments that have been achieved in the last 10 years. These include Bayesian approximate measurement invariance, the alignment method, measurement invariance testing within the multilevel modeling framework, mixture multigroup factor analysis, the measurement invariance explorer, and the response shift-true change decomposition approach. Furthermore, the contribution of survey methodological research to the construction of invariant measurement instruments is explicitly addressed and highlighted, including the issues of design decisions, pretesting, scale adoption, and translation. The paper ends with an outlook on future research perspectives.

* Corresponding author. University of Leipzig, Beethovenstraße 15, 04107 Leipzig, Germany.
E-mail address: heinz.leitgoeb@uni-leipzig.de (H. Leitgöb).

https://doi.org/10.1016/j.ssresearch.2022.102805
Received 17 May 2022; Received in revised form 30 September 2022; Accepted 2 October 2022
Available online 31 October 2022
0049-089X/© 2022 Elsevier Inc. All rights reserved.

1. Introduction

Émile Durkheim (1982; the original was published in 1895) emphasized that the comparative method is the core of social science research. Accordingly, social phenomena are often considered unique to a particular context and "indirect experimentation" (another term he coined for the comparative method) is the best suitable method to identify what is evident in one context and not in another and, thus, what might cause differences. Today, the increasing availability and quality of cross-national and longitudinal data from large-scale studies such as the European Social Survey (ESS), International Social Survey Program (ISSP), European Values Study (EVS), or World Values Survey (WVS) have stimulated comparative analyses in many fields of the social and behavioral sciences. A prerequisite for valid comparisons is that the same phenomena are actually observed or measured across contexts. This is the basis of the measurement invariance (MI) problem.

The social and behavioral sciences often consider theoretical constructs that are not directly observable (e.g., intelligence, motivations, attitudes, values) but measured by multiple observable indicators—with varying degrees of accuracy. This measurement process has been conceptualized with various measurement models. For example, the classical test theory (CTT) model (Lord and Novick, 1968) is based on the notion that an observed measurement score for a theoretical construct of interest can be decomposed into a true score and a random error component:

$$O_{bserved} = T_{rue} + E_{rror}. \tag{1}$$

The true score is sometimes equated with the construct score, that is, the real score on an underlying construct. However, this interpretation may be problematic for several reasons (Borsboom, 2005). Other measurement models allow theoretical constructs to be more accurately represented as latent variables that are considered to be the causes of the observed indicators (see Bollen, 2002). Accordingly, latent variables (sometimes referred to as *common factors*) are assumed to explain variation in observed indicators and residual variation is due to measurement error, which may be systematic or random. This view is the foundation of factor analytical approaches such as exploratory factor analysis (EFA; e.g., Lawley, 1943; Lawley and Maxwell, 1963; Spearman, 1904; Thurstone, 1947) and confirmatory factor analysis (CFA; e.g., Jöreskog, 1969; Wiley et al., 1973) as well as item response theory (IRT; e.g., de Ayala, 2022; Guttman, 1945; Lord, 1952, 1980; Rasch, 1960; Samejima, 1969).

From early on, researchers in the factor analytic field were aware of the MI problem (see Millsap and Meredith, 2007).[1] First, a theoretical line of thought emerged that assumes (non)invariance in factor analytic models is affected by selection (i.e., processes that created the subpopulations of interest). Selection may operate directly on the observed variables under study and on the factor structure (e.g., when subgroups are related to a subset of the observed variables) or on unobserved variables (e.g., characteristics such as gender, socioeconomic status, age, or ethnicity indicating differences between subpopulations). The selective composition of subpopulations may lead to correlations among specific factors (unique error components) and between specific and common factors, which is a violation of factor analytic assumptions (see also Meredith and Teresi, 2006). A factor structure is assumed invariant when the selection variables are independent from specific factors implying that the selection process is only operating via the common factors. A second line of thought was concerned with rotational strategies for identifying the best-fitting invariant factor patterns, which allowed analytical approaches to assess empirically whether factor structures are invariant. However, applying rotational procedures to establish invariance only affects a particular aspect of a factor model (i.e., the factor loadings) and does not offer any fit information beyond that of the original factor solution. This means that invariance is established only after the estimation of the factor solution and invariance assumptions are not tested. A more comprehensive and testable approach was proposed by Jöreskog (1971) who extended CFA to multiple group confirmatory factor analysis (MGCFA), which today is the predominant tool for testing the invariance of measurement model parameters across subpopulations.[2] In the half century since this seminal publication, a number of review articles have taken up the developments and recommendations regarding MI testing with MGCFA and described the state of the art at the respective time of writing (e.g., Davidov et al., 2014, 2018a; Hui and Triandis, 1985; Steenkamp and Baumgartner, 1998; Vandenberg and Lance, 2000; Vandenberg, 2002; van de Schoot et al., 2015). At the time of their publication, these excellent reviews provided the available knowledge of MI in a concise form to a wide readership of students, teachers, and researchers.

In this overview, we aim to pick up where others have left off, tracing the historical developments of the most common strategies for identifying and dealing with measurement (non)invariance. We think it is therefore necessary to begin by reiterating what MI actually is and how it can be tested with the first generation of approaches based on the traditional MGCFA and multiple indicators multiple causes (MIMIC) models (section 2).

We then present the next generation of approaches (section 3) that were developed to address the question that has long puzzled and still puzzles comparative researchers using the traditional approach: What if the empirical data provide no support for (exact and partial) MI? This situation often occurs in cross-cultural research when comparing a large number of groups (e.g., countries) or in longitudinal research when comparing many periods or periods far apart in time. The proposed solution is moving away from the

[1] We use the more general term of measurement invariance throughout this article. Some researchers refer to the invariance problem using the narrower concept of factorial invariance (see Meredith and Teresi, 2006).

[2] In the IRT framework, MI has been discussed in terms of item bias (Mellenbergh, 1989) and is sometimes referred to as differential item functioning (DIF; e.g., Millsap and Everson, 1993). We recognize that from this very similar perspective, the invariance problem can be approached with at least the same rigor as with the MGCFA approach (see, e.g., Teresi, 2006). However, we limit our presentation to the latter and refer to relevant sources for specific situations in the context of multiple indicators multiple causes (MIMIC) modeling (section 2.4) and categorical data analysis.

concept of exact MI and assuming instead approximate MI. The Bayesian approximate measurement invariance (BAMI) approach assumes that measurement parameter differences across groups or time may be present but that they are small and substantively insignificant. The multiple group alignment approach assumes that most measurement parameters are invariant across groups and only a small subset of the parameters is noninvariant, allowing that group means and variances may still be compared even in the presence of some tolerable degree of noninvariance.

We continue with the discussion of the multilevel approach for testing MI and the statistical explanation of identified noninvariance (section 4). Multilevel analysis is particularly useful when the number of groups (or clusters) is large and invariance assessments with MGCFA or MIMIC models would require evaluating a very large number of parameters. In addition, the multilevel framework can help understanding noninvariance findings in unique ways, such as explaining noninvariance using observed cluster-level variables.

The latest generation of approaches addresses situations in which MI does not hold across all clusters/groups but only for subsets or when measurement parameter differences across groups or time are to be examined with respect to their proportion of change that reflects true change in the construct. Subset-specific MI (section 5) may be relevant when groups are not independent, for example, when countries belong to a particular cultural region. One way to detect clustered invariance is mixture multigroup factor analysis, which assumes subset-specific measurement parameters that are shared by groups within subsets but may differ between subsets. The approach can also be used to explain noninvariance between subsets. Another way to find subsets of groups that display MI is to quantify information on the desired MI level across each pair of groups and use it to infer connections between groups in network-type graphs. Such visualization techniques are implemented in a tool called the Measurement Invariance Explorer. The decomposition of measurement parameter differences (section 6) is based on a longitudinal CFA model and separates the amount of the true change in measurement parameters over time from change that appears due to systematic response shifts. Thus, even in the case of noninvariant measurements, empirical data can provide useful information about changes in constructs and the amount and causes of noninvariance.

Finally, we consider approaches that do not deal with statistical solutions or modeling, but are based on *survey methodological* ideas that can address the MI problem before any data are collected or analyzed (section 7). This involves decisions about rating scales and survey modes, the use of cognitive pretesting and web probing approaches, and cross-cultural scale adoption and translation methods.

In the following sections, we will present the different approaches along the developmental lines outlined above. In the course of the discussion (section 8), we will also address other research areas that seem to be promising avenues toward a deeper understanding of measurement (non)invariance and its causes and consequences.

2. Measurement invariance and the traditional global testing approach

2.1. What is measurement invariance?

Measurement invariance (MI), sometimes also referred to as measurement equivalence or measurement comparability, was defined by Horn and McArdle (1992, p. 117) as "whether or not, under different conditions of observing and studying a phenomenon, measurement observations yield measures of the same attribute." In other words, MI of a theoretical construct of interest is a measurement

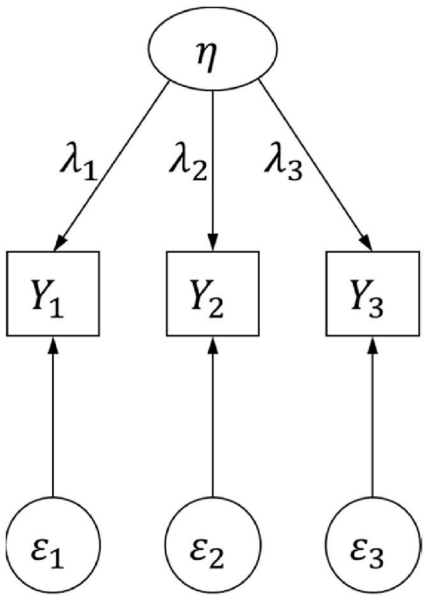


Fig. 1. Single group reflective measurement model for latent variable η

characteristic. It implies that in different contexts such as cultures, geographical areas, countries, time points, (Davidov et al., 2018a,b; Sokolov, 2018), language groups (Davidov and de Beuckelaer, 2010), methods of data collection (e.g. Cieciuch and Davidov, 2016; Davidov and Depner, 2011; Gordoni et al., 2012), or among other meaningful objects of comparison, the same concept is being measured in the same way (Billiet, 2003; He and van de Vijver, 2012; Johnson, 1998; Steenkamp and Baumgartner, 1998; Vandenberg and Lance, 2000; van de Vijver, 2018a,b; van de Vijver and Leung, 2011; van de Vijver and Poortinga, 1997). MI should not be confused with equality of measurement scores across groups. *Equality* of scores implies that the measurement scores in two groups are identical, that is, the two groups score similarly in some theoretical construct of interest, for instance, displaying the same level of negative attitudes toward immigration. MI implies that they are comparable. Comparability does not imply, however, that the scores are equal, but only that they may be compared meaningfully across groups. Thus, comparable scores may or may not be equal. MI simply guarantees that the measures at hand can be compared and that we are, in fact, not comparing "chopsticks with forks" (Chen, 2008).

A review study (Davidov et al., 2014) suggests that although the concept and crucial importance of MI have been determined already more than 30 years ago (Meredith, 1993), only a very small share of comparative studies in the social sciences, particularly in sociology and political science, examine MI before comparing theoretical constructs of interest and their measurement parameters. This is unfortunate, because the literature has repeatedly shown that if MI is not given, equal scores may reflect a methodological artefact when true scores are actually different across groups. At the same time, different scores may conceal equality of scores due to methodological bias (Davidov et al., 2014; Millsap, 2011; van de Vijver et al., 2019).

To describe MI in less abstract terms, let us first consider the simplest example with a single-group reflective measurement model for one latent factor η (e.g., attitudes toward immigration) measured by three manifest items, y_1 to y_3, also referred to as *indicators* (see Fig. 1). The indicators reflect the extent to which the respondents agree that immigrants should be welcome into the country (y_1), provided support upon arrival (y_2), and encouraged to come (y_3). At this stage, we assume that the responses are captured by some rating scale, ranging between 1 (fully agree) and 7 (totally disagree). Since measurements of opinions are susceptible to measurement errors, each item has a measurement error (ε_1 to ε_3, respectively). In addition, each item has a factor loading (λ_1 to λ_3) that relates it to the factor. If we further assume that we would like to compare both the mean of this latent variable η and its association (e.g., covariance) with the covariate "age" across two groups (e.g., two countries, Germany and the U.S.), then MI becomes relevant. Only if the measures of the latent variable "attitudes toward immigration" are invariant, can we compare its means and associations across countries meaningfully as explained below. Otherwise, estimated cross-country similarities or differences in its mean or covariance with the variable "age" may not only be a result of true differences but also a methodological artefact. The amount of tolerable measurement noninvariance bias can be studied using robustness studies either analytically or using Monte Carlo simulations (see, e.g., Kuha and Moustaki, 2015; Meuleman, 2012; Oberski, 2014).

2.2. Multiple group confirmatory factor analysis (MGCFA)

There are various techniques to examine MI (Braun and Johnson, 2010; for an overview see also Kim et al., 2017; Davidov et al., 2018a,b). However, the most popular way discussed in the literature is the multiple group confirmatory factor analysis approach (MGCFA; Jöreskog, 1971; Sörbom, 1974) within the structural equation modeling (SEM; Bollen, 1989) perspective. To formalize the standard MGCFA model, assume, $i = 1, ..., n$ is the number of individual observations, $p = 1, ..., P$ is the number of observed indicators, $g = 1, ...G$ is the number of groups to be compared, $\boldsymbol{y}_{ig}$ is the $P \times 1$ vector of observed indicator scores for individual i from group g, $\boldsymbol{\Lambda}_g$ is a $P \times 1$ group-specific vector of factor loadings, $\boldsymbol{\nu}_g$ is the respective vector of intercepts, η_{ig} is a scalar[3] representing the latent score of i from group g, and finally, $\boldsymbol{\varepsilon}_{ig}$ is a $P \times 1$ vector of error terms assumed to be normally distributed $N(\boldsymbol{0}, \boldsymbol{\Theta}_g)$, with $\boldsymbol{\Theta}_g = \mathrm{Cov}(\boldsymbol{\varepsilon}_g, \boldsymbol{\varepsilon}_g^{'})$ being the group-specific error covariance matrix. Then, the measurement model can be written as:

$$\boldsymbol{y}_{ig} = \boldsymbol{\nu}_g + \boldsymbol{\Lambda}_g \eta_{ig} + \boldsymbol{\varepsilon}_{ig.} \tag{2}$$

Furthermore, the respective *mean and covariance structure* (MACS; e.g., Little, 1997) equations are specified as follows:

$$\boldsymbol{\mu}_g = \boldsymbol{\nu}_g + \boldsymbol{\Lambda}_g \alpha_g \tag{3}$$

$$\mathrm{Cov}\left(\boldsymbol{y}_g, \boldsymbol{y}_g^{'}\right) = \boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \psi_g \boldsymbol{\Lambda}_g^{'} + \boldsymbol{\Theta}_g, \tag{4}$$

with $\boldsymbol{\mu}_g$ as the observed means of $\boldsymbol{y}_g$, as well as α_g and ψ_g as the latent mean and variance of η_g, which is also assumed to follow a normal distribution: $\eta_g \sim N(\alpha_g, \psi_g)$.

The MI literature, dating back to the seminal work of Meredith (1993), distinguishes between different levels of MI, following a hierarchical structure (for a summary, see Table 1). The first and lowest level is called configural invariance. Descriptively, it implies that the same indicators can be used in all G groups to measure the same underlying latent construct. We can test it by examining whether the model measuring the same latent variable fits the data in all groups, and whether the standardized factor loadings are also acceptable (e.g., above 0.3–0.4, see Brown, 2015). That is, the model fit can be evaluated using conventional tests within each group (see, e.g., Hu and Bentler, 1999; Marsh et al., 2004; West et al., 2012). However, even if configural invariance is supported by the data,

[3] Without loss of generality, we restrict the model to a single latent variable η for illustration purposes.

Table 1
Levels of measurement invariance.[4]

Invariance level	What it implies	Type of comparison across groups allowed	How the invariance level may be assessed
Configural invariance	The same items measuring the same constructs across groups	None	An MGCFA suggesting an acceptable fit to the data
(Full or partial) Metric invariance	The same items have the same factor loadings across groups (at least two equal factor loadings for partial metric invariance)	Unstandardized associations (covariances, unstandardized regression coefficients with other theoretical constructs of interest)	The model fit does not deteriorate considerably compared to the configural invariance model
(Full or partial) Scalar invariance	The same items have the same factor loadings *and* intercepts across groups (at least two items with equal factor loadings and intercepts for partial scalar invariance)	Unstandardized associations *and* latent means	The model fit does not deteriorate considerably compared to the (full or partial) metric invariance model
Strict invariance	The same items have the same factor loadings, intercepts, *and* error variances across groups	Unstandardized associations *and* latent means	The model fit does not deteriorate considerably compared to the (full or partial) scalar invariance model

it still does not allow us to make any meaningful comparisons. The reason is that meaningful comparisons require higher levels of invariance, as explained below, because similarities or differences may be due to methodological artefacts rather than due to true differences.

The next MI level is metric (or weak) invariance. Metric invariance implies that the factor loadings of the items measuring the latent variable are the same across groups:

$$\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2 = \ldots = \boldsymbol{\Lambda}_g = \ldots = \boldsymbol{\Lambda}_G \tag{5}$$

One can test whether metric invariance is given by comparing the fit of this more restricted model to the model fit of the configural model. If the model fit does not deteriorate considerably (e.g., if the chi-square different test is not significant, or if the CFI value does not deteriorate by more than approximately 0.01; see, e.g., Byrne, 2004; Byrne and Stewart, 2006; Chen, 2007; Cheung and Rensvold, 2002), one may assume that metric invariance is supported by the data. Indeed, previously, the chi-square difference test was applied to test different types of invariance. However, later simulations showed that this test is often too restrictive, and sensitive to sample size. Therefore, other criteria were proposed. If metric invariance is given, one may compare unstandardized associations (i.e., covariances or unstandardized regression coefficients between the latent variables of interest for which metric invariance holds). Considering our previous example, one may compare in that case the covariance or the unstandardized regression coefficient between the variable "age" and the latent variable "attitudes toward immigration." However, even if metric invariance is given, the model still does not allow comparing means with confidence. Meaningful comparisons of means require a higher level of invariance, scalar (or strong) invariance (Meredith, 1993).

Scalar invariance implies that not only the factor loadings but also the item intercepts are equal across groups:

$$\boldsymbol{\nu}_1 = \boldsymbol{\nu}_2 = \ldots = \boldsymbol{\nu}_g = \ldots = \boldsymbol{\nu}_G. \tag{6}$$

A state of the art implementation of the global testing approach, based on the described comparison of a sequence of nested MGCFA models, is provided in Putnick and Bornstein (2016).

In Table 1, we summarize the levels of MI and the types of substantive comparisons for which they are required.

One can test whether scalar invariance across groups is given by comparing the fit of this more restricted model to the model fit of the metric model. If the model fit does not deteriorate considerably (e.g., if the CFI value does not deteriorate by more than approximately 0.01, see Chen, 2007; see also Little et al., 2006, and Sörbom, 1974 on how to identify the model), one may assume that scalar invariance is supported by the data. Now also the latent means of the latent variable may be compared across groups with confidence. Little et al. (2006) have demonstrated different methods that can be used to identify the models to allow the comparison of latent means (see also Kim and Yoon, 2011; Raykov et al., 2013; Stark et al., 2006; Thompson et al., 2021; Yoon and Millsap, 2007).

The routine assessment of whether the underlying latent constructs meet the MI properties has increased considerably in all forms of comparative studies (i.e., cross-group/cultural/national or longitudinal) in recent years. Among others, these studies examined the cross-cultural properties of various measurements and scales such as nationhood and national identity (Davidov, 2009; Sarrasin et al., 2012), authoritarianism (Heyder and Schmidt, 2003), social trust (Coromina and Davidov, 2013; Freitag and Bauer, 2013; van der Veld and Saris, 2018), physical and mental health (e.g., Maskileyson et al., 2021a,b), attitudes toward immigration (e.g., Becker et al., 2020; Davidov et al., 2015; Davidov et al., 2018c; Munck et al., 2018), attitudes toward democracy (Ariely and Davidov, 2010; Davidov and Braun, 2012), left-right orientation (Weber, 2011), religiosity (e.g., Remizova et al., 2022) ageism (Seddig et al., 2020), gender role attitudes (e.g., Lomazzi, 2018; Seddig and Lomazzi, 2019), or basic human values (e.g., Cieciuch et al., 2014, 2018, 2019; Davidov et al., 2008b; Davidov and Siegers, 2010). In addition, various special issues have been dedicated to the topic (e.g., Davidov et al., 2018a; Meitinger et al., 2020; Meuleman et al., 2018a,b; van de Schoot et al., 2015). Furthermore, this procedure was generalized to higher-order factor analysis (Chen et al., 2005; Marsh and Hocevar, 1985), and recently applied to the alienation scale by Rudnev et al. (2018).

Finally, it should be noted that the literature also refers to the term "*strict* invariance", which indicates that measurement errors are also equivalent across groups (see, e.g., Steenkamp and Baumgartner, 1998; see also Table 1). However, since strict invariance does not

have any direct consequences for the comparability of structural parameters across groups, most applied studies refrain from its evaluation.

2.3. Partial measurement invariance

The models described above considered conditions in which all factor loadings (i.e., metric invariance) or all factor loadings and intercepts (i.e., scalar invariance) are exactly equal. This condition is referred to as full exact metric invariance or scalar invariance. Unfortunately, although survey research has invested considerable efforts to increase the comparability of survey data across nations or cultural groups (Harkness et al., 2010a; Johnson et al., 2019; Jowell et al., 2007; Lynn et al., 2006; Roberts et al., 2020; see also section 5), it is rarely the case that full invariance is supported when survey data are used (Byrne and van de Vijver, 2010). The absence of (exact) invariance may have different reasons on the individual or group level causing item parameters to behave differently across groups. This situation is often referred to as differential item functioning (DIF; see also section 2.4). The variability of item parameters across groups might potentially be explained by either group-level variables (e.g., the Human Development Index in a country; see, e.g., Davidov et al., 2012; see also section 4) or by individual level variables (e.g., education) or by both (see, e.g., Welkenhuysen-Gybels and Billiet, 2002). However, Byrne et al. (1989) as well as Steenkamp and Baumgartner (1998) suggested that it may be sufficient to guarantee that only two items display equal factor loadings and intercepts across groups. They termed this situation as partial (rather than full) invariance. Whereas a few studies found that relying on partial invariance is insufficient (e.g., De Beuckelaer and Swinnen, 2018; Steinmetz, 2013), recent simulations by Pokropek et al. (2019) reassessed this argument with a detailed simulation study and concluded that partial invariance is sufficient under various condition and performs as well as other more recent approaches discussed below.

2.4. Multiple indicator multiple causes modeling and measurement invariance

Within the SEM framework, MI (or DIF) can further be assessed by means of the multiple indicator multiple causes (MIMIC) modeling approach (Jöreskog and Goldberger, 1975; for an overview see, e.g., Brown, 2015; Lee et al., 2018). The advantages of the MIMIC approach over MGCFA to detect MI are smaller sample size requirements and greater parsimony (Brown, 2015). The standard MIMIC model extents the CFA model from Eqs. (2)–(4) (without group indicator g) by a structural equation, in which η, the latent construct of interest, is regressed on a set of $k = 1, \ldots, K$ observed covariates $\mathbf{X}$. To allow for the testing of differing intercept parameters (corresponding to the difficulty parameters in IRT; e.g., Lee et al., 2018) across groups, referred to as uniform DIF (e.g., Mellenbergh, 1989), Muthén (1988, 1989) specified the MIMIC DIF model by further regressing the manifest indicators $\mathbf{Y}$ on the covariates $\mathbf{X}$. The resulting two model equations can—again for the single latent construct case—be written as:

$$\text{Measurement}: \boldsymbol{y}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\eta_i + \mathbf{B}\mathbf{x}_i + \boldsymbol{\varepsilon}_i, \tag{7}$$

$$\text{Structural}: \ \eta_i = A + \boldsymbol{\Gamma}\mathbf{x}_i + \zeta_i, \tag{8}$$

with A as the intercept parameter, $\mathbf{B}$ as the $P \times K$ matrix of effect parameters capturing the linear influence of the covariates $\mathbf{X}$ on the observed responses $\mathbf{Y}$, $\boldsymbol{\Gamma}$ as the $1 \times K$ vector of effect parameters (path coefficients) representing the regression slopes of the latent factor η on the covariates $\mathbf{x}$, and ζ denotes the disturbance term of η, with $\zeta \sim N(0, \psi)$. In the typical multiple group case, the nominal group membership indicator X, which identifies the $g = 1, \ldots, G$ groups to be compared (e.g., the countries to be compared in a cross-country study), is split up into dummy variables (e.g., Kaplan, 2009; Thompson and Green, 2013). Then, the model contains a total of $K = G - 1$ covariates.

Fig. 2 offers a graphical representation of the MIMIC DIF model. For simplicity, we employ the $G = 2$ group case (e.g., with a treatment group and a control group), with the single binary group indicator x. The effect parameter γ represents the latent mean difference ($\Delta\alpha$) between the two underlying groups. It is also referred to as the *indirect effect* (Lee et al., 2018), because the impact of the group indicator X on the observed responses $\mathbf{Y}$ is mediated via the latent factor η. In contrast, the parameters from $\mathbf{B}$ represent the *direct effects* because they capture the direct (i.e., unmediated) impacts of X on $\mathbf{Y}$. In detail, β_p indicates whether the observed means in indicator p differ between the groups after having accounted for group differences in the respective latent means. Thus, $\beta_p \neq 0$ is a direct evidence for the presence of uniform DIF (Lee et al., 2018).

Analogous to the global MI testing strategy with MGCFA (section 2.2), one can simultaneously test whether any of the indicators is affected by uniform DIF. First, a *no-DIF* baseline model is estimated with zero constraints imposed on the direct effects, that is, $\mathbf{B} = \mathbf{0}$. Second, an unrestricted model is estimated, with $\mathbf{B}$ being freely estimated. Third, the model fit is compared. If both models fit the data equally well, the more parsimonious baseline model is preferred over the unrestricted model, indicating the absence of systematic uniform DIF. Alternatively, one could also specifically modify the no-DIF baseline model according to the evidence from the modification indices and expected parameter changes. Furthermore, several "one item at a time" testing strategies exist. For an overview of different MIMIC DIF-based testing strategies see, for example, Chun et al. (2016), Lee et al. (2018), as well as Thompson and Green (2013).

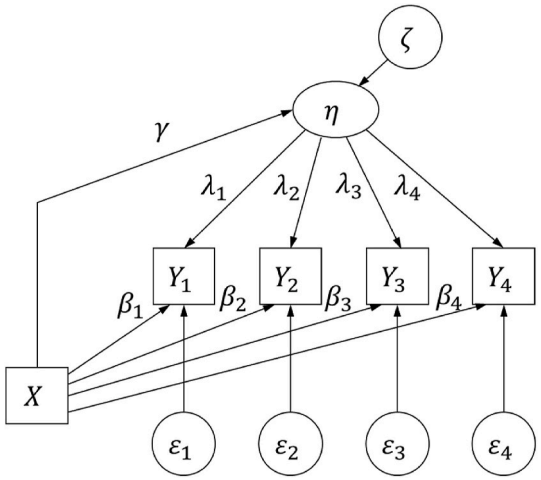


Fig. 2. The MIMIC DIF model with binary grouping variable.

As noted by various authors (e.g., Brown, 2015; Lee et al., 2018; Woods and Grimm, 2011), the standard MIMIC DIF model specification from Eqs. (7) and (8) introduces common factor loadings (corresponding to the discrimination parameters in IRT; e.g., Lee et al., 2018) for all G groups, which does not allow testing for noninvariance in the factor loadings across groups (nonuniform DIF). Thus, a good fit of the standard MIMIC DIF model is not indicative of invariant factor loadings (Kim et al., 2011). This limitation can be resolved by extending the model with an interaction between the latent factor η and the group indicator X (or the respective dummy variables; e.g., Barendse et al., 2010, 2012; Montoya and Jeon, 2020; Woods and Grimm, 2011). Furthermore, Hildebrandt et al. (2009) proposed the latent moderated structural (LMS) equations approach and the local structural equation modeling (LSEM; see also Olaru et al., 2019) to test naturally continuous covariates (e.g., age) for MI to avoid the arbitrary and thus, artificial categorization indispensable in MGCFA.

Recent innovations in MIMIC DIF modeling cover, among others, the specification of multilevel MIMIC DIF models (e.g., Kim and Cao, 2015; Kim et al., 2015; see also section 4), the elimination of the impact of extreme response styles in DIF detection (Jin and Chen, 2020), and the specification of MIMIC DIF models without having to select DIF-free indicators as reference indicators (also referred to as anchoring items; Chen et al., 2021).

3. Approximate measurement invariance methods

Approximate measurement invariance methods aim to accommodate noninvariance in the measurement model, while still providing a proper fit for the data and a proper comparison of factor means and variances across the groups. Here, we discuss the Bayesian approximate measurement invariance (BAMI) method and the alignment method. Underlying these methods is a substantive assumption regarding the scope of the noninvariance in the parameters. The BAMI method is based on the assumption that the measurement parameters may not be *exactly* identical across the groups (as assumed in section 2.2) but the differences are expected to be small and substantively insignificant. The alignment method is based on the assumption that most measurement parameters are invariant across groups while a small subset of the parameters is not invariant.

3.1. Bayesian approximate measurement invariance

With strict MI, the goal is to fit a measurement model in which any *small* measurement differences are assumed to be precisely zero. Any difference of factor loadings or intercepts between groups is forced to be zero. Such exact zero constraints may be overly strict, especially when fitting a model with many groups with minor parameter differences that cancel each other out, both within and between groups. As a consequence, a rigid MI model is frequently rejected. On the other end of the spectrum, there is the configural model, assuming there is no invariance, and all intercepts and factor loadings between groups are freely estimated. Such a model might fit the data best, but the latent factor means cannot be used for comparing groups. The Bayesian toolbox[5] can be used to gently push small differences in intercept and factor loadings between groups closer to zero, still yielding a well-fitting model. This has been called Bayesian approximate measurement invariance (BAMI), first described by Muthén and Asparouhov (2012), and later also by van de Schoot et al. (2013).

[4] In the longitudinal panel case, the equality of error covariances of identical indicators across time is sometimes introduced as a further invariance level after strict invariance (e.g., Leitgöb et al., 2021).

[5] For an introduction and overview of Bayesian (SEM) analysis, we refer the interested reader to Depaoli (2021), Gelman et al. (2004), Kaplan (2014), Kaplan and Depaoli (2012), Lee (2007), Song and Lee (2012), van de Schoot et al. (2014, 2021).

The BAMI method provides a model that allows comparing latent variables across (many) groups while allowing for some "wiggle room" to accommodate minor differences in intercept and factor loadings between groups.[6] The amount of wiggle room is determined by the degree of precision of the prior. That is, instead of restricting the differences between measurement parameters to zero (see Fig. 3A), BAMI assumes that these differences follow a (normal) distribution with a mean of zero and a prespecified prior variance: $N(0, \sigma_0)$. Such a prior balances model fit on the one hand and MI restrictions on the other (see Fig. 3B). Therefore, the BAMI option falls in between full and no MI, which could mean that one can still compare the latent factor means (as MI holds approximately) while the model also fits the data.

Bayesian statistics can be used to estimate the parameters from the MGCFA model formalized in Eqs. (2)–(4). Unique for Bayesian statistics is that all observed and unobserved parameters in a statistical model are given a joint probability distribution, termed the prior and data distributions. A Bayesian workflow captures available knowledge about a given parameter in a statistical model via the prior distribution. Information about the parameters available in the observed data is captured by the likelihood function. Both are combined using estimation techniques, like the Gibbs Sampler, in the form of the posterior distribution. The posterior distribution reflects one's updated knowledge, balancing prior knowledge with observed data, and is used to conduct inferences. For an extensive Bayesian literature overview, see van de Schoot et al. (2021).

When using the BAMI method, priors are not put on the parameters themselves (e.g., $\boldsymbol{\Lambda}_g$ or $\boldsymbol{\nu}_g$), but are placed on the covariances of a parameter between groups. Consider λ_{pg}, the factor loading for indicator p for two arbitrary groups g and g', the prior is put on $\lambda_{pg} - \lambda_{pg'}$ using

$$V(\lambda_{pg} - \lambda_{pg'}) = V(\lambda_{pg}) + V(\lambda_{pg'}) - 2\mathrm{Cov}(\lambda_{pg}, \lambda_{pg'}) \tag{9}$$

If we assume the variances to be 0.5, and the covariance 0.495, then $V(\lambda_{pg} - \lambda_{pg'}) = 0.01$. This method translates to a prior put on the difference for $\lambda_{pg} - \lambda_{pg'} \sim N(0, \sigma_0 = 0.01)$. Such a small variance prior results in $\lambda_{pg} \approx \lambda_{pg'}$ instead of $\lambda_{pg} = \lambda_{pg'}$ as would be the case with strict MI, or $\lambda_{pq} \neq \lambda_{pg'}$, as is the case with no (or partial) MI. The smaller the prior variance, σ_0, the more the results will be closer to strict MI, seeFig. 3B. Vice versa, the larger the prior variance, the more the results will reflect a configural model. Thus, using the largest prior variance possible allows for larger deviations from strict invariance while still being able to claim (approximate) measurement invariance.

The prior is highly informative because of the high precision of σ_0. It will significantly impact the posterior difference, which is intentional because the goal is to push the differences between groups toward zero. Therefore, it is essential to select the hyperparameters carefully (i.e., values for σ_0 of λ_{pg} or ν_{pg} for all group differences) and provide a rationale for the choice of the hyperparameters. Arts et al. (2021) provide an overview of studies applying BAMI with hyperparameter specifications and the underlying rationale of this choice—if present at all; see their supplementary materials at the Open Science Framework: https://osf.io/t3h9e. First of all, users of BAMI and any other Bayesian method should always include a rationale of their prior specification; see also the Bayesian analysis reporting guidelines by Kruschke (2021) or the "When to Worry and How to Avoid the Misuse of Bayesian Statistics" checklist by Depaoli and van de Schoot (2017).

One way of selecting σ_0 is by using the results of simulation studies (Kim et al., 2017; Lai et al., 2022; Lek et al., 2019; Muthén and Asparouhov, 2012; Pokropek et al., 2019, 2020; Shi et al., 2017; van de Schoot et al., 2013). The simulation results are not conclusive on which value to use for σ_0. Of course, it depends on the specifications of the model and the characteristics of the dataset. Therefore, the results of the simulation studies can only be used when the new dataset and model features are similar to the specifications of the original simulation design.

Another approach is to employ a sensitivity analysis (see, e.g., van Erp et al., 2018) and run the same model with different specifications of σ_0. The model with the smallest value for σ_0 but still yielding a good fit is preferred because it will bring the model closest to a strict MI model (see also Asparouhov et al., 2015). If multiple priors are used, it might be helpful to visualize the results of different models so that the effect of different prior specifications on group means and rankings becomes clear (see, e.g., Arts et al., 2021). They show that even when models appear to be a good fit to the data, there might still be an unwanted impact on the rank ordering of countries. Another strategy is to evaluate different models using metrics like the posterior predictive p-value (ppp-value; Gelman et al., 2004). However, as Hoijtink and van de Schoot (2018) show, the ppp-value is not suited for evaluating models with different prior specifications, and they propose the prior-posterior predictive p-value (pppp-value). A generalized version was implemented in Mplus by Asparouhov and Muthén (2017). The pppp-value can be used to evaluate the small variance parameters as demonstrated, for example, by Sideridis et al. (2020). As Asparouhov and Muthén (2017) point out: If the pppp-value does not reject the model, it means that there is no evidence in the data for the parameters in the model to be outside the prior distribution.

In conclusion, the advantages of the BAMI method have been demonstrated in cases with a large number of groups having many small differences in the intercepts and factor loadings between the groups, and these differences cancel each other out both within and between groups (Kim et al., 2017; Lai et al., 2022; Lek et al., 2019; Muthén and Asparouhov, 2012; Pokropek et al., 2019, 2020; Shi et al., 2017; van de Schoot et al., 2013). When differences are systematic or relatively large, one should be cautious in applying the approximate measurement testing procedure. In such situations, the small variance prior tends to pull the strongly deviating parameter estimates toward the average across groups resulting in the deviating parameter being smaller. In contrast, the invariant parameters will be larger than their true values (Muthén and Asparouhov, 2012). To reduce this bias, BAMI can be combined with the alignment procedure (Asparouhov and Muthén, 2014; see also section 3.2).

[6] Seddig and Leitgöb (2018) have introduced the BAMI method also for the longitudinal case within the CFA panel model.

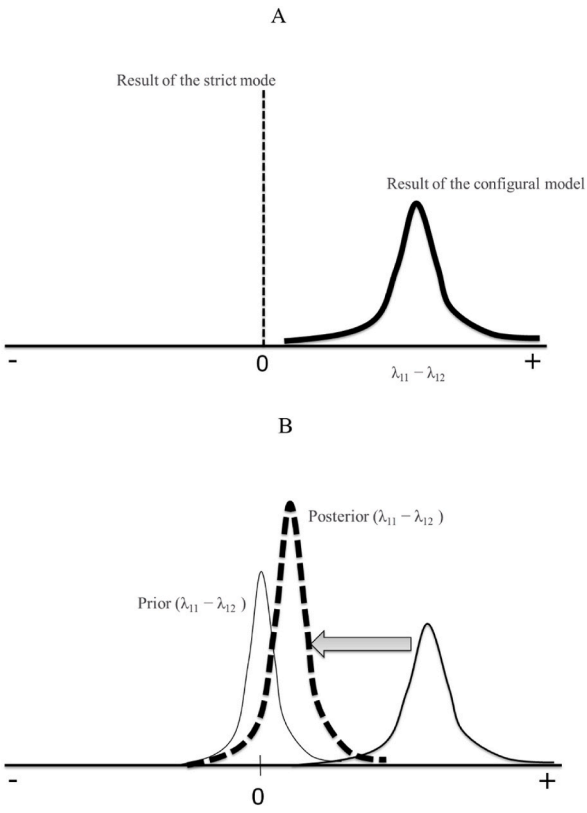


Fig. 3. The principles of exact (A) and Bayesian approximate (B) MI.

Future research should focus on how to best specify the prior for σ_0. That is, the aforementioned sensitivity and visualization methods are post hoc methods. As is recommended for Bayesian analyses in general, and BAMI is no exception, the specification of the hyperparameters for the priors should ideally be specified before running the model. Expert elicitation methods (O'Hagan et al., 2006) can be used to obtain background knowledge to specify the hyperparameters of a Bayesian model. Such methods have been applied in the field of SEM (see Garthwaite et al., 2013; Veen et al., 2020) but have not (yet) been applied to the degree of prior precision (i.e., values for σ_0) in BAMI models. Another method to obtain information about the amount of wiggle room is to use web probing, a mixed-methods approach to assess the comparability of items via collecting qualitative data from the respondents. For an application of web probing in the field of MI, see Meitinger (2017; see also section 7.2), and responses of the probing method could be processed to specify the prior on $\lambda_{pg} - \lambda_{pg'}$ following the aforementioned prior elicitation methods. However, as Davidov et al. (2015) rightly mentioned, no method to deal with noninvariance performs magic; there is a point at which one must conclude that MI simply does not hold (see, e.g., Lommen et al., 2013).

3.2. *Multiple group alignment*

The alignment methodology was introduced in Asparouhov and Muthén (2014) for continuous indicators and in Muthén and Asparouhov (2014) for categorical indicators. This method aims to compare latent variables across groups without requiring exact MI. The differences in the observed indicators across groups are primarily attributed to differences in the latent variables. Remaining differences that cannot be explained by latent variable differences are interpreted as evidence of partial noninvariance. The alignment method automates this process within a single stage estimation. The model fit is the same as the model fit of the configural model (see section 2), that is, the fit of the alignment model is as good as or better than any other MI model. The alignment method attempts to minimize the amount of noninvariance without altering the fit of the model.

Alignment utilizes the exploratory factor analysis (EFA) methodology in the following sense. In EFA, an unrotated model is estimated as a first step, which determines the best fitting variance covariance matrix for the observed variables given a fixed number of factors. The unrotated model can be rotated with an infinite number of rotations without altering the model fit. This provides an indeterminacy in the model, that is, there is no information in the data that can illuminate the best possible rotation for the factors. This indeterminacy is resolved by specifying a rotation criterion. The role of the rotation criterion is to eliminate the indeterminacy by quantifying our preference for simple loading structures. These are the loading structures where each observed variable loads primarily on one factor only and the number and size of cross-loadings is minimized. Alignment uses the same logic. The configural model plays the role of the unrotated solution, that is, this is the best fitting model given the number of factors and factor structure among all

models with all possible MI assumptions. The configural model can be reparameterized to include arbitrary values for the factor means and variances, without altering the model fit. The factor means and variances are unidentifiable. This indeterminacy is resolved by specifying an alignment criterion. The role of the alignment criterion is to eliminate the indeterminacy in the model by quantifying our preference for measurement invariant structures. The alignment method gives preference to as many invariant parameters and as few noninvariant parameters as possible. This parallelism between EFA and alignment can be very useful in understanding the alignment methodology.

Next, we provide a brief description of the alignment method. For illustration purposes, we use a multiple group factor analysis model with a single factor η measured by $p = 1, \ldots, P$ observed indicators in $g = 1, \ldots, G$ groups. Let y_{ipg} be the p-th observed indicator for individual i in group g. The factor model is given by the following equation:

$$y_{ipg} = \nu_{pg} + \lambda_{pg}\eta_{ig} + \varepsilon_{ipg}. \tag{10}$$

Analogous to Eq. (2), ν_{pg} and λ_{pg} are the intercept and loading parameters, $\varepsilon_{ipg} \sim N(0, \theta_{pg})$ is the residual variable, and $\eta_{ig} \sim N(\alpha_g, \psi_g)$ is the factor for individual i in group g. The alignment method estimates all of the parameters ν_{pg}, λ_{pg}, α_g, ψ_g, θ_{pg} as group-specific parameters. In particular, the method estimates group-specific factor mean and variance without assuming exact MI. The assumption of exact MI that allows us to estimate α_g and ψ_g is replaced by the assumption of approximate MI. That is, α_g and ψ_g are chosen as to minimize the overall differences between the measurement parameters. The overall measurement parameter differences are evaluated in a special way which gives preferences to models with large number of invariant parameters and as few noninvariant parameters as possible. Formal details are provided in Appendix A.

The alignment method can be viewed as an automation of the partial invariance method of Byrne et al. (1989). This method relies on estimating a sequence of nested models, where noninvariant parameters are selected in a step-wise process using modification indices. The alignment methodology automatically selects the noninvariant parameters within a single step estimation and it removes the somewhat subjective and time-consuming nature of the partial invariance method.

The alignment methodology works very well when most of the measurement parameters are invariant. The method will automatically separate the invariant and noninvariant parameters and all estimates will be consistent. It is somewhat difficult to quantify, however, the amount of noninvariance for which the alignment method will perform well. A rule of thumb is that as long as the number of noninvariant parameters is less than 20%, spread out across the groups, we can expect the alignment method to work correctly. However, the exact percentage of noninvariant parameters is not really what determines the alignment performance. Asparouhov and Muthén (2022), for example, show that the alignment method works well even in situations where half of all measurement parameters are noninvariant. The performance of the method is determined by the following question. Is the true parameter set, that is, the data generating parameter set, the simplest and most invariant representation of the data? Alignment will always pick α_g and ψ_g that produce the smallest amount of noninvariance. If a data-generating parameter set has a simpler alternative we cannot expect alignment to produce estimates consistent with the data-generating parameter set. Alignment will converge to the simpler alternative instead. If 100% of the data generating parameters are not invariant, we know that a simpler representation exists (at least one indicator can be made group invariant by adding factor means and variances) and so alignment will not recover the data generating parameters. On the other hand, in most situations when less than 20% of the data generating parameters are noninvariant, a simpler alternative will likely not exist and the alignment estimates will be consistent.

Consider as an example the situation where data is generated using intercept and loading parameters that are group-specific random effects, see Muthén and Asparouhov (2018). In this case, all parameters are noninvariant and the data-generating parameters will have a simpler (more invariant) alignment alternative. Alignment is not expected to recover the data-generating parameters. This phenomenon is exactly as in EFA. EFA produces simple loading structures. If the data-generating loadings include a large amount of cross-loadings, EFA will not recover the parameters and will produce a simpler loading structure instead. Alignment and EFA parameters are not biased in such situations. Both methods would produce more optimal (in terms of their optimization criterion) representation of the data than the data-generating parameters. Furthermore, because the intercepts and slopes are random, the alignment results are expected to show many noninvariant parameters, which probably will become challenging to interpret. The alternative methodology of estimating the model with random intercepts and loadings will in fact recover the data-generating parameters and will provide a simpler model interpretation. Thus, we conclude that the alignment methodology is not universally applicable for all MI studies. If the amount of noninvariance found with alignment is so large that model interpretation is challenging, alternative MI methodologies should be pursued (see sections 4 to 6).

The alignment methodology has been implemented since 2013 in Mplus version 7.1, Asparouhov and Muthén (2014). Subsequently, the SIRT package in R (Robitzsch, 2021) has also implemented the methodology. Several extensions have been added in Mplus since the original release. In Version 7.2, the alignment methodology was extended to binary items and in Version 7.3 to ordered polytomous items (Muthén and Asparouhov, 2014). There are generally 3 estimation methods that have been developed: maximum-likelihood (ML), Bayesian estimation, as well as BSEM based alignment. The BSEM based alignment starts with a BSEM measurement invariance estimation, instead of with the configural model, and is followed up with alignment. In the most recent release of Mplus, version 8.8, the alignment methodology is implemented also for the WLS estimators. The Mplus alignment language has also been simplified. Standard multiple-group specification can be used instead of knownclass mixture modeling specification. The alignment procedure implemented in Mplus prior to version 8.8 applies only to factor analysis models with multiple factors and simple loading structures, i.e., without cross-loadings. In Mplus 8.8, the alignment procedure is extended to factor models with complex loading structures, that is, models with cross-loadings and bi-factor models. Furthermore, in Mplus 8.8, the alignment procedure is extended to the general SEM model. Residual covariances among the factor indicators can be included in this model. Factor predictors

can also be included in the generalized alignment model, as well as direct effects from the predictors to the factor indicators. The generalized alignment imposes certain restrictions on the SEM model. For example, the factor regressions must have group-specific parameters. If such restrictions are not desired, an alternative two-stage estimation outlined in Marsh et al. (2018) can be used. Further discussion on the alignment procedure for the general SEM model is available in Asparouhov and Muthén (2022).

Multiple groups EFA and ESEM (Asparouhov and Muthén, 2009), can also be combined with alignment. To do that, the configural ESEM model is estimated first and then the rotated loading structure is aligned. This approach essentially parallels the aligned SEM estimation. The method can also be viewed as minimizing a joint simplicity function. The rotation simplicity function is added to the alignment loss function to form a joint simplicity function. The joint simplicity function is then minimized as a function of the factor means, the factor variances and the factor rotation. The key issue in the joint simplicity function is how to weigh the two components. The approach implemented in Mplus 8.8 essentially uses an infinitely large weight for the rotation part. This is to reflect the fact that we rotate the configural model without considering the alignment loss function. Conditional on these rotated results, the alignment is then conducted. It is in principle possible to combine rotation and alignment in a more equitable way. However, such an estimation will be more complex and it would still have the uncertainty about how to weigh the two components. Presumably, a more equitable rotation/alignment approach can have an MSE advantage in some situations. However, we can view the equitable rotation/alignment procedure as an alternative simplicity function, which is more or less going to yield similar results in most situations.

The alignment procedure can also be used with panel/longitudinal CFA models where MI does not necessarily hold across time. Such models cannot be formulated as multiple group models because the variables are correlated across time/groups. The forthcoming release of Mplus 8.9 will include the possibility to use the alignment method in panel CFA models.

The alignment procedure with categorical variables presents some additional challenges. One issue is related to the parameterization/metric of the model. Two different parameterizations are implemented in Mplus with the WLS estimator: the delta and theta parameterizations, see Muthén and Asparouhov (2002). In the theta parameterization the residual variances for the categorical items are fixed to 1 for identification purposes, while in the delta parameterization the total variances are fixed to 1. With the WLS estimator, a factor analysis model can be estimated in either one of these two metrics and the models will be equivalent. The ML and Bayes estimators currently can estimate the model only in the theta parameterization. In multiple group situations, MI in one metric does not translate into MI in the other metric. Because the goal of alignment is MI, the alignment results will depend on the metric. The model can be estimated in both metrics and perhaps the more invariant metric would be preferable. Some practical guidelines must be developed, however, in this regard.

The second important issue regarding alignment with categorical data is related to estimated residual variances in the theta parameterization and estimated delta parameters in the delta parameterization. Currently, for alignment models with categorical variables, these parameters remain fixed to 1. Conceivably, however, allowing the alignment function to be minimized with respect to these parameters as well, we may find an even better MI. In fact, MI models for categorical data without the alignment methodology typically involves estimating these parameters for all but the first group and is done by default in Mplus for the scalar invariance model. Our attempts to pursue this idea in the context of alignment, however, have fallen short so far. It appears that the sample size requirements make such an approach impractical and the additional parameters are rarely significantly different from 1, that is, categorical variables alignment with residual variances fixed to 1 is expected to be sufficient in most situations.

Extensive simulation studies on the alignment methodology can be found in Flake and McCoach (2018) and practical illustrations can be found in Munck et al. (2018) and Lomazzi (2018). A brief tutorial on the alignment method is provided in Rudnev (2019).

4. Measurement invariance with multilevel data

The evaluation of MI using MGCFA or MIMIC models (see section 2) may be complicated when a researcher is interested in MI over a large number of groups. For example, if a researcher wants to evaluate MI of a particular scale across 50 countries, a MGCFA would involve a comparison of measurement parameters across 50 groups, and the MIMIC approach would involve the evaluation of the effects of at least 49 dummy variables (see section 2.4). In these cases, it can be practical to treat group membership as random instead of fixed, and use multilevel SEM (MLSEM) to evaluate MI. For MLSEM, the number of groups (referred to as clusters in multilevel analysis) should be large enough to be able to obtain valid results at the cluster level. Although the exact number is dependent on factors such as the complexity of the model and the amount of variance present at the cluster level, minimum numbers of 50 (Maas and Hox, 2005) and 30 (Muthén and Asparouhov, 2018) have been suggested as rough guidelines.

We will discuss testing invariance within the random intercept framework and within the random intercept and slope framework, while limiting our presentation to two-level structures of individuals (Level 1 or the within level) in clusters (Level 2 or the between level).

4.1. Random intercept framework (within–between formulation)

MLSEM decomposes the observed variables into a within component and a between component (Muthén, 1989, 1994; Schmidt, 1969). Given the multivariate response vector $\boldsymbol{y}_{ig}$, with scores from subject i in cluster (group) g, the scores are decomposed into a vector of cluster means ($\boldsymbol{\mu}_g$), and a vector of individual deviations from the respective cluster means ($\boldsymbol{\eta}_{ig} = \boldsymbol{y}_{ig} - \boldsymbol{\mu}_g$):

$$\boldsymbol{y}_{ig} = \boldsymbol{\mu}_g + \boldsymbol{\eta}_{ig}, \tag{11}$$

where $\boldsymbol{\mu}_g$ and $\boldsymbol{\eta}_{ig}$ are independent. The overall covariances of $\boldsymbol{y}_{ig}$ (Σ_{TOTAL}) can be written as the sum of the covariances of these two

components:

$$\begin{aligned}\Sigma_{TOTAL} &= COV(\boldsymbol{\mu}_g, \boldsymbol{\mu}_g) + COV(\boldsymbol{\eta}_{ig}, \boldsymbol{\eta}_{ig}) \\ &= \Sigma_{BETWEEN} + \Sigma_{WITHIN}.\end{aligned} \tag{12}$$

This model specification is denoted the within/between formulation (Muthén, 1989, 1994). With two-level data, MI can be evaluated with respect to Level 1 variables (e.g., student gender in data from students in school classes), Level 2 variables (e.g., teacher gender), or the clustering variable (e.g., school classes). If MI across clusters holds, then observed mean differences between clusters reflect differences in the means of common factors across groups (Muthén, 1990; Rabe-Hesketh et al., 2004). Jak et al. (2013) refer to measurement bias with respect to the clustering variable as *cluster bias* and show how restrictions across groups in a multigroup model relate to restrictions across levels and on between-level residual variances in a two-level model. Specifically, weak factorial invariance across clusters (i.e., equal factor loadings across clusters in a multigroup model) implies equal factor loadings across levels in a two-level model. Strong factorial invariance across clusters (i.e., equal factor loadings and intercepts across clusters in a multigroup model), implies equal factor loadings across levels and zero residual variance at the between-level in a two-level factor model.

If cluster bias is found in one or more of the indicators, this indicates that there is measurement bias with respect to some (unmeasured) between-level variable. The amount of cluster bias in the indicator could be quantified by calculating the proportion of residual between-level variance of the total variance (see Jak, 2017). Sometimes researchers may have a measure of the between-level variable that possibly explains the cluster bias. Such a variable could then be included in the model as a covariate, regressing the common factor(s) and the indicator(s) with cluster bias on the covariate. This would actually represent a MIMIC model at the between level, and any significant direct effect of the covariate on an indicator would represent uniform measurement bias with respect to the covariate, similar to single level MIMIC models (Kim and Cao, 2015). Examples of this practice are included in Davidov et al. (2012), Jak et al. (2014), Jak (2017), Jang et al. (2017), or Seddig and Lomazzi (2019). If the variable that may explain the cluster bias is a grouping variable, then a researcher might fit a multigroup multilevel model (Asparouhov and Muthén, 2012; Muthén et al., 1997). If the grouping variable indeed explains the bias, then one would find zero between-level residual variances (i.e., no cluster bias) within the groups. Note that testing for cluster bias thus serves as an omnibus test for measurement bias with respect to any between-level variable. However, the between-level MIMIC model has more statistical power than the test for cluster bias (Jak and Oort, 2015).

Accounting for the measurement bias by freely estimating the residual variances at the between-level when needed already takes the measurement bias into account. This means that the variance in the between-level common factor will reflect "true" or purified differences across groups. If one includes a between-level covariate in a MIMIC model, then correctly specifying partial invariance by regressing (some) indicators on the covariate leads to an unbiased effect of the covariate on the latent variable, similar to single level models (Byrne et al., 1989; Guenole and Brown, 2014; Hsiao and Lai, 2018).

Testing measurement bias with respect to a Level 1 variable such as student gender in data from students in classrooms is more complicated. Splitting the data into two groups based on student gender would create dependent groups, because students who share a classroom will be in different groups based on their gender. Therefore, a simple multigroup model is not appropriate (see Asparouhov and Muthén, 2012; Ryu, 2014). Instead, one can apply multilevel factor mixture modeling or multilevel MIMIC modeling (Kim et al., 2015; Son and Hong, 2021).

4.2. Random intercepts and slopes

In the random intercept framework described above, the random intercepts refer to unconditional intercepts. The observed variables are decomposed into a within- and a between part before a factor model is fitted on the observed variables. An alternative approach to evaluating MI with many groups is to let the measurement intercepts and factor loadings themselves be random across clusters. Such an approach was first introduced in the IRT literature (de Jong et al., 2007; Fox, 2010; Verhagen and Fox, 2013), but can also be applied to continuous factor indicators (Asparouhov and Muthén, 2016). These models need Bayesian estimation, because maximum likelihood estimation would involve integration over more dimensions than is computationally feasible. The random intercept and random slope model is an extension of the model discussed in the previous section. If strong factorial invariance over clusters holds, then the model for the observed indicator p for individual i in cluster g is:

$$y_{ipg} = \nu_p + \lambda_p\left(\xi_{pg} + \xi_{ipg}\right) + \varepsilon_{ipg}, \tag{13}$$

where ξ_{pg} represents the mean factor score in cluster g and ξ_{ipg} and represents individual i's deviation from the cluster mean ξ_{pg}. Because the cluster means of the residual factors ε_{ipg} are assumed to be zero, and the intercepts and factor loadings are equal across clusters, the only source of variance at the between level is the common factor. This model translates to the two-level SEM model with equal factor loadings across levels and zero residual variance at the between level as discussed in the previous section.

The model in Eq. (13) can be extended by allowing the intercepts and factor loadings to vary across clusters. Instead of estimating one fixed value ν, one would then estimate the mean intercept across clusters and the variance across clusters: $\nu_g \sim N\left(\mu_\nu, \sigma_\nu^2\right)$. Similarly, one would estimate a mean factor loading and its variance across clusters: $\lambda_g \sim N\left(\mu_\lambda, \sigma_\lambda^2\right)$. Testing MI across clusters in this framework involves testing whether σ_ν^2 and/or σ_λ^2 are nonzero (Asparouhov and Muthén, 2016).[7] So, in the random intercepts and random loadings framework, measurement bias is operationalized as random intercepts and factor loadings with nonzero variances. To explain

[7] See Fox and Smink (2021) as well as Fox et al. (2020) for advancements of this approach that avoid testing variance parameters.

possible heterogeneity in intercepts or factor loadings across clusters, the random effects could be regressed on other cluster-level variables (Verhagen and Fox, 2013).

An advantage of this approach over the within/between model with random intercepts only, is that this model distinguishes nonuniform bias and uniform bias. In the within/between model, in contrast, differences in factor loadings and intercepts may both appear as residual variance at Level 2 (Jak et al., 2013). As a result, detecting nonzero residual variance at Level 2 provides no information about whether the source is a difference in intercepts or in factor loadings.

5. Measurement invariance in subsets of groups

Recent methodological advancements addressed the situation in which the exact invariance of all parameters across all groups is not supported, including partial invariance (section 2.3), the Bayesian approximate invariance approach (section 3.1), the alignment optimization (section 3.2), or their combinations. The alignment approach assumes that the majority of measurement parameters is invariant, whereas the minority does not conform to it. Likewise, the Bayesian approximate approach assumes that the measurement parameters (factor loadings and item intercepts) are normally distributed across groups, which implies, again, that the majority of groups are closer to a given level of invariance. Crudely, both methods assume two categories of groups—invariant and noninvariant, by each parameter. This assumption does not always hold, in particular, when the groups are not independent but rather nested within subsets of groups. This is often the case in comparative research, where countries can be nested within regions, religions, or cultural zones (e.g., Johnson et al., 2019). In this case, both the alignment and the Bayesian approximate invariance approaches would identify the largest subset of countries as the one possessing invariance and discard the others as outliers. The clustering problem then pertains to finding (unknown) subsets of groups for which MI holds *within* subsets, but not *between* subsets.

The clustering problem has been addressed early on by Rensvold and Cheung (2000), who suggested using similarity of factor loadings to find subsets. However, their approach was neither fully implemented nor widely used. The multilevel SEM described above could only address the clustering problem when the clustering is known/observed, in which case it can be entered as a categorical group-level predictor. More recently, two novel approaches were suggested that allow for finding subsets of groups in which MI holds, namely mixture multigroup factor analysis (De Roover et al., 2022) and Measurement Invariance Explorer (MIE; Rudnev, 2018–2022). The current section discusses these approaches.

5.1. Mixture multigroup factor analysis

When MI is evaluated across many groups (e.g., countries in cross-national surveys), it often does not hold (especially scalar invariance; Boer et al., 2018). When a certain level of invariance is rejected, resorting to pairwise comparisons of group-specific parameters in an attempt to find subsets of groups for which invariance holds is problematic. The many pairwise comparisons make it hard to unravel invariant from noninvariant parameters and for which groups they apply (Byrne and van de Vijver, 2010) and fosters the false detection of noninvariance (Rutkowski and Svetina, 2014).

In case of many groups, it is likely that "latent classes" of groups emerge with respect to the measurement parameters (Byrne and van de Vijver, 2010). Recently, mixture multigroup factor analysis (MMG-FA) was proposed (De Roover, 2021; De Roover et al., 2022), which combines multigroup factor analysis and mixture modeling (McLachlan and Peel, 2000) to capture subsets of groups with equivalent measurement parameters.

As described in section 2.2, different levels of measurement (non)invariance have different implications in terms of which comparisons are (in)valid (Meredith, 1993). Therefore, MMG-FA classifies groups such that a user-specified level of invariance is obtained within each subset. It can be used to scrutinize group subsets and the noninvariances between them in a level-by-level manner, or it can directly aim for the required level of invariance for a particular research question. In particular, MMG-FA allows researchers to classify groups based on their loadings, intercepts, or residual variances. Classifying groups on the following combinations is also enabled: on loadings and intercepts, on intercepts and residual variances, or on loadings, intercepts, and residual variances. For the classification to focus on these parameters specifically, they are made "class-specific", implying that they are shared by groups in the same class and differ between classes. Measurement parameters pertaining to a lower level of invariance than the class-specific parameters are set to be invariant across all groups (e.g., the loadings when classifying on intercepts and residual variances), and measurement parameters pertaining to a higher level as well as structural parameters (i.e., factor (co)variances and factor means) are allowed to vary freely across groups. It is worth noting that MMG-FA only captures differences and similarities in measurement parameters, whereas related group-level mixture models—most notably, multilevel factor mixture modeling (Kim et al., 2017)—would classify groups based on the measurement and structural parameters (see De Roover, 2021). MMG-FA can also be used with exploratory factor analysis (EFA) to capture all kinds of loading differences (i.e., including configural noninvariance).

Generally, the MMG-FA model for the data $\mathbf{Y}_g$ of group g is written as follows:

$$f\left(\mathbf{Y}_g;\boldsymbol{\theta}\right)=\sum_{k=1}^{K}\pi_k\prod_{i=1}^{n_g}MVN\left(\boldsymbol{y}_{ig};\boldsymbol{\mu}_{gk},\boldsymbol{\Sigma}_{gk}\right)\,, \tag{14}$$

where the total population density function f is modeled as a mixture of K multivariate normal distributions, where $\boldsymbol{\theta}$ refers to all parameters and π_k denotes the mixing proportion (i.e., the prior probability of a group belonging to class k, with $\sum_{k=1}^{K}\pi_k = 1$) and where, for example, $\boldsymbol{\Sigma}_{gk}=\boldsymbol{\Lambda}_k\boldsymbol{\Psi}_{gk}\boldsymbol{\Lambda}_k^{'}+\boldsymbol{\Theta}_g$ and $\boldsymbol{\mu}_{gk}=\boldsymbol{\nu}_k+\boldsymbol{\Lambda}_k\boldsymbol{\alpha}_{gk}$ to classify groups based on loadings $\boldsymbol{\Lambda}_k$ and intercepts $\boldsymbol{\nu}_k$ (but not on

residual variances and structural parameters). It is worth noting that the factor (co)variances $\boldsymbol{\Psi}_{gk}$ and means $\boldsymbol{\alpha}_{gk}$ depend on the group and the class it is assigned to. The factor (co)variances are group- and class-specific because $\boldsymbol{\Lambda}_k$ can differ strongly across classes and, in case of EFA, rotational freedom is present per class. Similarly, assuming that a group's factor means take on the same values in all classes would be too restrictive when they depend on class-specific intercepts. To classify the groups on the intercepts only, the loadings are made invariant across all groups—also replacing $\boldsymbol{\Psi}_{gk}$ by $\boldsymbol{\Psi}_g$. To classify on residual variances, they are specified as class-specific (i.e., $\boldsymbol{\Theta}_k$) rather than group-specific—also replacing $\boldsymbol{\nu}_k$ by $\boldsymbol{\nu}$ (and $\boldsymbol{\alpha}_{gk}$ by $\boldsymbol{\alpha}_g$) when intending to classify purely on residual variances. For details on model identification, the interested reader is referred to De Roover (2021) and De Roover et al. (2022). It is of note that the identification of MMG-FA does not depend on the choice of a "marker variable" or "referent item".

Due to the parameters being group- and/or class-specific, obtaining the parameter estimates $\widehat{\boldsymbol{\theta}}$ by Newton-Raphson, Fisher scoring, or Quasi-Newton optimization methods—that is, methods used in commercial software such as Latent GOLD (Vermunt and Magidson, 2013, 2016) and Mplus—is very slow and more sensitive to starting values. To find the parameter estimates in a time-efficient and stable manner, an expectation–maximization algorithm was developed (De Roover, 2021; De Roover et al., 2022), which is available in Latent GOLD 6.0 (as the "emfa" option) and in the *mixmgfa* R package (De Roover, 2022).

As is the case for any mixture model, the results of MMG-FA depend on the choice of the number of classes. It is important to guide users in making this choice, because estimating too few classes means that invariance is not guaranteed within each subset and estimating too many classes means that invariance may even exist across (some) classes. Extensive simulation studies established that a combination of the BIC (Schwarz, 1978; using the number of groups as the sample size) and the CHull method (Ceulemans and Kiers, 2006) mostly indicates the correct number of classes (De Roover, 2021; De Roover et al., 2022). These methods are integrated in the *mixmgfa* package. Additionally, one can apply multigroup factor analysis to each class of groups (or the "clustered invariance" model mentioned in section 5.2) to confirm whether invariance holds within the classes. Based on these results, the number of classes can be increased, if needed. Of course, it may also be that the groups within a class are identical with respect to most—but not all—measurement parameters (i.e., partial invariance; section 2.3) or that approximate invariance exists (section 3.1).

Thus, MMG-FA can be used to generate subsets of groups for which a specific level of invariance holds and for which valid comparisons or further invariance testing can be performed. Note that the classification of the groups into the subsets is probabilistic, but groups are commonly assigned to their classes with a probability of (around) one. To explain the noninvariance between subsets, group-level covariates can be related to the classification with the three-step (Vermunt, 2010) or two-step approach (Bakk and Kuha, 2018).

MMG-FA also ties down the number of parameters to inspect when searching for a selection of items (or partial invariance set up) that can be used to compare *all* groups. For this, modification indices (Sörbom, 1989) and item-deletion strategies (Byrne and van de Vijver, 2010; De Roover et al., 2014, 2017; Gvaladze et al., 2020) can be used. To optimally compare measurement parameters between subsets and test for significant differences, the alignment method (section 3.2) can be applied to the group-subsets (rather than the individual groups). When using EFA, multigroup factor rotation (De Roover and Vermunt, 2019) is also relevant to optimally compare class-specific loadings and perform hypothesis testing.

Though it is convenient that MMG-FA allows us to go from merely configural invariance across all groups (or not even that, when using EFA) to any level of invariance within classes of groups in one analysis, it is important to consider that classifying groups on more than one set of parameters—say, on loadings and intercepts—implies the assumption that one classification is underlying all of them. However, it may be that the differences in intercepts are explained by another classification than the loading differences—possibly with more classes—or they may even be group-specific. In that case, MMG-FA would need many classes for capturing the loading and intercept differences at the same time. It is then better to apply MMG-FA in steps; that is, to first find subsets of groups based on the loadings and then to further classify the groups per subset based on the intercepts.

Currently, MMG-FA is limited to maximum likelihood estimation, which assumes continuous items and multivariate normality whereas many questionnaire items are Likert scale items and nonnormality is often present. However, having five or more response categories allows for the items to be analyzed as continuous (in case of nonsevere nonnormality; Dolan, 1994). The effects of nonnormality are alleviated by the fact that multiple groups are gathered within classes, increasing the sample size within a class (Lei and Lomax, 2005). Another limitation is that an equal number of factors is assumed for all (subsets of) groups. An extension of MMG-FA with a class-specific number of factors is certainly interesting, but it requires tackling an extensive model selection problem (i.e., selecting not only the number of classes, but also the number of factors for each class).

In the future, MMG-FA will also be extended to build on partial invariance across all groups when looking for subsets of groups (e.g., allowing for loadings to be partially group-specific, when classifying on intercepts). Accommodating partial invariance within subsets (e.g., classifying groups on most, but not all, intercepts) is also an interesting venture.

5.2. Measurement invariance explorer

A simpler way to find the subsets of groups is to quantify information on the target level of MI across each pair of groups. Such information can be used to infer connections between groups in terms of their MI, which, in turn, makes it possible to find clusters, outliers, and visualize these links in a network-type graph. This strategy was implemented in the R package Measurement Invariance Explorer (MIE; Rudnev, 2018–2022), which provides a procedure to detect the subsets of groups that hold the target level of MI.

The MIE approach is applicable when the configural invariance has been supported across all groups, but some of the higher levels of invariance were rejected. It includes two key steps. In the first step, the degree of measurement noninvariance between each possible pair of groups is quantified. In the second step, this information can be used to visualize the distances between groups by means of

multidimensional scaling or a network analysis, and to find the subsets of groups that hold a target level of MI.

Measurement noninvariance across groups can be quantified in various ways. The most straightforward way, building on the idea by Rensvold and Cheung (2000), involves parameters estimated in a MGCFA with supported lower level of invariance, for instance, factor loadings from the configural invariance model. In this case, the differences in factor loadings can be aggregated into distances between groups. One downside of such a measure is that it involves point estimates of the parameters (e.g., factor loadings) and ignores the standard errors. An alternative measure, the *dMACS* effect size (Nye and Drasgow, 2011) represents the expected standardized differences in observed item scores due to noninvariance. Because of their standardized nature, the *dMACS* effects can be safely compared across groups and items, thus they can be transformed into the aggregate distances between countries. This measure is limited in its dependency on the reference items, which are initially assumed to be invariant across groups, as well as in its applicability to one factor at a time. Since Nye et al. (2019) provided cutoff values for *dMACS*, the distances can be used for the network visualization.

MI tests can be computed for every possible pair of groups, and the differences in fit measures (e.g., CFI difference between configural and metric invariance models) can be used as an MI-based distance between groups. The drawback of this measure is that most fit indices are probably not distributed normally, so the distances between groups can be biased. The advantage of this approach is that differences in fit indices have cutoff values (Chen, 2007), which can be used to visualize the confirmed invariance between groups as a network (see Fig. 4B). Such representation directly shows the results of the invariance tests. It is more informative compared to the conventional omnibus invariance tests. Moreover, such approach accounts for the heterogeneity of invariance solutions. The latter occurs when, despite a global support of MI across many groups, some pairs of these groups reject MI.

An optional step of MIE is to identify subsets of groups on the graph. The classification of groups can be obtained by using some ad hoc clustering of the groups, such as the partitioning around medoids method (Reynolds et al., 2006) based on one of the quantifications of invariance described above. The number of clusters could be detected by a parallel analysis and the "gap" statistic (Tibshirani et al., 2001).

To plot the distances, MIE calculates the distances based on measures of noninvariance, which are then passed to a multidimensional scaling, and finally visualized as a scatterplot with points representing groups (see Fig. 4A for an example). When the cutoff values are available for a measure of distance, the network plots become available – here, the edges represent statistically supported MI between groups (see Fig. 4B).

As a final step, one can run a single "stratified invariance model" in which the measurement parameters are constrained to equality within subsets of groups but not between. Such a model should satisfy the common conditions of the invariant model. For instance, the stratified metric invariance model should not be much worse than the configural model in terms of the model fit decrease—akin to the Chen (2007) criteria. If these criteria are satisfied, the clustering described above was successful. It implies that the comparisons of groups are possible within each cluster, but not between clusters. The stratified invariance model can also be used by itself to test hypotheses about invariance within known or assumed clusters.

An interesting application of the stratified invariance model may be of use with longitudinal cross-national surveys. In this case, the invariance can be tested within every country (but not between) for several countries simultaneously. If such a pattern of invariance holds, one can estimate country-specific change, whereas configural invariance across countries would still allow generalizing an overall direction of change.

The MIE approach is altogether a promising way to detect the patterns of invariance across groups. It can involve continuous or categorical indicators, employ various measures of noninvariance, look for subsets of groups using different parameters (e.g., both loadings and intercepts), and apply different methods of visualization of the distances and clusters of groups. Both MIE and MMG-FA can be applied whenever the groups are not independently sampled or when there are some apparent relations between groups which may, in theory, affect the test of invariance.

Since MIE is a young approach, its extensive evaluation is yet to be conducted. In particular, simulations are needed to compare different quantifications of noninvariance, to detect appropriate clustering techniques, as well as to provide specific cutoff values of the fit decrement in the stratified invariance model. Besides, the MIE approach is a heuristic rather than a formal model (such as MMG-FA from section 5.1) and intended to provide researchers with insights into the possible subsets of groups rather than complete solutions. The MIE approach does not provide information on the item-level noninvariance, as it is focused on the entire model's invariance. This limitation makes MIE especially useful with a small number of indicators, where an elimination of items or relaxation of the invariance constraints is not feasible (e.g., models with three indicators).

Currently, MIE accommodates the exact and stratified invariance models only. Further development can combine the MIE approach with Bayesian approximate invariance models, which could make the invariance tests even more flexible. Finally, the MIE approach can be expanded to include more general quantifications of the degree of noninvariance, such as Bayesian regions (Zhang et al., 2022) and scale-level effects (Meade, 2010).

6. Response shift theory and the decomposition approach

Many of the previously described procedures emanate from the idea that at least some invariance assumptions hold. In contrast, the following approach focuses explicitly on longitudinal noninvariance, providing a theoretical explanation for its presence and the analytical framework to decompose shifts in the measurement parameters and *true* change in the latent means across time. The theoretical fundament is *response shift theory* (RST; e.g., Rapkin and Schwartz, 2004, 2019; Schwartz and Sprangers, 1999; Sprangers and Schwartz, 1999) which has been developed in the field of quality of life research to explain conceptual changes in some components of a measurement model as the result of response shifts in intraindividual cross-time comparisons. The term *response shift*

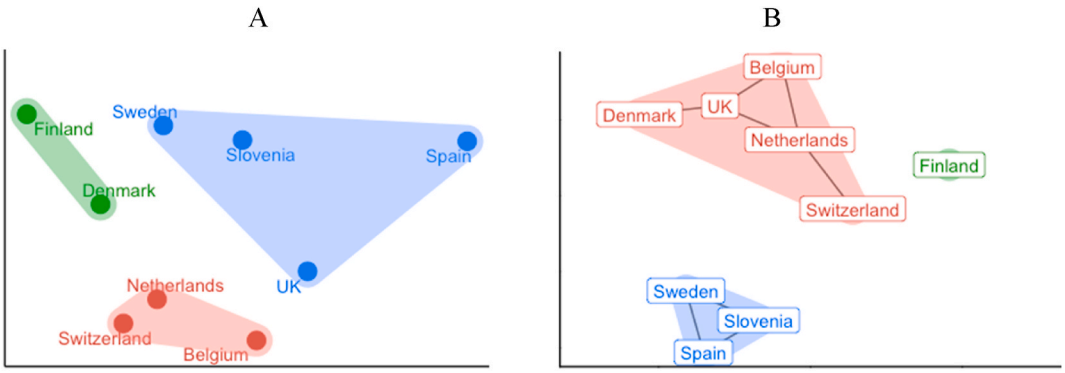


Fig. 4. Representation of a degree of measurement noninvariance as a distance between groups.[8]

"refers to a change in the meaning of one's self-evaluation of a target construct" (Schwartz and Sprangers, 1999, p. 1532), and thus, "is about individual cognitive change" (Rapkin and Schwartz, 2019, p. 2623).

In general, response shifts are theorized to be the outcome of three different mental processes: (i) *scale recalibration* as the change in a respondent's internal standards of measurement, (ii) *reprioritization* as the change in a respondent's values, and (iii) *reconceptualization* as the "big bang" change (Golembiewski et al., 1976, p. 138), representing a respondent's fundamental redefinition of the target construct. Leitgöb et al. (2021) employed the repeated measurement of the construct of "job satisfaction" to illustrate these processes. In this case, scale recalibration could be the result of an exorbitant wage increase (e.g., due to an unexpected job change) to a level outside the initially perceived range of a respondent's internal scale on the dimension "wage satisfaction", because she or he did not expect to ever earn that much. Then, recalibration refers to the cognitive process of updating the internal scale to a level that adequately reflects the new wage situation. Reprioritization is expected to occur as the consequence of a change in a respondent's value orientations from materialistic to post-materialistic over the life course. The importance the respondent attaches to the dimension "wage satisfaction" for overall job satisfaction will then decrease in favor of nonmonetary dimensions (e.g., reconciliation of work and family life, having enough leisure time, positive working atmosphere) with age. Reconceptualization will take place if the dimension "reconciliation of work and family life" is irrelevant for a childless respondent's job satisfaction. Entering parenthood, however, will result in a radical change in the respondent's internal definition of the concept of job satisfaction, with time for the fulfillment of care obligations then becoming one of the most relevant factors.

Oort (2005) established the conceptual link between RST and the different levels of MI (see Table 1) within the CFA panel modeling framework.[9] Configural invariance, as the least restrictive MI level, assumes that the underlying factor structure is equivalent across time. Technically, the identical pattern of zero and non-zero factor loadings is expected to exist as a stable representation of the construct(s) by the respective indicators ($\text{Patt}(\mathbf{\Lambda}_t) = \text{Patt}(\mathbf{\Lambda}_{t'})$, with $t < t'$). Violations of configural invariance across time result from respondents' internal processes of concept reconceptualization (Oort, 2005).[10] Metric invariance assumes the factor loadings of identical indicators to be equivalent across time. Varying factor loadings suggest that an indicator differs in its relevance of defining the underlying construct (Newsom 2015) and reflect reprioritization processes (Oort, 2005). Scalar invariance places additional cross-time equality constraints on the intercepts of identical indicators. It is violated if the same level of ability or degree of an attitude as indicated by identical latent scores at time points t and t' ($\eta_t = \eta_{t'} = \eta = 0$) is associated with systematically differing observable scores on some indicator p representing η ($\Delta y_p = y_{pt'} - y_{pt} \neq 0$). As argued by Oort (2005), this shift in scaling is the consequence of respondents' internal recalibration. He differentiates between two types of recalibration processes: (i) *Uniform* recalibration captures the respondents' change in the interpretation of the response options "in the same direction and to the same extent" (Oort, 2005, p. 592). It implies a constant perception of the measurement scale with regard to metric and range, but with a shift to the left or right.

[8] The figure is based on the European Social Survey data, round 7, a single-factor model used seven items from the Center for Epidemiological Studies Depression Scale (Radloff, 1977). Left panel shows the distances based on differences in the intercepts. Right panel shows distances based on ΔCFI between metric and scalar models, the lines represent ΔCFI <0.01.

[9] The CFA panel models differs from the MGCFA model as outlined in section 2 in a few aspects: First, the group indicator g is replaced by the time indicator $t = 1, \ldots, T$, to emphasize that the underlying data are no longer structured along G groups of independently drawn (random) samples. Rather, the data represent the T repeated measurements of the latent construct (via the set of manifest indicators) for the same, initially drawn (random) sample of individuals—organized in wide format. Second, to consider autocorrelations of errors terms from identical indicators across time, the respective off-diagonal elements of $\mathbf{\Theta}$ are typically estimated from the data, while they are fixed to zero by design in the cross-sectional multiple group case. The same holds true for the autocorrelative factor structure at the latent level. For further details, see Leitgöb et al. (2021), Little (2013) and Newsom (2015).

[10] Switching the perspective from the respondent level to the respondent population level is owed to the fact that the measurement parameters of the underlying measurement model refer to the latter. This implies that a sufficiently large number of individuals from the respondent population have to experience some response shift in the same direction to observe it as systematic change in the respective parameters between the time points of measurement. For a brief discussion of detecting individual change based on aggregate parameters, see Oort (2005).

Thus, uniform recalibration affects only the mean structure of the indicators and invalidates the assumption of scalar invariance. (ii) *Nonuniform* recalibration refers to the process of perceiving the measurement scale as distorted (either stretched or shrunk) between the different time points of measurement. Under such conditions, the mean and the covariance structure of the indicators could be affected. Regarding the latter, it is the difference in error variances of identical indicators across time that is indicative of the occurrence of nonuniform recalibration processes (Oort, 2005).

To allow at least some attenuated conclusions to be drawn about the amount and direction of *true change* in some construct under study from repeated measurements affected by systematic response shifts—implying measurement *non*invariance—, Oort (2005) formalized a frequently applied (Sajobi et al., 2018) threefold linear decomposition model based on the principles of Blinder-Oaxaca (BO) decomposition (Blinder, 1973; Oaxaca, 1973; Fortin et al., 2011). It decomposes the observed mean differences in each of the $p = 1, \ldots, P$ indicators between arbitrary time points t and t' into one term reflecting true change (actually occurring change in latent means) and the two other terms capturing the effects of (uniform) recalibration and reprioritization. By doing so, the model extracts information about true change in the underlying latent construct from the observed cross-time mean differences in the indicators in the presence of any form of measurement noninvariance:

$$\underbrace{\Delta\mu_p}_{\substack{\text{observed}\\ \text{change}}} = \underbrace{\Delta\nu_p}_{\text{recalibaration}} + \underbrace{\Delta\lambda_p\alpha_{t'}}_{\text{reprioritization}} + \underbrace{\Delta\alpha\lambda_{pt}}_{\textit{true}\text{ change}}, \tag{15}$$

with $\Delta\nu_p = \nu_{pt'} - \nu_{pt}$ as the intercept change capturing recalibration, $\Delta\lambda_p = \lambda_{pt'} - \lambda_{pt}$ as the change in factor loadings representing reprioritization, and $\Delta\alpha = \alpha_{t'} - \alpha_t$ as the true change in latent means. A formal derivation of the model is provided in Appendix B.

For statistical inference, Leitgöb et al. (2021) derived standard errors for the decomposition components based on the delta method. To assess the components' relevance, Oort (2005) proposed the computation of Cohen's d as an absolute effect size measure for each component. Additionally, Leitgöb et al. (2021) advocate—in line with the econometric tradition of BO decomposition—determining the relative contributions of recalibration, reprioritization, and true change to the observed change in p across time by calculating the respective proportions.

Recently, Leitgöb and Seddig (forthcoming) adopted the decomposition approach to the repeated cross-sectional data case by treating the repeated samples across time as independent groups and specifying the MGCFA instead of the CFA panel model as the baseline model for the decomposition. In contrast to the four-step approach of model identification proposed by Oort (2005), which relies on the fixed factor method in combination with MI-induced equality constraints on the measurement parameters, Leitgöb and Seddig (forthcoming) recommend specifying only the configural model (all parameters are freely estimated) and refer to the reference indicator method to determine the scale of the latent variables.[11] As formally demonstrated by Leitgöb and Seddig (forthcoming), this model specification will result in parameter estimates as input for the decomposition model that ensure an unbiased decomposition solution for the remaining $P - 1$ indicators, if the selected reference indicator is indeed invariant.

Furthermore, Leitgöb and Seddig (forthcoming) extended the threefold RSTC decomposition model by decomposing the reprioritization effect $\Delta\lambda_p\alpha_{t'}$ into (i) a *pure reprioritization* component and an (ii) *interaction* component.[12] For the panel case, the resulting fourfold model can be written as

$$\underbrace{\Delta\mu_p}_{\substack{\text{observed}\\ \text{change}}} = \underbrace{\Delta\nu_p}_{\text{recalibration}} + \underbrace{\underbrace{\Delta\lambda_p\alpha_t}_{\substack{\text{pure}\\ \text{reprioritization}}} + \underbrace{\Delta\lambda_p\Delta\alpha}_{\text{interaction}}}_{\text{reprioritization}} + \underbrace{\lambda_{pt}\Delta\alpha}_{\textit{true}\text{ change}}. \tag{16}$$

While the pure reprioritization effect captures the amount of mean change in p caused exclusively by a shift in the loading parameter across time, the interaction effect represents the impact of the coincidence of reprioritization and true change on the mean of p. However, the term provides only information about the joint occurrence of $\Delta\alpha$ and $\Delta\lambda_p$ and does *not* give rise to any causal interpretation (e.g., that $\Delta\lambda_p$ occurred as a consequence of $\Delta\alpha$).

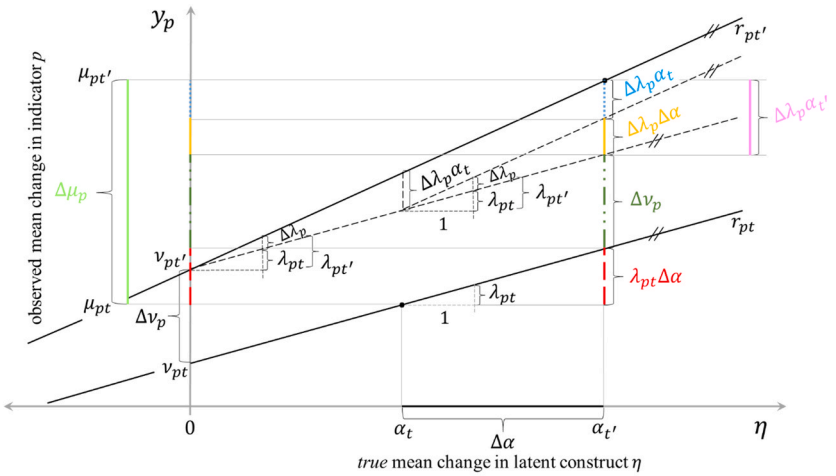
For the purpose of graphic illustration, Fig. 5 provides the geometrical derivation of both the three- and fourfold RSTC decomposition models.

To sum up, the objectives associated with the decomposition approach are twofold: First, it advances the theoretically informed investigation of longitudinal measurement noninvariance. Second, it still allows drawing some inferences about true change even if a high degree of noninvariance exists, given that the measurement model is correctly specified, the reference indicator is time invariant, and represents the underlying latent construct extremely well.

Finally, we outline several limitations and desiderata of the RSTC decomposition approach to define core areas of the future research agenda. First, because it is based on the pairwise comparison of time points, the approach is not well suited for the full

[11] The reference factor method is based on the idea of fixing the scale of a latent variable by constraining its distributional parameters, typically to $\alpha = 0$ and $\psi = 1$. In contrast, the reference indicator method follows the strategy of constraining the measurement parameters of some predetermined indicator r. In case of $\nu_r = 0$ and $\lambda_r = 1$, the latent variable adopts the scale of the reference indicator. For further details see, e.g., Brown (2015) and Kline (2016).

[12] Note that the fourfold decomposition is possible only if α_t is not fixed to zero. Otherwise, the reprioritization components from Eq. (16) reduce to $\Delta\lambda_p\alpha_t = 0$ and $\Delta\lambda_p\Delta\alpha = \Delta\lambda_p\alpha_{t'}$, leading to the threefold model from footnote 17.



Legend: $\Delta\mu_p$ … observed mean change in indicator p, $\lambda_{pt}\Delta\alpha$ … true change effect, $\Delta\nu_p$ … recalibration effect, $\Delta\lambda_p\alpha_{t'}$ … total reprioritization effect, $\Delta\lambda_p\alpha_t$ … pure reprioritization effect, $\Delta\lambda_p\Delta\alpha$ … interaction effect

Fig. 5. Geometrical derivation of the three- and fourfold RSTC decomposition models.

investigation of (very) many time points. Efficient multiple comparison strategies and the handling of their consequences, such as alpha error accumulation, can be inspired either by the respective discussion in the field of quality of life research (e.g., Barclay--Goddard et al., 2009; Sébille et al., 2021; Verdam and Oort, 2014, 2019) or by the general econometric literature on linear decomposition methods. Second, the approach focuses exclusively on decomposing the changes in the first moments of the distributions of the indicators. An extension to the second moments appears reasonable to separate the true change in the latent variance across time from reprioritization (changes in factor loadings) and nonuniform recalibration (changes in error variances) effects. Third, while RST is an individual-level theory that reflects intraindividual changes in response behavior, response shifts are identified at the aggregate level by evaluating changes in the measurement parameters. This has implications for the validity of inferences in the case of interindividual heterogeneity in response shifts (Sébille et al., 2021) and systematic panel mortality, calling for more complex classes of models and decomposition solutions. Fourth, it is obvious to adopt the approach for the cross-sectional multiple group case. A first step in this direction is the formal implementation for the repeated cross-sectional case based on the MGCFA model (Leitgöb & Seddig, forthcoming). However, at least two challenges remain unsolved. Theoretically, an adaptation of RST from intraindividual response shifts to interindividual cross-group differences in response behavior is required. Technically, the "unstructuredness" of the independent groups—in contrast to the temporal ordering of time points in the longitudinal case—complicates the identification of a unique decomposition solution, because it is not invariant to the choice of the reference group (index number problem; e.g., Fortin et al., 2011). Fifth, the approach is not yet implemented in the standard software packages.

7. The contribution of survey methodology to construct invariant measurement instruments

In the previous sections we have elaborated on the statistical approaches (i) to assess the empirical level of MI obtained by some measurement instrument(s), (ii) to compensate for identified parameter noninvariance, and (iii) to investigate its causes. It appears indispensable, however, to additionally cover the underemphasized but extremely relevant findings of survey methodological research on the development of invariant items and instruments. In fact, for comparative surveys of any form, the well-informed item and instrument construction represents the basic prerequisite for achieving a high level of measurement quality in general and MI in particular. Poor data quality caused by dysfunctional noninvariant instruments cannot be fully compensated by advanced statistical modeling. Put differently, the accuracy of inferences about *true* differences/changes in latent constructs drawn from estimation results based on data from actual invariant measurements is hard to reach in cases of noninvariant measurements, even when the most elaborated statistical measurement models are applied. For this reason, we subsequently highlight the existing evidence on the potential effects of scale formats, survey mode, item wording, and item translation on MI. Furthermore, we discuss the application of pretesting procedures not only as diagnostic tools to identify potential noninvariance problems prior to the main data collection but also to learn how respondents from different groups may vary in their perception, understanding, processing, and interpretation of the item content.

7.1. Design decisions and measurement invariance

As discussed, the application of exact MI testing has provided disappointing results for many instruments and data, showing that

metric and particularly scalar invariance are rarely supported (e.g., Davidov et al., 2008a,b, 2014, 2018c; Lee et al., 2020; Wu et al., 2007; Zercher et al., 2015). Besides implementing less restrictive measurment models, researchers recognize that little is known about the methodological or psychological sources of lacking MI (Meitinger, 2017; Roberts et al., 2020). In the following, we summarize research findings that address design decisions at the data collection stage that are associated with significant violations of MI.

Several studies addressed the design of rating scales as source of comparability problems. A rating scale consists of response options, ordered along a continuum, such as agreement, application, satisfaction, importance, or other evaluation dimensions (e.g., ranging from strongly disagree to strongly agree). As raised earlier by Krosnick and Fabrigar (1997) and later by other researchers (e.g., Alwin, 2007; Krebs, 2012; Menold 2020a; Menold and Raykov, 2016; Saris and Gallhofer, 2007; Weng, 2004), the choice of the rating scale format influences the reliability of measurement, which also implies that the variance covariance structures are affected. Consequently utilizing a different number of rating scale points for the same instrument led to noninvariance, as variations of 5, 7, or 11 scale points were associated with rejections of metric and scalar MI (Menold and Kemper, 2015; Menold and Tausch, 2016; Menold and Toepoel, 2022). In a similar fashion, varying the degree of rating scale verbalization or use of numbers to mark rating scale points violated MI (Menold and Kemper, 2015; Menold and Tausch, 2016). However, if the variations of rating scales were rather minor, scalar MI could be supported. This was shown for different verbalizations of the middle category (Höhne and Krebs, 2021), different verbalizations of extreme categories (Roberts et al., 2020), and incremental versus decremental presentation order (Höhne and Krebs, 2018).

As another potential source of noninvariance, the *mode of survey administration* has been addressed. It refers to the method and medium of data collection, that is, face-to-face and telephone as interviewer-administered survey modes or paper-and-pencil and web as self-administered survey modes (for an overview, see Hox et al., 2015; Roberts et al., 2020). The comparability of measurement instruments across different survey modes is of particular relevance when several modes are systematically combined within a survey, referred to as mixed-mode survey design (e.g., Heerwegh and Loosveldt, 2011; Hox et al., 2015; Klausch et al., 2013; Revilla, 2013; Roberts et al., 2020; Sakshaug et al., 2022). Mixed-mode surveys are typically implemented to reduce survey errors (Groves et al., 2009) and costs by referring to push-to-web methodology (e.g., Lynn, 2020). The assessment of (exact) MI serves as a tool to evaluate the potential of combining survey modes without generating artificial comparability shifts. In non-experimental studies, Revilla (2013) as well as Heerwegh and Loosveldt (2011) obtained scalar MI for modes with and without interviewers, whereas MI was found to be violated in experimental studies (e.g., Hox et al., 2015; Klausch et al., 2013). Remarkably, Klausch et al. (2013) found—as expected—configural, metric, and scalar invariance to hold *within* self-administered modes, while MI could not be achieved *between* self- and interviewer administered modes. *Within* interviewer-administered modes, scalar MI could be reached as well (Roberts et al., 2020). Comparing administration on different devices (e.g., tablets, smartphones, and desktop PCs) for the same (web) mode revealed scalar invariance (Menold and Toepoel, 2022; Höhne et al., 2021). Visual design in mixed-device surveys would impact MI between devices, however. Menold and Toepoel (2022) could demonstrate that the use of radio buttons did not affect MI, while MI could not be achieved for Visual Analog Scales.

Other potential sources of noninvariance, such as language effects (for cross- but also mono-language surveys), have been studied less frequently. Meitinger (2017) has shown that lack of scalar MI was associated with substantial differences in the understanding of concepts across countries (see also section 7.2). In randomized experiments, Menold (2020b) compared splits of double barreled questions by conventional omnibus invariance testing and the alignment method and found strong violations of metric and scalar MI.

To summarize, methodological artefacts, known to influence responses and data, have also been found to negatively influence measurement invariance. This is shown for the conception of rating scales, the visual design, survey mode, device and item wording.

7.2. Pretesting

When answering a survey question, respondents should comprehend the question's meaning, retrieve relevant information from their memory, form a judgment, and report their response (Tourangeau et al., 2000), and these cognitive processes should be possible and comparable for all cultural groups involved. Survey measures should be as equivalent as possible, but comparability might be reduced due to different biases. Biases are "nuisance factors that jeopardize the validity of instruments applied in different cultures" (He and van de Vijver, 2012: p. 3). For example, construct bias signifies that the measured construct varies across cultures or might not exist at all (van de Vijver and Poortinga, 1997). Specific methods and measurement contexts (e.g., differences in sampling procedures) can create a method bias (He and van de Vijver, 2012). Finally, item bias can appear for each survey question due to poor item translations, ambiguous source items, inapplicability of item contents or differences in associations of key terms across cultures (van de Vijver and Leung, 2021).

To achieve a high level of MI, it is necessary to understand the sources of bias and to proactively apply strategies that increase the equivalence of measures during the questionnaire development (Meitinger et al., 2020). Bias can be reduced if the questionnaire undergoes thorough pretesting before the data collection. Pretesting can potentially identify the existence and source of problems, provide insights on how language and culture influence the question response process, and reveal whether questions and response options are interpreted differently across cultural groups (Aizpurua, 2020). Different testing approaches can be applied such as expert assessments (e.g., content or methodological experts or the Survey Quality Predictor Tool (SQP; Zavala-Rojas et al., 2019)), lab or web-based approaches (e.g., focus groups, cognitive interviews, and web probing), and field-based approaches (e.g., experiments, behavior coding, vignettes) as well as statistical approaches to assess data quality (e.g., multitrait-multimethod experiments; de Jong et al., 2019). For an overview of different quantitative and qualitative pretesting approaches for cross-cultural research, see Aizpurua (2020), Caspar et al. (2016), and Smith (2019). We will focus in the following on two qualitative pretesting approaches that are particularly useful for assessing equivalence of measurement: cross-cultural cognitive interviewing and web probing (Behr et al.,

2017).

Cognitive interviewing "entails administering draft survey questions while collecting additional verbal information about survey responses, which is used to evaluate the quality of the response or to help determine whether the question is generating the information that its author intends" (Beatty and Willis, 2007, p. 287). Think aloud and verbal probing are the two dominant approaches in cognitive interviewing. With think aloud, the cognitive interviewer encourages the respondents to verbalize their thoughts while answering a question. Probes are questions designed to elicit additional relevant information about question functioning (Willis and Miller, 2011). For example, a category selection probe asks respondents for the reasons why they selected a certain answer category (e.g., "Why did you choose 'I completely agree'?") and a comprehension probe requests a definition of a key term of a question (e.g., "What does the term 'civil disobedience' mean to you?") (Willis, 2005; Willis and Miller, 2011). Probing can reveal silent misinterpretations, where respondents are unaware that they have misunderstood the item (Behr et al., 2017). Sample sizes are usually small (e.g., 5–15 respondents), but the questionnaire is often tested in iterative rounds (Willis, 2005). The interviewer motivates the respondents, allows for spontaneous follow-up questions, and an in-depth assessment of questions, which makes it particularly suitable for newly developed questions (Edgar et al., 2016; Meitinger and Behr, 2016).

When conducted with multiple cultural groups and/or languages, cross-cultural cognitive interviewing (CCCI) is a useful tool to determine "whether the range of interpretations associated with the evaluated items varies acceptably between cultural or language groups, given the survey measurement objectives" (Willis, 2015, p. 363). It reveals how respondents in different cultures and languages process and answer survey questions (Willis and Miller, 2011) and helps to evaluate the questionnaire translations (Harkness, 2003). Although sample sizes tend to be higher in CCCI (Willis, 2015), low case numbers per country (e.g., around 20 cases per country in (Fitzgerald et al., 2011)) do not allow for an assessment of the prevalence of identified issues, and conclusions on the differences between country-specific patterns might be difficult (Braun et al., 2019). In addition, setting up a CCCI study comes with practical (e.g., recruitment and training of interviewers) and harmonization challenges (e.g., comparability of results across countries) (Miller, 2019; Willis, 2015).

To address these challenges, web probing has been proposed as a complementary approach. Web probing "is the implementation of probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey questions" (Behr et al., 2017, p. 1). In an international context, web probing can assess comparability (Behr et al., 2020). During web probing, respondents receive on the first screen the target question and then receive, on a second screen, the probe. For example, Behr et al. (2014) assessed a potentially biased item of the "rights in a democracy scale" of the 2004 ISSP. The item asked about the importance of civil disobedience. In their web probing study, they evaluated this item with a category selection probe and a comprehension probe in Canada, Denmark, Germany, Hungary, the U.S., and Spain ($n = 3695$). Based on the probing results, they revealed that associations with the key term "civil disobedience" differed across countries. In Canada and the U.S., respondents associated violent actions with it which is not the case for the other countries in this study.

The implementation in web surveys facilitates the recruitment of respondents, speeds up the data collection (Neuert et al., 2021), and easily increases sample sizes when compared to cognitive interviewing (Meitinger and Behr, 2016). This allows for an evaluation of the prevalence of themes and can explain the response patterns of specific subpopulations. Since all respondents receive the same probe, the procedure is very standardized (Braun et al., 2015), which reduces the harmonization issues of CCCI (Meitinger, 2017). On the downside, web probing might not elicit sufficient information to explain the lack of comparability (Behr et al., 2020). There is no interviewer that could ask spontaneous follow-up questions to fully explore the respondents' understandings of a question. The absence of an interviewer also means that fewer questions can be tested than in CCCI (Meitinger and Behr, 2016). For a detailed discussion of strength and weaknesses of both approaches see Behr et al. (2017, 2020).

CCCI and web probing can also be combined with quantitative MI testing. For example, Meitinger (2017) assessed items measuring constructive patriotism and nationalism in the 2013 ISSP Module on National Identity with MI testing using MGCFA and with web probing. She compared five countries: Germany, Great Britain, Mexico, Spain, and the U.S. and found metric, but no scalar MI. The web probing results identified differences in associations with the item asking about the pride in the country's social security system. In Mexico, many respondents silently misunderstood the term as referring to the general security situation in their country (e.g., "there is a lot of violence"). In the remaining countries, the scope of perceived benefits varied with respondents in the U.S. mostly thinking about retirement benefits, Spaniards associating it with health benefits, and respondents in Germany and Great Britain linking it with many different social security benefits.

The combination of web probing or CCCI with quantitative MI tests circumvents the shortcomings of each approach. The latter allow for the testing of many countries but do not provide explanations for a lack of comparability. The qualitative approaches can reveal these reasons but are rather time and labor intensive because probe answers need to be translated, and coded which restricts its application to a few countries (Behr, 2015). Therefore, MI tests can also be used to identify countries and items for a follow-up web probing study to improve survey questions or develop new items that better match the respondents' reality for future data collection (Braun et al., 2015; Meitinger, 2017).

So far, CCCI and web probing have either been implemented as pretest before the data collection or as a follow-up study (Behr et al., 2017; Braun et al., 2015). However, web probing could also be implemented easily during the actual data collection if the main survey is a web survey. Following the random probe approach proposed by Schuman (1966), randomly selected respondents could receive probes for a subset of questions (Behr et al., 2017). With the establishment of cross-national online survey panels (e.g., CRONOS[13]), the

[13] For further information, see https://www.europeansocialsurvey.org/methodology/methodological_research/modes_of_data_collection/cronos.html.

extension of web probing to in-field assessment might be promising. Whereas CCCI already has been conducted in diverse linguistic and cultural contexts (e.g., Martin et al., 2017; Mneimneh et al., 2018), web probing mostly focused on Western countries. Future studies should apply this approach with a culturally diverse set of countries and especially include non-Western countries (Behr et al., 2020).

7.3. Cross-cultural scale translation and adaptation

A necessary, even though not sufficient precondition for invariant survey instruments, is the sound translation and/or adaptation of a given measurement instrument. In the following, we will be using the term *translation*, given its usage in cross-cultural survey methodology (Harkness et al., 2010b; Lyberg et al., 2021) and in large-scale studies such as the ESS, the ISSP, EVS, etc. In these studies, the source instruments are typically designed with cross-cultural implementation in mind so that cross-cultural relevance and translatability are considered early on (Dorer, 2020; Smith, 2004), paving the way for a more or less smooth translation.[14] The term translation, however, should not lead to taking translation lightly or to misunderstanding it as a mere automatic word replacement exercise (Lyberg et al., 2021). Translation is based on in-depth understanding and complex decision-making, taking into account the function of the measurement instrument, the target population, the mode of implementation, the co-text, and potentially further project specifications. Translation competence largely surpasses knowledge of two languages (Behr, 2018). The translation method chosen is considered to be crucial when it comes to ensuring translation quality. In cross-cultural survey methodology, the TRAPD method is regarded as best practice (Harkness, 2003; Harkness et al., 2010b)[15]: It combines a multi-step process with interdisciplinary collaboration. The translation step (T) in a prototypical TRAPD implementation[16] uses parallel translation whereby two translators independently from each other produce a translation of the instrument; this process ensures greater objectivity, offers stylistic variants, and helps to identify obvious or more subtle errors. In the review step (R), the translators and a reviewer jointly discuss and reconcile these versions, thereby taking decisions for each individual item as to whether any (or a combination) of the initial translations is suitable or whether a new translation should be produced from scratch. In the adjudication phase (A), an adjudicator, oftentimes the same person as the reviewer, takes final decisions and signs off the translation, before it goes into a pretest (P). Documentation (D) of particular decisions, difficulties, deviations, etc., as well as of the overall process (see, for details, Behr and Zabal, 2020) supplements the entire process. Lack of documentation regarding translation is deplored in the cross-cultural community (Rios and Sireci, 2014), hindering external researchers to independently assess the quality of a study. Furthermore, documentation of particular decisions or challenges may be linked to the outcome from invariance testing and may provide potential reasons for lack of invariance. The TRAPD method calls for interdisciplinary collaboration between translation professionals, substantive experts (on the constructs measured), and questionnaire design experts. The team review is the place to pool the expertise (Harkness, 2003; Harkness et al., 2010b, see also International Test Commission, 2016). Given that item bias can results from, for instance, poor translation or culture-specific connotations (van de Vijver and Leung, 2011), the review is the place to weigh different translation options, discuss different interpretations and perspectives, and to come to a solution that works best in a given context. The 'P' in TRAPD suggests, however, that despite its value diverse expertise cannot substitute for a (qualitative/quantitative) test among the target population, which is essential for ensuring validity of an instrument. Thus, Harkness et al. (2010a, b, p. 138) argue: "The obvious route to take is to make translations as good as one can using design and translation strategies and qualitative testing and then test whether statistical analysis verifies qualitative assessments or not or reveals new aspects not found in qualitative appraisal." Item bias can also be triggered by low item appropriateness or ambiguities in original items (van de Vijver and Leung, 2011; Byrne, 2016). This brings us back to the beginning of this section, namely that cross-national research, when planned right from the beginning, should set up appropriate procedures that ensure the development of an instrument that – at least theoretically – can be valid and reliable for the cultures concerned. The ensuing statistical tests will show whether this is indeed the case.

Besides developing one universal (etic) instrument for all, recent years have also witnessed a greater call for a combination of a universal and a culture-specific (emic) approach (van de Vijver, 2013; Behr and Zabal, 2019) to do justice to the heterogeneity of cultures. Culture-specific instruments contain items that are developed for only one culture in mind (examples provided by Cheung et al., 2011). The future is also calling for more empirical testing to better understand the effects of different translation options, deviations, and adaptations, on MI, as well as to better delineate between successful and less successful translation methods (Lyberg et al., 2021).

8. Discussion and future perspectives

This paper summarizes the current state of the art of statistical and (survey) methodological research on measurement invariance

[14] In cross-cultural psychology, the term test or scale *adaptation* is preferred to stress the importance of decisions on the linguistic, cultural, and psychometric level; to sensitize to various types of adaptations, i.e., intentional deviations from the source instrument, when transferring an instrument from one culture to another, and to highlight the importance of profound knowledge of the construct studied among those adapting an instrument (van de Vijver, 2013; van de Vijver and Leung, 2011; ITC, 2016).

[15] TRAPD was set up as a counter model to the back translation method popular since the 1970s, with TRAPD focusing on in-depth and joint assessment of the translation itself rather of a back translation. Discussions and empirical tests can be found in Behr (2017), Behr and Braun (2022), Bolaños-Medina and González-Ruiz (2012), Epstein et al. (2015), and Hagell et al. (2010).

[16] Lyberg et al. (2021) list variants of a TRAPD implementation, taking into account contextual constraints such as time, costs, and feasibility.

(MI), which is considered a core challenge for the comparative social sciences. As demonstrated, much progress has been achieved in the last 10 years, ranging from the specification of statistical models with milder assumptions than exact MI that still allow drawing valid conclusions about latent variables to survey methodological tools that support the development of invariant measurement instruments. Finally, what future developments can be expected in the subject area? From our point of view, at least four lines of development are recognizable, besides the ongoing advancements in the established approaches described above.

(i) *Conceptual and theoretical developments*: The recent debate in *Sociological Methods & Research* (Fischer et al., 2022; Meuleman et al., 2022; Welzel et al., 2021, 2022) on the relevance of MI in comparative research and how it is routinely assessed uncovers the still existing differences in the conceptual understanding of MI in the scientific community. This exchange of the diverging perspectives should be regarded as a constructive process of joining forces to further develop and more precisely specify the concepts of (non)comparability, (non)invariance, and item bias. Furthermore, the ubiquity of measurement noninvariance in empirical comparative research is expected to stimulate the theoretical reasoning about its causes. In this sense, noninvariance as an independent phenomenon of interest will probably receive more attention in the future.

(ii) *Analytical and statistical developments*: Boer et al. (2018) point out the potential of exploratory structural equation modeling (ESEM; e.g., Asparouhov and Muthén, 2009; Marsh et al., 2014; see also the related approach proposed by Dolan et al., 2009) for MI testing. ESEM represents the systematic integration of the (MG)CFA approach and exploratory factor analysis (EFA). As mentioned in section 3.1, it can also be combined with the alignment method. However, little application experience is available to date, making it difficult to assess the additional practical benefits of ESEM for MI testing. Further research is needed to evaluate its applicability (Boer et al., 2018). Another promising approach is making use of machine learning (ML) algorithms. This trend concerns the specification of SEM models in general (e.g., van Kesteren and Oberski, 2022) and the assessment of MI in particular. The latter aims, for example, at the detection of invariant reference indicators for model identification by means of lasso regularization methods (Belzak and Bauer, 2020) or the detection of noninvariance based on deep neural networks (Pokropek and Pokropek, 2022). Furthermore, Jankowsky et al. (2020), Olaru and Danner (2020), as well as Olaru et al. (2019) proposed utilizing the *ant colony optimization* (ACO) algorithm for item selection to develop short scales for comparative research by optimizing model fit and MI. Due to the increasing popularity of ML in the social sciences, it is expected to represent one of the future core areas in the field of MI assessment. However, this development will depend substantially on the extent to which substantive researchers are enabled to apply the procedures properly in academic training (e.g., Friedrich et al., 2021).

(iii) *Methodological developments*: Due to the continuing interest in the comparative study of social change and the increasing availability of large-scale cross-national studies in repeated cross-sectional format with repeated measurements at several points in time (e.g., ESS, ISSP, EVS), systematic strategies for the testing of combined cross group-cross time MI are expected to gain relevance (e.g., Koc and Pokropek, 2022). This holds also true for the panel case with the assessment of MI within (across time) and between individuals (e.g., Adolf et al., 2014). Furthermore, the elaboration of a tailored methodological framework for longitudinal MI testing is another desideratum. Although many of the basic methodological principles of cross-group invariance testing apply equally to the repeated cross-sectional case, this does not hold entirely true for the panel case with its, for example, time-related dependencies at the manifest and latent levels (e.g., Leitgöb et al., 2021; Seddig and Leitgöb, 2018). Contributions to longitudinal MI methodology may further help raising the level of awareness of the relevance of MI testing in substantive fields.

(iv) *Survey methodological developments*: Following the general *big data* trend in the social sciences, it appears obvious to train supervised ML models on the mass of existing comparative data to identify the specific features of items responsible for noninvariance across groups or time. This allows generating substantive evidence on the determinants of noninvariance from all available data. Such information would substantially support the tasks of item formulation, instrument construction, and design selection in comparative research.

To sum up, statistical and (survey) methodological research has come a long way in expanding and improving the tool kid for the assessment and dealing with (non)invariance. As illustrated, we expect this trend also to continue in future. This review intends to contribute to the literature by equipping applied researchers with a state of the art guidance through the various aspects of the topic to find tailored solutions and to raise further awareness for its relevance in comparative research.

Funding

This work was partly supported by the Dutch Research Council (NWO), project number VI.Vidi.201.009, awarded to Suzanne Jak.

Appendix A. The alignment model

The first step in the alignment method is the estimation of the configural model. In the configural model $\alpha_g = 0, \psi_g = 1$ for every g, and all loading, intercept, and residual variance parameters are estimated as group-specific parameters. Denote the configural model estimates by $\nu_{pg,0}$, $\lambda_{pg,0}$ and $\theta_{pg,0}$, and let the configural factor be $\eta_{ig,0}$. Because the aligned model has the same fit as the configural model, the following relationships must hold:

$$\eta_{ig} = \alpha_g + \sqrt{\psi_g}\eta_{ig,0}, \tag{A1}$$

$$V(y_{ipg}) = \lambda_{pg}^2 \psi_g + \theta_{pg} = \lambda_{pg,0}^2 + \theta_{pg,0}, \tag{A2}$$

$$E(y_{ipg}) = \nu_{pg} + \lambda_{pg}\alpha_g = \nu_{pg,0}, \tag{A3}$$

where $E(y_{ipg})$ and $V(y_{ipg})$ are the model estimated mean and variance for Y_{ipg}. Setting $\theta_{pg,0} = \theta_{pg}$, we get

$$\lambda_{pg} = \frac{\lambda_{pg,0}}{\sqrt{\psi_g}}, \tag{A4}$$

$$\nu_{pg} = \nu_{pg,0} - \alpha_g \frac{\lambda_{pg,0}}{\sqrt{\psi_g}}. \tag{A5}$$

The aligned model chooses α_g and ψ_g to minimize the amount of measurement noninvariance, that is, the differences in λ_{pg} and ν_{pg} across groups. To formalize this, we minimize the alignment function F, which accumulates all measurement noninvariance:

$$F = \sum_p \sum_{g<g'} \omega_{g,g'} f(\lambda_{pg} - \lambda_{pg'}) + \sum_p \sum_{g<g'} \omega_{g,g'} f(\nu_{pg} - \nu_{pg'}), \tag{A6}$$

where f is a component loss function and $\omega_{g,g'}$ are weights. The weights $\omega_{g,g'}$ are set to reflect the group size and the amount of certainty we have in the group estimates for a particular group. We use $\omega_{g,g'} = \sqrt{N_g N_{g'}}$. With these weights, larger groups will contribute more to the total loss function than smaller groups. The component loss function is set to

$$f(x) = \sqrt[4]{x^2 + \varepsilon}, \tag{A7}$$

where ε is a small number such as 0.0001. This function is approximately equal to $\sqrt{|x|}$. We use a positive ε so that F has a continuous first derivative, which makes the optimization easier and more stable. This choice of f, as compared to other choices such as x and x^2, has the advantage that it overemphasizes the penalty for medium-size losses/noninvariance and underemphasizes the penalty for larger losses/noninvariance. Thus, the optimal invariance losses are expected to be either close to zero (invariant parameters) or not zero (noninvariant parameters). The medium-range losses are meant to be eliminated with this choice of f. This is a key feature of the alignment methodology that distinguishes the method from other methods. BSEM measurement invariance (section 3.1) or multilevel models with random intercepts and slopes (section 4) tend to minimize mean squared error functions, which can lead to many parameters with medium-sized noninvariance. The alignment method typically will result in many approximately invariant measurement parameters, a few large noninvariant measurement parameters, and no medium-sized noninvariant measurement parameters. This is similar to the fact that EFA rotation functions aim for either large or small loadings, but not mid-sized loadings. Minimizing the loss function F will generally identify the parameters α_g and ψ_g in all but the first group. In the first group, these parameters remain fixed to 0 and 1, respectively.

The alignment estimation for models with complex loading structure is adjusted as follows. Eq. (A5) is replaced by

$$\nu_{pg} = \nu_{pg,0} - \sum_{m=1}^{M} \alpha_{mg} \frac{\lambda_{pmg,0}}{\sqrt{\psi_{mg}}}, \tag{A8}$$

where M is the number of factors, α_{mg} and ψ_{mg} are the m-th factor mean and variance in group g, while λ_{pmg} is the loading of the p-th indicator on the m-th factor.

Appendix B. Response shift-true change (RSTC) decomposition model

Leitgöb et al. (2021) provide a model derivation based on the principles of counterfactual reasoning (e.g., Morgan and Winship, 2015). Starting point is the scalar form of Eq. (3) for some arbitrary indicator p and a single latent construct η at time points t and t':

$$\mu_{pt} = \nu_{pt} + \lambda_{pt}\alpha_t \tag{A9}$$

$$\mu_{pt'} = \nu_{pt'} + \lambda_{pt'}\alpha_{t'}, \tag{A10}$$

with $\mu_{pt^{(\prime)}}$ as the observed mean of p at $t^{(\prime)}$, $\nu_{pt^{(\prime)}}$ and $\lambda_{pt^{(\prime)}}$ as the intercept and loading parameters, and $\alpha_{t^{(\prime)}}$ representing the mean of the latent variable $\eta_{t^{(\prime)}}$. Next, $\mu^c_{pt'}$, the counterfactual mean for p at t', given the *hypothetical* situation that longitudinal scalar invariance is satisfied, is introduced. The respective equality assumptions allow $\nu_{pt'}$ and $\lambda_{pt'}$ to be replaced by ν_{pt} and λ_{pt}, their initial values at t:

$$\mu^c_{pt'} = \nu_{pt} + \lambda_{pt}\alpha_{t'}. \tag{A11}$$

Then, the observed mean difference in p across time can be formalized as

$$\Delta\mu_p = \mu_{pt'} - \mu_{pt} = \underbrace{\mu_{pt'} - \mu^c_{pt'}}_{\text{response shift}} + \underbrace{\mu^c_{pt'} - \mu_{pt}}_{\textit{true}\text{ change}}. \tag{A12}$$

First, the response shift term $\mu_{pt'} - \mu^c_{pt'}$ represents the difference between the observed and the counterfactual mean of p at t'. Because the latent mean equals $\alpha_{t'}$ in both cases, the difference deviates from zero only if the scalar invariance assumption does not hold, that is, if the measurement parameters change across time. Thus, it captures the fraction of $\Delta\mu_p$ caused by recalibration ($\nu_{pt'} \neq \nu_{pt}$) and/or reprioritization ($\lambda_{pt'} \neq \lambda_{pt}$). Second, the true change term $\mu^c_{pt'} - \mu_{pt}$ reflects the difference between the counterfactual mean of p at t' and the observed mean at t. Because the set of measurement parameters is equal to $[\nu_{pt}, \lambda_{pt}]$ in both conditions, the term takes a nonzero value only in case of true change, that is, a shift in latent means between t and t' ($\alpha_{t'} \neq \alpha_t$). Inserting Eqs. (A9) to (A11) into Eq. (A12) and reorganizing terms leads finally to the threefold RSTC decomposition model:

$$\underbrace{\Delta\mu_p}_{\text{observed change}} = \underbrace{\Delta\nu_p}_{\text{recalibaration}} + \underbrace{\Delta\lambda_p\alpha_{t'}}_{\text{reprioritization}} + \underbrace{\Delta\alpha\lambda_{pt}}_{\textit{true}\text{ change}}, \tag{A13}$$

with $\Delta\nu_p = \nu_{pt'} - \nu_{pt}$, $\Delta\lambda_p = \lambda_{pt'} - \lambda_{pt}$, and $\Delta\alpha = \alpha_{t'} - \alpha_t$.

For identification purposes, Oort (2005) proposed to constrain $\alpha_t = 0$. Consequently, $\Delta\alpha = \alpha_{t'} - 0 = \alpha_{t'}$ changing the right-hand-side of Eq. (A13) to $\Delta\nu_p + \Delta\lambda_p\alpha_{t'} + \alpha_{t'}\lambda_{pt}$, which is identical to the formulation of Eq. (8) in Oort (2005, p. 594).

References

Adolf, J., Schuurman, N.K., Borkenau, P., Borsboom, D., Dolan, C.V., 2014. Measurement invariance within and between individuals: a distinct problem in testing the equivalence of intra- and inter-individual model structures. Front. Psychol. 5, 883.

Aizpurua, E., 2020. Pretesting methods in cross-cultural research. In: Sha, M., Gabel, T. (Eds.), The Essential Role of Language in Survey Research. RTI Press, Research Triangle Park, pp. 129–150.

Alwin, Duane F., 2007. Margins of Error: A Study of Reliability in Survey Measurement. Wiley, New York.

Ariely, G., Davidov, E., 2010. Can we rate public support for democracy in a comparable way? Cross-national equivalence of democratic attitudes in the world value survey. Soc. Indicat. Res. 104 (2), 271–286.

Arts, I., Fang, Q., Meitinger, K., van de Schoot, R., 2021. Approximate measurement invariance of willingness to sacrifice for the environment across 30 countries: the importance of prior distributions and their visualization. Front. Psychol. 2911.

Asparouhov, T., Muthén, B.O., 2009. Exploratory structural equation modeling. Struct. Equ. Model. 16 (3), 397–438.

Asparouhov, T., Muthén, B.O., 2012. *Multiple Group Multilevel Analysis* (Mplus Web Notes No. 16). Retrieved from. http://www.statmodel.com/examples/webnote.shtml.

Asparouhov, T., Muthén, B.O., 2014. Multiple-group factor analysis alignment. Struct. Equ. Model. 21 (4), 495–508.

Asparouhov, T., Muthén, B.O., 2016. General random effect latent variable modeling: random subjects, items, contexts, and parameters. In: Harring, J.R., Stapleton, L. M., Beretvas, S.N. (Eds.), Advances in Multilevel Modeling for Educational Research: Addressing Practical Issues Found in Real-World Applications. Information Age Publishing, Charlotte, pp. 163–192.

Asparouhov, T., Muthén, B.O., 2017. *Prior-posterior Predictive P-Values* (Mplus Web Notes No. 22). Retrieved from. http://www.statmodel.com/examples/webnote.shtml.

Asparouhov, T., Muthén, B.O., 2022. Multiple Group Alignment for Exploratory and Structural Equation Models. Retrieved from. https://www.statmodel.com/download/alignment.pdf.

Asparouhov, T., Muthén, B., Morin, A.J.S., 2015. Bayesian structural equation modeling with cross-loadings and residual covariances: comments on Stromeyer et al. J. Manag. 41 (6), 1561–1577.

Bakk, Z., Kuha, J., 2018. Two-step estimation of models between latent classes and external variables. Psychometrika 83 (4), 871–892.

Barclay-Goddard, R., Lix, L.M., Tate, R., Weinberg, L., Mayo, N.E., 2009. Response shift was identified over multiple occasions with a structural equation modeling framework. J. Clin. Epidemiol. 62 (11), 1181–1188.

Barendse, M.T., Oort, F.J., garst, G.J.A., 2010. Using restricted factor analysis with latent moderated structures to detect uniform and nonuniform measurement bias: a simulation study. Adv. Stat. Anal. 94 (2), 117–127.

Barendse, M.T., Oort, F.J., Werner, C.S., Ligtvoet, R., Schermelleh-Engel, K., 2012. Measurement bias detection through factor analysis. Struct. Equ. Model. 19 (4), 561–579.

Beatty, P.C., Willis, G.B., 2007. Research synthesis: the practice of cognitive interviewing. Publ. Opin. Q. 71 (2), 287–311.

Becker, C.C., Davidov, E., Cieciuch, J., Algesheimer, R., Kindschi, M., 2020. Measuring school children's attitudes toward immigrants in Switzerland and Poland. Meas. Instrum. Soc. Sci. 2, 9.

Behr, D., 2015. Translating answers to open-ended survey questions in cross-cultural research: a case study on the interplay between translation, coding, and analysis. Field Methods 27 (3), 284–299.

Behr, D., 2017. Assessing the use of back translation: the shortcomings of back translation as a quality testing method. Int. J. Soc. Res. Methodol. 20 (6), 573–584.

Behr, D., 2018. Translating questionnaires for cross-national surveys: a description of a genre and its particularities based on the ISO 17100 categorization of translator competences. Transl. Interpr. 10 (2), 5–20.

Behr, D., Braun, M., 2022. How does back translation fare against team translation? An experimental case study in the language combination English–German. J. Surv. Stat. Methodol. Advance online publication.

Behr, D., Braun, M., Kaczmirek, L., Bandilla, W., 2014. Item comparability in cross-national surveys: results from asking probing questions in cross-national web surveys about attitudes towards civil disobedience. Qual. Quantity 48 (1), 127–148.

Behr, D., Meitinger, K., Braun, M., Kaczmirek, L., 2017. *Web Probing: Implementing Probing Techniques from Cognitive Interviewing in Web Surveys with the Goal to Assess the Validity of Survey Questions* (GESIS–Survey Guidelines). Retrieved from. https://www.gesis.org/gesis-survey-guidelines/instruments/qualitaet-von-umfragedaten/web-probing.

Behr, D., Meitinger, K., Braun, M., Kaczmirek, L., 2020. Cross-national web probing: an overview of its methodology and its use in cross-national studies. In: Beatty, P. C., Collins, D., Kaye, L., Padilla, J.-L., Willis, G.B., Wilmot, A. (Eds.), Advances in Questionnaire Design, Development, Evaluation and Testing. Wiley, Hoboken, pp. 521–543.

Behr, D., Zabal, A., 2019. A meeting report: OECD-GESIS Seminar on translating and adapting instruments in large-scale assessments (2018). Meas. Instrum. Soc. Sci. 1, 10.

Behr, D., Zabal, A., 2020. *Documenting Survey Translation* (GESIS–Survey Guidelines). Retrieved from. https://www.gesis.org/en/gesis-survey-guidelines/open-science/documenting-survey-translation-1.
Belzak, W.C.M., Bauer, D.J., 2020. Improving the assessment of measurement invariance: using regularization to select anchor items and identify differential item functioning. Psychol. Methods 25 (6), 673–690.
Billiet, J., 2003. Cross-cultural equivalence with structural equation modeling. In: Harkness, J.A., van de Vijver, F.J.R., Mohler, P.P. (Eds.), Cross-cultural Survey Methods. Wiley, Hoboken, pp. 247–264.
Blinder, A.S., 1973. Wage discrimination: reduced form and structural estimates. J. Hum. Resour. 8 (4), 436–455.
Boer, D., Hanke, K., He, J., 2018. On detecting systematic measurement error in cross-cultural research: a review and critical reflection on equivalence and invariance tests. J. Cross Cult. Psychol. 49 (5), 713–734.
Bolaños-Medina, A., González-Ruiz, V., 2012. Deconstructing the translation of psychological tests. Meta: J. des Traducteurs/Translators' J. 57 (3), 715–739.
Bollen, K.A., 1989. Structural Equations with Latent Variables. Wiley, New York.
Bollen, K.A., 2002. Latent variables in psychology and the social sciences. Annu. Rev. Psychol. 53 (1), 605–634.
Borsboom, D., 2005. Measuring the Mind: Conceptual Issues in Contemporary Psychometrics. Cambridge University Press, Cambridge.
Braun, M., Behr, D., Kaczmirek, L., Bandilla, W., 2015. Evaluating cross-national item equivalence with probing questions in web surveys. In: Engel, U., Jann, B., Lynn, P., Scherpenzeel, A., Sturgis, P. (Eds.), Improving Survey Methods. Lessons from Recent Research. Routledge, New York, pp. 184–199.
Braun, M., Behr, D., Meitinger, K., Raiber, K., Repke, L., 2019. Using web probing to elucidate respondents' understanding of minorities in cross-cultural comparative research. Ask: Res. Methods 28 (1), 3–20.
Braun, M., Johnson, T.P., 2010. An illustrative review of techniques for detecting inequivalences. In: Harkness, J.A., Braun, M., Edwards, B., Johnson, T.P., Lyberg, L., Mohler, P.P., Smith, T.W. (Eds.), Survey Methods in Multinational, Multiregional, and Multicultural Contexts. Wiley-Blackwell, Hoboken, pp. 375–393.
Brown, T.A., 2015. Confirmatory Factor Analysis for Applied Research, second ed. Guilford Press, New York.
Byrne, B.M., 2004. Testing for multigroup invariance using AMOS graphics: a road less traveled. Struct. Equ. Model. 11 (2), 272–300.
Byrne, B.M., 2016. Adaptation of assessment scales in cross-national research: issues, guidelines, and caveats. Int. Perspect. Psychol.: Res. Pract. Consult. 5 (1), 51–65.
Byrne, B.M., Shavelson, R.J., Muthén, B.O., 1989. Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. Psychol. Bull. 105 (3), 456–466.
Byrne, B.M., Stewart, S.M., 2006. The MACS approach to testing for multigroup invariance of a second-order structure: a walk through the process. Struct. Equ. Model. 13 (2), 287–321.
Byrne, B.M., van de Vijver, F.J.R., 2010. Testing for measurement and structural equivalence in large-scale cross-cultural studies: addressing the issue of nonequivalence. Int. J. Test. 10 (2), 107–132.
Caspar, R., Peytcheva, E., Yan, T., Lee, S., Liu, M., Hu, M., 2016. *Pretesting* (Cross-Cultural Survey Guidelines). Retrieved from. https://ccsg.isr.umich.edu/chapters/pretesting.
Ceulemans, E., Kiers, H.A., 2006. Selecting among three-mode principal component models of different types and complexities: a numerical convex hull based method. Br. J. Math. Stat. Psychol. 59 (1), 133–150.
Chen, F.F., 2007. Sensitivity of goodness of fit indexes to lack of measurement invariance. Struct. Equ. Model. 14 (3), 464–504.
Chen, F.F., 2008. What happens if we compare chopsticks with forks? The impact of making inappropriate comparison in cross-cultural research. J. Pers. Soc. Psychol. 95 (5), 1005–1018.
Chen, Y., Li, C., Xu, G., 2021. DIF Statistical Inference and Detection without Knowing Anchoring Items. Retrieved from. https://arxiv.org/abs/2110.11112.
Chen, F.F., Sousa, K.H., West, S.G., 2005. Testing measurement invariance of second-order factor models. Struct. Equ. Model. 12 (3), 471–492.
Cheung, G.W., Rensvold, R.B., 2002. Evaluating goodness-of-fit indexes for testing measurement invariance. Struct. Equ. Model. 9 (2), 233–255.
Cheung, F.M., van de Vijver, F.J.R., Leong, F.T.L., 2011. Toward a new approach to the study of personality in culture. Am. Psychol. 66 (7), 593–603.
Chun, S., Stark, S., Kim, E.S., Chernyshenko, O.S., 2016. MIMIC methods for detecting DIF among multiple groups: exploring a new sequential-free baseline procedure. Appl. Psychol. Meas. 40 (7), 486–499.
Cieciuch, J., Davidov, E., 2016. Establishing measurement invariance across online and offline samples: a tutorial with the software packages Amos and Mplus. Stud. Psychol. 15 (2), 83–99.
Cieciuch, J., Davidov, E., Algesheimer, R., Schmidt, P., 2018. Testing for approximate measurement invariance of human values in the European Social Survey. Socio. Methods Res. 47 (4), 665–686.
Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., 2019. How to obtain comparable measures for cross-national comparisons. Kölner Z. Soziol. Sozialpsychol. 71 (Suppl. 1), 157–186.
Cieciuch, J., Davidov, E., Vecchione, M., Beierlein, C., Schwartz, S.H., 2014. The cross-national invariance properties of a new scale to measure 19 basic human values: a test across eight countries. J. Cross Cult. Psychol. 45 (5), 764–779.
Coromina, L., Davidov, E., 2013. Evaluating measurement invariance for social and political trust in western Europe over four measurement time points (2002-2008). ASK Res. Methods 22 (1), 37–54.
Davidov, E., 2009. Measurement equivalence of nationalism and constructive patriotism in the ISSP: 34 countries in a comparative perspective. Polit. Anal. 17 (1), 64–82.
Davidov, E., Braun, M., 2012. What do citizens expect from a democracy? An invariance test and comparison between East and West Germany with the ISSP 2004. In: Salzborn, S., Davidov, E., Reinecke, J. (Eds.), Methods, Theories, and Empirical Applications in the Social Sciences. Festschrift for Peter Schmidt. Springer VS, Wiesbaden, pp. 213–219.
Davidov, E., Cieciuch, J., Meuleman, B., Schmidt, P., Algesheimer, R., Hausherr, M., 2015. The comparability of measurements of attitudes toward immigration in the European social survey: exact versus approximate measurement equivalence. Publ. Opin. Q. 79 (1), 244–266.
Davidov, E., Cieciuch, J., Schmidt, P., 2018c. The cross-country measurement comparability in the immigration module of the European social survey 2014–15. Surv. Res. Methods 12 (1), 15–27.
Davidov, E., De Beuckelaer, A., 2010. How harmful are survey translations? A test with Schwartz's human values instrument. Int. J. Publ. Opin. Res. 22 (4), 485–510.
Davidov, E., Depner, F., 2011. Testing for measurement equivalence of human values across online and paper-and-pencil surveys. Qual. Quantity 45 (2), 375–390.
Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., Meulemann, B., 2012. Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. J. Cross Cult. Psychol. 43 (4), 558–575.
Davidov, E., Meuleman, B., Billiet, J., Schmidt, P., 2008a. Values and support for immigration: a cross-country comparison. Eur. Socio Rev. 24 (5), 583–599.
Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., Billiet, J., 2014. Measurement equivalence in cross-national research. Annu. Rev. Sociol. 40, 55–75.
Davidov, E., Muthén, B.O., Schmidt, P., 2018a. Measurement invariance in cross-national studies: challenging traditional approaches and evaluating new ones. Socio. Methods Res. 47 (4), 631–636.
Davidov, E., Schmidt, P., Billiet, J., Meuleman, B., 2018b. Cross-cultural Analysis: Methods and Applications, second ed. Routledge, New York.
Davidov, E., Schmidt, P., Schwartz, S.H., 2008b. Bringing values back in: the adequacy of the European Social Survey to measure values in 20 countries. Publ. Opin. Q. 72 (3), 420–445.
Davidov, E., Siegers, P., 2010. Comparing basic human values in East and West Germany. In: Beckers, T., Birkelbach, K., Hagenah, J., Rosar, U. (Eds.), Komparative Empirische Sozialforschung. Springer VS, Wiesbaden, pp. 43–63.
de Ayala, R.J., 2022. The Theory and Practice of Item Response Theory, second ed. Guilford Press, New York.
De Beuckelaer, A., Swinnen, G., 2018. Biased latent variable mean comparisons due to measurement noninvariance: a simulation study. In: Davidov, E., Schmidt, P., Billiet, P., Meuleman, B. (Eds.), Cross-cultural Research: Methods and Applications, second ed. Routledge, New York, pp. 127–156.
de Jong, J.A.J., Dorer, B., Lee, S., Yan, T., Villar, A., 2019. Overview of questionnaire design and testing. In: Johnson, T.P., Pennell, B.-E., Stoop, I.A.L., Hoboken, B. Dorer (Eds.), Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC). Wiley, New York, pp. 115–133.

de Jong, M.G., Steenkamp, J.B.E., Fox, J.P., 2007. Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. J. Consum. Res. 34 (2), 260–278.
Depaoli, S., 2021. Bayesian Structural Equation Modeling. Guilford Press, New York.
Depaoli, S., van de Schoot, R., 2017. Improving transparency and replication in Bayesian statistics: the WAMBS-Checklist. Psychol. Methods 22 (2), 240–261.
De Roover, K., 2021. Finding clusters of groups with measurement invariance: unraveling intercept non-invariance with mixture multigroup factor analysis. Struct. Equ. Model. 28 (5), 663–683.
De Roover, K., 2022. Mixmgfa. Retrieved from. https://github.com/KimDeRoover/mixmgfa.
De Roover, K., Timmerman, M.E., Ceulemans, E., 2017. How to detect which variables are causing differences in component structure among different groups. Behav. Res. Methods 49 (1), 216–229.
De Roover, K., Timmerman, M.E., De Leersnyder, J., Mesquita, B., Ceulemans, E., 2014. What's hampering measurement invariance: detecting non-invariant items using clusterwise simultaneous component analysis. Front. Psychol. 5, 604.
De Roover, K., Vermunt, J.K., 2019. On the exploratory road to unraveling factor loading non-invariance: a new multigroup rotation approach. Struct. Equ. Model. 26 (6), 905–923.
De Roover, K., Vermunt, J.K., Ceulemans, E., 2022. Mixture multigroup factor analysis for unraveling factor loading noninvariance across many groups. Psychological Methods 27 (3), 281–306.
Dolan, C.V., 1994. Factor analysis of variables with 2, 3, 5, and 7 response categories: a comparison of categorical variable estimators using simulated data. Br. J. Math. Stat. Psychol. 47 (2), 309–326.
Dolan, C.V., Oort, F.J., Stoel, R.D., Wicherts, J.M., 2009. Testing measurement invariance in the target rotated multigroup exploratory factor model. Struct. Equ. Model. 16 (2), 295–314.
Dorer, B., 2020. Advance Translation as a Means of Improving Source Questionnaire Translatability? Findings from a Think-Aloud Study for French and German. Frank & Timme, Berlin.
Durkheim, É, 1982. In: Lukes, S. (Ed.), The Rules of Sociological Method. The Free Press, New York.
Edgar, J., Murphy, J., Keating, M., 2016. Comparing traditional and crowdsourcing methods for pretesting survey questions. Sage Open 6 (4), 1–14.
Epstein, J., Osborne, R.H., Elsworth, G.R., Beaton, D.E., Guillemin, F., 2015. Cross-cultural adaptation of the health education impact questionnaire: experimental study showed expert committee, not back-translation, added value. J. Clin. Epidemiol. 68 (4), 360–369.
Fischer, R., Karl, J.A., Fontaine, J.R.J., Poortinga, Y.H., 2022. Evidence of validity does *not* rule out systematic bias: a commentary on nomological noise and cross-cultural invariance. Sociaol. Methods Res. Adv. online publication.
Fitzgerald, R., Widdop, S., Gray, M., Collins, D., 2011. Identifying sources of error in cross-national questionnaires: application of an error source typology to cognitive interview data. J. Off. Stat. 27 (4), 569–599.
Flake, J.K., McCoach, D.B., 2018. An investigation of the alignment method with polytomous indicators under conditions of partial measurement invariance. Struct. Equ. Model. 25 (1), 56–70.
Fortin, N., Lemieux, T., Firpo, S., 2011. Decomposition methods in econometrics. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 4. North Holland, Amsterdam, pp. 1–102. *Part A*.
Fox, J.-P., 2010. Bayesian Item Response Modeling: Theory and Applications. Springer, New York.
Fox, J.-P., Koops, J., Feskens, R., Beinhauer, L., 2020. Bayesian covariance structure modelling for measurement invariance testing. Behaviormetrika 47 (2), 385–410.
Fox, J.-P., Smink, W.A.C., 2021. Assessing an alternative for "negative variance components": a gentle introduction to Bayesian covariance structure modeling for negative associations among patients with personalized treatments. Psychol. Methods. Advance online publication.
Freitag, M., Bauer, P.C., 2013. Testing for measurement equivalence in surveys: dimensions of social trust across cultural contexts. Publ. Opin. Q. 77 (S1), 24–44.
Friedrich, S., Antes, G., Behr, S., Binder, H., Brannath, W., Dumpert, F., Lederer, J., Leitgöb, H., Ickstadt, K., Kestler, H., Pauly, M., Steland, A., Wilhelm, A., Friede, T., 2021. Is there a role for statistics in artificial intelligence? Adv. Data Anal. Classif. Advance online publication.
Garthwaite, P.H., Al-Awadhi, S.A., Elfadaly, F.G., Jenkinson, D.J., 2013. Prior distribution elicitation for generalized linear and piecewise-linear models. J. Appl. Stat. 40 (1), 59–75.
Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2004. Bayesian Data Analysis, second ed. Chapman & Hall CRC, London.
Golembiewski, R.T., Billingsley, K., Yeager, S., 1976. Measuring change and persistence in human affairs: types of change generated by OD designs. J. Appl. Behav. Sci. 12 (2), 133–157.
Gordoni, G., Schmidt, P., Gordoni, Y., 2012. Measurement invariance across face-to-face and telephone modes: the case of minority-status collectivistic-oriented groups. Int. J. Publ. Opin. Res. 24 (2), 185–207.
Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., Tourangeau, R. (Eds.), 2009. Survey Methodology, second ed. Wiley, New York.
Guenole, N., Brown, A., 2014. The consequences of ignoring measurement invariance for path coefficients in structural equation models. Front. Psychol. 5, 980.
Guttman, L., 1945. A basis for analyzing test–retest reliability. Psychometrika 10 (4), 255–282.
Gvaladze, S., De Roover, K., Tuerlinckx, F., Ceulemans, E., 2020. Detecting which variables alter component interpretation across multiple groups: a resampling-based method. Behav. Res. Methods 52 (1), 236–263.
Hagell, P., Hedin, P.J., Meads, D.M., Nyberg, L., McKenna, S.P., 2010. Effects of method of translation of patient reported health outcome questionnaires: a randomized study of the translation of the rheumatoid arthritis quality of life (Raqol) instrument for Sweden. Value Health 13 (4), 424–430.
Harkness, J.A., 2003. Questionnaire translation. In: Harkness, J.A., van de Vijver, F.J.R., Mohler, P.P. (Eds.), Cross-cultural Survey Methods. Wiley, Hoboken, pp. 35–56.
Harkness, J.A., Braun, M., Edwards, B., Johnson, T.P., Lyberg, L., Mohler, P.P., Pennell, B.-E., Smith, T.W. (Eds.), 2010a. Survey Methods in Multinational, Multiregional, and Multicultural Contexts. Wiley, Hoboken.
Harkness, J.A., Villar, A., Edwards, B., 2010b. Translation, adaptation, and design. In: Harkness, J.A., Braun, M., Edwards, B., Johnson, T.P., Lyberg, L., Mohler, P.P., Pennell, B.-E., Smith, T.W. (Eds.), Survey Methods in Multinational, Multiregional, and Multicultural Contexts. Wiley, Hoboken, pp. 117–140.
Heerwegh, D., Loosveldt, G., 2011. Assessing mode effects in a national crime victimization survey using structural equation models: social desirability bias and acquiescence. J. Off. Stat. 27 (1), 49–63.
He, J., van de Vijver, F., 2012. Bias and equivalence in cross-cultural research. Online Readings in Psychology & Culture 2 (2), 1–19.
Heyder, A., Schmidt, P., 2003. Authoritarianism and ethnocentrism in east and west Germany: does the system matter? In: Alba, R., Schmidt, P., Wasmer, M. (Eds.), Germans or Foreigners? Attitudes toward Ethnic Minorities in Post-reunification Germany. Palgrave, St. Martin's Press, New York, pp. 97–104.
Hildebrandt, A., Wilhelm, O., Robitzsch, A., 2009. Complementary and competing factor analytic approaches for the investigation of measurement invariance. Rev. Psychol. 16 (2), 87–102.
Höhne, J.K., Krebs, D., 2018. Scale direction effects in agree/disagree and item-specific questions: a comparison of question formats. Int. J. Soc. Res. Methodol. 21 (1), 91–103.
Höhne, J.K., Krebs, D., 2021. Mismatching middle options: consequences for attitude measurement in smartphone surveys. Int. J. Soc. Res. Methodol. 24 (3), 381–386.
Höhne, J.K., Krebs, D., Kühnel, S.-M., 2021. Measurement properties of completely and end labeled unipolar and bipolar scales in Likert-type questions on income (in) equality. Soc. Sci. Res. 97, 102544.
Hoijtink, H., van de Schoot, R., 2018. Testing small variance priors using prior-posterior predictive p values. Psychol. Methods 23 (3), 561–569.
Horn, J.L., McArdle, J.J., 1992. A practical and theoretical guide to measurement invariance in aging research. Exp. Aging Res. 18 (3–4), 117–144.
Hox, J.J., De Leeuw, E.D., Zijlmans, E.A.O., 2015. Measurement equivalence in mixed mode surveys. Front. Psychol. 6, 87.
Hsiao, Y.-Y., Lai, M.H.C., 2018. The impact of partial measurement invariance on testing moderation for single and multi-level data. Front. Psychol. 9, 740.
Hu, L., Bentler, P.M., 1999. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct. Equ. Model. 6 (1), 1–55.
Hui, C.H., Triandis, H.C., 1985. Measurement in cross-cultural psychology: a review and comparison of strategies. J. Cross Cult. Psychol. 16 (2), 131–152.

International Test Commission (ITC), 2016. ITC Guidelines for Translating and Adapting Tests, second ed. Retrieved from. http://www.psyssa.com/wp-content/uploads/2015/11/ITC-Guidelines-Translating-and-Adapting-Tests-v2-3.pdf.
Jak, S., 2017. Testing and explaining differences in common and residual factors across many countries. J. Cross Cult. Psychol. 48 (1), 75–92.
Jak, S., Oort, F.J., 2015. On the power of the test for cluster bias. Br. J. Math. Stat. Psychol. 68 (3), 434–455.
Jak, S., Oort, F.J., Dolan, C.V., 2013. A test for cluster bias: detecting violations of measurement invariance across clusters in multilevel data. Struct. Equ. Model. 20 (2), 265–282.
Jak, S., Oort, F.J., Dolan, C.V., 2014. Measurement bias in multilevel data. Struct. Equ. Model. 21 (1), 31–39.
Jang, S., Kim, E.S., Cao, C., Allen, T.D., Cooper, C.L., Lapierre, L.M., et al., 2017. Measurement invariance of the satisfaction with life scale across 26 countries. J. Cross Cult. Psychol. 48 (4), 560–576.
Jankowsky, K., Olaru, G., Schroeders, U., 2020. Compiling measurement invariant short scales in cross-cultural personality assessment using ant colony optimization. Eur. J. Pers. 34 (3), 470–485.
Jin, K.-Y., Chen, H.-F., 2020. MIMIC approach to assessing differential item functioning with control of extreme response style. Behav. Res. Methods 52 (1), 23–35.
Johnson, T.P., 1998. Approaches to equivalence in cross-cultural and cross-national survey research. In: Harkness, J.A. (Ed.), *Cross-cultural Survey Equivalence* (Zuma-Nachrichten Spezial 3). Zuma, Mannheim, pp. 1–40.
Johnson, T.P., Pennell, B.-E., Stoop, I.A.L., Dorer, B. (Eds.), 2019. *Advances in Comparative Survey Methods: Multinational, Multiregional and Multicultural Contexts* (3MC). Wiley, Hoboken.
Jöreskog, K.G., 1969. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika 34 (2), 183–202.
Jöreskog, K.G., 1971. Simultaneous factor analysis in several populations. Psychometrika 36 (4), 409–426.
Jöreskog, K.G., Goldberger, A.S., 1975. Estimation of a model with multiple indicators and multiple causes of a single latent variable. J. Am. Stat. Assoc. 70 (351), 631–639.
Jowell, R., Roberts, C., Fitzgerald, R., Eva, G., 2007. Measuring Attitudes Cross-Nationally: Lessons from the European Social Survey. Sage, London.
Kaplan, D., 2009. Structural Equation Modeling: Foundations and Extensions, second ed. Sage, Thousand Oaks.
Kaplan, D., 2014. Bayesian Statistics for the Social Sciences. Guilford Press, New York.
Kaplan, D., Depaoli, S., 2012. Bayesian structural equation modeling. In: Hoyle, R.H. (Ed.), Handbook of Structural Equation Modeling. Guilford Press, New York, pp. 650–673.
Kim, E.S., Cao, C., 2015. Testing group mean differences of latent variables in multilevel data using multiple-group multilevel CFA and multilevel MIMIC modeling. Multivariate Behav. Res. 50 (4), 436–456.
Kim, E.S., Cao, C., Wang, Y., Nguyen, D.T., 2017. Measurement invariance testing with many groups: a comparison of five approaches. Struct. Equ. Model. 24 (4), 524–544.
Kim, E.S., Yoon, M., 2011. Testing measurement invariance: a comparison of multiple group categorical CFA and IRT. Struct. Equ. Model. 18 (2), 212–228.
Kim, E.S., Yoon, M., Lee, T., 2011. Testing measurement invariance using MIMIC: likelihood ratio test with a critical value adjustment. Educ. Psychol. Meas. 72 (3), 469–492.
Kim, E.S., Yoon, M., Wen, Y., Luo, W., Kwok, O.M., 2015. Within-level group factorial invariance with multilevel data: multilevel factor mixture and multilevel MIMIC models. Struct. Equ. Model. 22 (4), 603–616.
Klausch, T., Hox, J.J., Schouten, B., 2013. Measurement effects of survey mode on the equivalence of attitudinal rating scale questions. Socio. Methods Res. 42 (3), 227–263.
Kline, R., 2016. Principles and Practice of Structural Equation Modeling. Guilford Press, New York.
Koc, P., Pokropek, A., 2022. Accounting for cross-country-cross-time variations in measurement invariance testing. A case of political participation. Surv. Res. Methods 16 (1), 79–96.
Krebs, D., 2012. The impact of response format on attitude measurement. In: Salzborn, S., Davidov, E., Reinecke, J. (Eds.), Methods, Theories, and Empirical Applications in the Social Sciences. Festschrift for Peter Schmidt. Springer VS, Wiesbaden, pp. 105–113.
Krosnick, J.A., Fabrigar, L.R., 1997. Designing rating scales for effective measurement in surveys. In: Lyberg, L., Biemer, P.P., Collins, M., De Leeuw, E.D., Dippo, C., Schwarz, N. (Eds.), Survey Measurement and Process Quality. Wiley, New York, pp. 141–164.
Kruschke, J.K., 2021. Bayesian analysis reporting guidelines. Nat. Human Behav. 5 (10), 1282–1291.
Kuha, J., Moustaki, I., 2015. Non-equivalence of measurement in latent variable modeling of multigroup data: a sensitivity analysis. Psychol. Methods 20 (4), 523–536.
Lai, M.H., Liu, Y., Tse, W.W.-Y., 2022. Adjusting for partial invariance in latent parameter estimation: comparing forward specification search and approximate invariance methods. Behav. Res. Methods 54, 414–434.
Lawley, D.N., 1943. On problems connected with item selection and test construction. Proc. R. Soc. Edinb. Sect. A (Math. Phys. Sci.) 61 (3), 74–82.
Lawley, D.N., Maxwell, A.E., 1963. Factor Analysis as a Statistical Method. Butterworth, London.
Lee, S.-Y., 2007. Structural Equation Modeling. A Bayesian Approach. Wiley, Chichester.
Lee, J., Little, T.D., Preacher, K.J., 2018. Methodological issues in using Structural equation models for testing differential item functioning. In: Davidov, E., Schmidt, P., Billiet, J., Meulemann, B. (Eds.), Cross-cultural Analysis: Methods and Applications, second ed. Routledge, New York, pp. 65–94.
Lee, S., Vasquez, E., Ryan, L., Smith, J., 2020. Measurement equivalence of subjective bell-being scales under the presence of acquiescent response style for the racially and ethnically diverse older population in the United States. Surv. Res. Methods 14 (4), 417–437.
Lei, M., Lomax, R.G., 2005. The effect of varying degrees of nonnormality in structural equation modeling. Struct. Equ. Model. 12 (1), 1–27.
Leitgöb, H., & Seddig, D. (forthcoming). Identifying true change and response shifts across time. A multi-group confirmatory factor analysis approach for repeated cross-sectional data.
Leitgöb, H., Seddig, D., Schmidt, P., Sosu, E., Davidov, E., 2021. Longitudinal measurement (non-)invariance in latent constructs: conceptual insights, model specifications, and testing strategies. In: Cernat, A., Sakshaug, J. (Eds.), Measurement Error in Longitudinal Data. Oxford University Press, Oxford, pp. 211–257.
Lek, K., Oberski, D., Davidov, E., Cieciuch, J., Seddig, D., Schmidt, P., 2019. Approximate measurement invariance. In: Johnson, T.P., Pennell, B.-E., Stoop, I.A.L., Dorer, B. (Eds.), Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC). Wiley, Hoboken, pp. 911–928.
Little, T.D., 1997. Mean and covariance structures (MACS) analysis of cross-cultural data: practical and theoretical issues. Multivariate Behav. Res. 32 (1), 53–76.
Little, T.D., 2013. Longitudinal Structural Equation Modeling. Guilford Press, New York.
Little, T.D., Slegers, D.W., Card, N.A., 2006. A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. Struct. Equ. Model. 13 (1), 59–72.
Lomazzi, V., 2018. Using alignment optimization to test the measurement invariance of gender role attitudes in 59 countries. Methods, Data, Analyses 12 (1), 77–103.
Lommen, M.J.J., Engelhard, I.M., Sijbrandij, M., van den Hout, M.A., Hermans, D., 2013. Pre-trauma individual differences in extinction learning predict posttraumatic stress. Behav. Res. Ther. 51 (2), 63–67.
Lord, F.M., 1952. A Theory of Test Scores. Psychometric Society, New York.
Lord, F.M., 1980. Applications of Item Response Theory to Practical Testing Problems. Lawrence Erlbaum Associates, Mahwah.
Lord, F.M., Novick, M.R., 1968. Statistical Theories of Mental Test Scores. Addison-Wesley, Reading.
Lyberg, L., Pennell, B.-E., Cibelli Hibben, K., de Jong, J., Behr, D., Burnett, J., Fitzgerald, R., Granda, P., Luz Guerrero, L., Gyuzalyan, H., Johnson, T., Kim, J., Mneimneh, Z., Moynihan, P., Robbins, M., Schoua-Glusberg, A., Sha, M., Smith, T.W., Zechmeister, E.J., 2021. AAPOR/WAPOR Task Force Report in Comparative Surveys. Retrieved from. https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/images/AAPOR-WAPOR-Task-Force-Report-on-Quality-in-Comparative-Surveys_Full-Report.pdf.
Lynn, P., 2020. Evaluating push-to-web methodology for mixed-mode surveys using address-based samples. Surv. Res. Methods 14 (1), 19–30.
Lynn, P., Japec, L., Lyberg, L., 2006. What's so special about cross-national surveys? In: Harkness, J.A. (Ed.), Conducting Cross-National and Cross-Cultural Surveys. Zuma, Mannheim, pp. 7–21.

Maas, C.J.M., Hox, J.J., 2005. Sufficient sample sizes for multilevel modeling. Methodology 1 (3), 86–92.
Marsh, H.W., Guo, J., Parker, P.D., Nagengast, B., Asparouhov, T., Muthén, B.O., Dicke, T., 2018. What to do when scalar invariance fails: the extended alignment method for multi-group factor analysis comparison of latent means across many groups. Psychol. Methods 23 (3), 524–545.
Marsh, H.W., Hocevar, D., 1985. Application of confirmatory factor analysis to the study of self-concept: first- and higher order factor models and their invariance across groups. Psychol. Bull. 97 (3), 562–582.
Marsh, H.W., Kit-Tai, H., Wen, Z., 2004. In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. Struct. Equ. Model. 11 (3), 320–341.
Marsh, H.W., Morin, A.J.S., Parker, P.D., Kaur, G., 2014. Exploratory structural equation modeling: an integration of the best features of exploratory and confirmatory factor analyses. Annu. Rev. Clin. Psychol. 10, 85–110.
Martin, S.L., Birhanu, Z., Omotayo, M.O., Kebede, Y., Pelto, G.H., Stoltzfus, R.J., Dickin, K.L., 2017. I can't answer what you're asking me. Let me go, please: cognitive interviewing to assess social support measures in Ethiopia and Kenya. Field Methods 29 (4), 317–332.
Maskileyson, D., Seddig, D., Davidov, E., 2021a. The EURO-D measure of depressive symptoms in the aging population: comparability across European countries and Israel. Frontiers in Political Science 3, 665004.
Maskileyson, D., Seddig, D., Davidov, E., 2021b. The comparability of perceived physical and mental health measures across immigrants and natives in the United States. Demography 58 (4), 1423–1443.
McLachlan, G.J., Peel, D., 2000. Mixtures of factor analyzers. In: Langley, P. (Ed.), Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann, San Francisco, pp. 599–606.
Meade, A.W., 2010. A taxonomy of effect size measures for the differential functioning of items and scales. J. Appl. Psychol. 95 (4), 728–743.
Meitinger, K., 2017. Necessary but Insufficient: why measurement invariance tests need online probing as a complementary tool. Publ. Opin. Q. 81 (2), 447–472.
Meitinger, K., Behr, D., 2016. Comparing cognitive interviewing and online probing: do they find similar results? Field Methods 28 (4), 363–380.
Meitinger, K., Davidov, E., Schmidt, P., Braun, M., 2020. Measurement invariance: testing for it and explaining why it is absent. Surv. Res. Methods 14 (4), 345–349.
Mellenbergh, G.J., 1989. Item bias and item response theory. Int. J. Educ. Res. 13 (2), 127–143.
Menold, N., 2020a. Rating-scale labeling in online surveys: an experimental comparison of verbal and numeric rating scales with respect to measurement quality and respondents' cognitive processes. Socio. Methods Res. 49 (1), 79–107.
Menold, N., 2020b. Double barreled questions: an analysis of the similarity of elements and the measurement quality. J. Off. Stat. 36 (4), 855–886.
Menold, N., Kemper, C.J., 2015. The impact of frequency rating scale formats on the measurement of latent variables in web surveys: an experimental investigation using a measure of affectivity as an example. Psihologija 48 (4), 431–449.
Menold, N., Raykov, T., 2016. Can reliability of multiple component measuring instruments depend on response option presentation mode? Educ. Psychol. Meas. 76 (3), 454–469.
Menold, N., Tausch, A., 2016. Measurement of latent variables with different rating scales: testing reliability and measurement equivalence by varying the verbalization and number of categories. Socio. Methods Res. 45 (4), 678–699.
Menold, N., Toepoel, V., 2022. Do different devices perform equally well with different numbers of scale points and response formats? A test of measurement invariance and reliability. Socio. Methods Res. Advance online publication.
Meredith, W., 1993. Measurement invariance, factor analysis, and factorial invariance. Psychometrika 58 (4), 525–543.
Meredith, W., Teresi, J.A., 2006. An essay on measurement and factorial invariance. Med. Care 44 (11), 69–77. Suppl 3.
Meuleman, B., 2012. When are intercept differences substantively relevant in measurement invariance testing? In: Salzborn, S., Davidov, E., Reinecke, J. (Eds.), Methods, Theories, and Empirical Applications in the Social Sciences. Festschrift for Peter Schmidt. Springer VS, Wiesbaden, pp. 97–104.
Meuleman, B., Davidov, E., Seddig, D., 2018a. Editorial: comparative survey analysis: comparability and equivalence of measures. Methods, Data, Analyses 12 (1), 3–6.
Meuleman, B., Davidov, E., Seddig, D., 2018b. Editorial: comparative survey analysis: models, techniques, and applications. Methods, Data, Analyses 12 (2), 181–184.
Meuleman, B., Žóltak, T., Pokropek, A., Davidov, E., Muthén, B., Oberski, D.L., Billiet, J., Schmidt, P., 2022. Why measurement invariance is important in comparative research. A response to Welzel et al, 2021 Socio. Methods Res. Advance online publication.
Miller, K., 2019. Conducting cognitive interviewing studies to examine survey question comparability. In: Johnson, T.P., Pennell, B.-E., Stoop, I.A.L., Dorer, B. (Eds.), Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC). Wiley, New York, pp. 203–226.
Millsap, R.E., 2011. Statistical Approaches to Measurement Invariance. Taylor & Francis Group, New York.
Millsap, R.E., Everson, H.T., 1993. Methodology review: statistical approaches for assessing measurement bias. Appl. Psychol. Meas. 17 (4), 297–334.
Millsap, R.E., Meredith, W., 2007. Factorial invariance: historical perspectives and new problems. In: Cudeck, R., MacCallum, R.C. (Eds.), Factor Analysis at 100: Historical Developments and Future Directions. Lawrence Erlbaum Associates, Mahwah, pp. 131–152.
Mneimneh, Z., Hibben, K.C., Bilal, L., Hyder, S., Shahab, M., Binmuammar, A., Altwaijri, Y., 2018. Probing for sensitivity in translated survey questions: differences in respondent feedback across cognitive probe types. Trans. Interpr. 10 (2), 73–88.
Montoya, A.K., Jeon, M., 2020. MIMIC models for uniform and nonuniform DIF as moderated mediation models. Appl. Psychol. Meas. 44 (2), 118–136.
Morgan, S.L., Winship, C., 2015. Counterfactuals and Causal Inference, second ed. Cambridge University Press, New York.
Munck, I., Barber, C., Torney-Purta, J., 2018. Measurement invariance in comparing attitudes toward immigrants among youth across Europe in 1999 and 2009: the alignment method applied to IEA CIVED and ICCS. Socio. Methods Res. 47 (4), 687–728.
Muthén, B.O., 1988. Some uses of structural equation modeling in validity studies: extending IRT to external variables. In: Wainder, H., Braun, H. (Eds.), Test Validity. Lawrence Erlbaum Associates, Hillsdale, pp. 213–238.
Muthén, B.O., 1989. Latent variable modeling in heterogeneous populations. Psychometrika 54 (4), 557–585.
Muthén, B.O., 1990. *Mean And Covariance Structure Analysis of Hierarchical Data* (UCLA Statistics Series, No. 62). UCLA, Los Angeles.
Muthén, B.O., 1994. Multilevel covariance structure analysis. Socio. Methods Res. 22 (3), 376–398.
Muthén, B.O., Asparouhov, T., 2002. *Latent Variable Analysis with Categorical Outcomes: Multiple-Group and Growth Modeling* (Mplus Web Notes No. 4). Retrieved from. http://www.statmodel.com/examples/webnote.shtml.
Muthén, B.O., Asparouhov, T., 2012. Bayesian structural equation modeling: a more flexible representation of substantive theory. Psychol. Methods 17 (3), 313–335.
Muthén, B.O., Asparouhov, T., 2014. IRT studies of many groups: the alignment method. Front. Psychol. 5, 978.
Muthén, B.O., Asparouhov, T., 2018. Recent methods for the study of measurement invariance with many groups: alignment and random effects. Socio. Methods Res. 47 (4), 637–664.
Muthén, B.O., Khoo, S.T., Gustafsson, J.E., 1997. Multilevel Latent Variable Modeling in Multiple Populations. Retrieved from. http://www.statmodel.com/papers.shtml.
Neuert, C.E., Meitinger, K., Behr, D., 2021. Open-ended versus closed probes: assessing different formats of web probing. Socio. Methods Res. Advance online publication.
Newsom, J.T., 2015. Longitudinal Structural Equation Modeling: A Comprehensive Introduction. Routledge, New York.
Nye, C.D., Bradburn, J., Olenick, J., Bialko, C., Drasgow, F., 2019. How big are my effects? Examining the magnitude of effect sizes in studies of measurement equivalence. Organ. Res. Methods 22 (3), 678–709.
Nye, C.D., Drasgow, F., 2011. Effect size indices for analyses of measurement equivalence: understanding the practical importance of differences between groups. J. Appl. Psychol. 96 (5), 966–980.
Oaxaca, R., 1973. Male-female wage differentials in urban labor markets. Int. Econ. Rev. 14 (3), 693–709.
Oberski, D.L., 2014. Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. Polit. Anal. 22 (1), 45–60.
O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., Rakow, T., 2006. Uncertain Judgements: Eliciting Experts' Probabilities. Wiley, Hoboken.
Olaru, G., Danner, D., 2020. Developing cross-cultural short scales using ant colony optimization. Assessment 28 (1), 199–210.

Olaru, G., Schroeders, U., Hartung, J., Wilhelm, O., 2019. Ant colony optimization and local weighted structural equation modeling. A tutorial on novel item and person sampling procedures for personality research. Eur. J. Pers. 33 (3), 400–419.
Oort, F.J., 2005. Using structural equation modeling to detect response shifts and true change. Qual. Life Res. 14 (3), 587–598.
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. Struct. Equ. Model. 26 (5), 724–744.
Pokropek, A., Pokropek, E., 2022. Deep neural networks for deteting statistical model misspecifications. The case of measurement invariance. Struct. Equ. Model. 29 (3), 394–411.
Pokropek, A., Schmidt, P., Davidov, E., 2020. Choosing priors in Bayesian measurement invariance modeling: a Monte Carlo simulation study. Struct. Equ. Model. 27 (5), 750–764.
Putnick, D.L., Bornstein, M.H., 2016. Measurement invariance conventions and reporting: the state of the art and future directions for psychological research. Dev. Rev. 41, 71–90.
Rabe-Hesketh, S., Skrondal, A., Pickles, A., 2004. Generalized multilevel structural equation modeling. Psychometrika 69 (2), 167–190.
Radloff, L.S., 1977. The CES-D scale: a self-report depression scale for research in the general population. Appl. Psychol. Meas. 1 (3), 385–401.
Rapkin, B.D., Schwartz, C.E., 2004. Toward a theoretical model of quality-of-life appraisal: implications of findings from studies of response shift. Health Qual. Life Outcome 2, 14.
Rapkin, B.D., Schwartz, C.E., 2019. Advancing quality-of-life research by deepening our understanding of response shift: a unifying theory of appraisal. Qual. Life Res. 28 (10), 2623–2630.
Rasch, G., 1960. Probabilistic Models for Some Intelligence and Attainment Tests. Paedagogiske Institut, Copenhagen.
Raykov, T., Marcoulides, G.A., Millsap, R.E., 2013. Factorial invariance in multiple populations: a multiple testing procedure. Educ. Psychol. Meas. 73 (4), 713–727.
Remizova, A., Rudnev, M., Davidov, E., 2022. In search of a comparable measure of generalized individual religiosity in the world values survey. Socio. Methods Res. Advance online publication.
Rensvold, R.B., Cheung, G.W., 2000. Beyond two-group comparisons: identifying sets of invariant groups. Acad. Manag. Proc. 2000 (1), A1–A6.
Revilla, M.A., 2013. Measurement invariance and quality of composite scores in a face-to-face and a web survey. Surv. Res. Methods 7 (1), 17–28.
Reynolds, A.P., Richards, G., de la Iglesia, B., Rayward-Smith, V.J., 2006. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. J. Math. Model. Algorithm. 5 (4), 475–504.
Rios, J.A., Sireci, S.G., 2014. Guidelines versus practices in cross-lingual assessment: a disconcerting disconnect. Int. J. Test. 14 (4), 289–312.
Roberts, C., Sarrasin, O., Stähli, E.M., 2020. Investigating the relative impact of different sources of measurement non-equivalence in comparative surveys: an illustration with scale format, data collection mode and cross-national variations. Surv. Res. Methods 14 (4), 399–415.
Robitzsch, A., 2021. *Sirt: Supplementary Item Response Theory Models* (R Package Reference Manual. Retrieved from, version 3.11-21. https://cran.r-project.org/web/packages/sirt/index.html.
Rudnev, M., 2018-2022. Measurement Invariance Explorer. Retrieved from. https://github.com/MaksimRudnev/MIE.package.
Rudnev, M., 2019. Elements of Cross-Cultural Research. Retrieved from. https://maksimrudnev.com/2019/05/01/alignment-tutorial.
Rudnev, M., Lytkina, E., Davidov, E., Schmidt, P., Zick, A., 2018. Testing measurement invariance for a second-order factor: a cross-national test of the alienation scale. Methods, Data, Analyses 12 (1), 47–76.
Rutkowski, L., Svetina, D., 2014. Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. Educ. Psychol. Meas. 74 (1), 31–57.
Ryu, E., 2014. Factorial invariance in multilevel confirmatory analysis. Br. J. Math. Stat. Psychol. 67 (1), 172–194.
Sajobi, T.T., Brahmbatt, R., Lix, L.M., Zumbo, B.D., Sawatzky, R., 2018. Scoping review of response shift methods: current reporting practices and recommendations. Qual. Life Res. 27 (5), 1133–1146.
Sakshaug, J., Cernat, A., Silverwood, R.J., Calderwood, L., Ploubidis, G.B., 2022. Measurement equivalence in sequential mixed-mode surveys. Surv. Res. Methods 16 (1), 29–43.
Samejima, F., 1969. Estimation of latent ability using a response pattern of graded scores. In: (Psychometrika Monograph No. 17). Psychometric Society, Richmond.
Saris, W.E., Gallhofer, G., 2007. Estimation of the effects of measurement characteristics on the quality of survey questions. Surv. Res. Methods 1 (1), 29–43.
Sarrasin, O., Green, E.G.T., Berchtold, A., Davidov, E., 2012. Measurement equivalence across subnational groups: an analysis of the conception of nationhood in Switzerland. Int. J. Publ. Opin. Res. 25 (4), 522–534.
Schmidt, W.H., 1969. Covariance Structure Analysis of the Multivariate Random Effects Model. Unpublished doctoral dissertation, University of Chicago.
Schuman, H., 1966. The random probe: a technique for evaluating the validity of closed questions. Am. Socio. Rev. 31 (2), 218–222.
Schwartz, C.E., Sprangers, M.A.G., 1999. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. Soc. Sci. Med. 48 (11), 1531–1548.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464.
Sébille, V., Lix, L.M., Ayilara, O.F., Sajobi, T.T., Janssens, A.C.J.W., Sawatzky, R., Sprangers, M.A.G., Verdam, M.G.E., 2021. Critical examination of current response shift methods and proposal for advancing new methods. Qual. Life Res. 30 (12), 3325–3342.
Seddig, D., Leitgöb, H., 2018. Approximate measurement invariance and longitudinal confirmatory factor analysis: concept and application with panel data. Surv. Res. Methods 12 (1), 29–41.
Seddig, D., Lomazzi, V., 2019. Using cultural and structural indicators to explain measurement noninvariance in gender role attitudes with multilevel structural equation modeling. Soc. Sci. Res. 84, 102328.
Seddig, D., Maskileyson, D., Davidov, E., 2020. The comparability of measures in the ageism module of the fourth round of the European Social Survey, 2008–2009. Surv. Res. Methods 14 (4), 351–364.
Shi, D., Song, H., Liao, X., Terry, R., Snyder, L.A., 2017. Bayesian SEM for specification search problems in testing factorial invariance. Multivariate Behav. Res. 52 (4), 430–444.
Sideridis, G.D., Tsaousis, I., Alamri, A.A., 2020. Accounting for differential item functioning using Bayesian approximate measurement invariance. Educ. Psychol. Meas. 80 (4), 638–664.
Smith, T.W., 2004. Developing and evaluating cross-national survey instruments. In: Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., Singer, E. (Eds.), Methods for Testing and Evaluating Survey Questionnaires. Wiley, Hoboken, pp. 431–452.
Smith, T.W., 2019. Optimizing questionnaire design in cross-national and cross-cultural surveys. In: Beatty, P.C., Collins, D., Kaye, L., Padilla, J.-L., Willis, G.B., Wilmot, A. (Eds.), Advances in Questionnaire Design, Development, Evaluation and Testing. Wiley, Hoboken, pp. 473–488.
Sokolov, B., 2018. The index of emancipative values: measurement model misspecifications. Am. Polit. Sci. Rev. 112 (2), 395–408.
Son, S., Hong, S., 2021. Multiple group analysis in multilevel data across within-level groups: a comparison of multilevel factor mixture modeling and multilevel multiple-indicators multiple-causes modeling. Educ. Psychol. Meas. 81 (5), 904–935.
Song, X.-Y., Lee, S.-Y., 2012. Basic and Advanced Bayesian Structural Equation Modeling. With Applications in the Medical and Behavioral Sciences. Wiley, Chichester.
Sörbom, D., 1974. A general method for studying differences in factor means and factor structure between groups. Br. J. Math. Stat. Psychol. 27 (2), 229–239.
Sörbom, D., 1989. Model modification. Psychometrika 54 (3), 371–384.
Spearman, C., 1904. General intelligence. objectively determined and measured. *American Journal of Psychology* 15 (2), 201–293.
Sprangers, M.A.G., Schwartz, C.E., 1999. Integrating response shift into health-related quality of life research: a theoretical model. Soc. Sci. Med. 48 (11), 1507–1515.
Stark, S., Chernyshenko, O.S., Drasgow, F., 2006. Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. J. Appl. Psychol. 91 (6), 1292–1306.
Steenkamp, J.-B.E.M., Baumgartner, H., 1998. Assessing measurement invariance in cross-national consumer research. J. Consum. Res. 25 (1), 78–90.
Steinmetz, H., 2013. Analyzing observed composite differences across groups: is partial measurement invariance enough? Methodology 9 (1), 1–12.

Teresi, J.A., 2006. Overview of quantitative measurement methods: equivalence, invariance, and differential item functioning in health applications. Med. Care 44 (11), S39–S49.
Thompson, M.S., Green, S.B., 2013. Evaluating between-group differences in latent variable means. In: Hancock, G.R., Mueller, R.O. (Eds.), Structural Equation Modeling: A Second Course, second ed. Information Age Publishing, Greenwich.
Thompson, Y.T., Song, H., Shi, D., Liu, Z., 2021. It matters: reference indicator selection in measurement invariance tests. Educ. Psychol. Meas. 81 (1), 5–38.
Thurstone, L.L., 1947. Multiple Factor Analysis. University of Chicago Press, Chicago.
Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. J. Roy. Stat. Soc. B 63 (2), 411–423.
Tourangeau, R., Rips, L.J., Rasinski, K.A., 2000. The Psychology of Survey Response. Cambridge University Press, Cambridge, UK.
Vandenberg, R.J., 2002. Toward a further understanding of and improvement in measurement invariance methods and procedures. Organ. Res. Methods 5 (2), 139–158.
Vandenberg, R.J., Lance, C.E., 2000. A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. Organ. Res. Methods 3 (1), 4–70.
van der Veld, W., Saris, W., 2018. Measurement equivalence 2.0. In: Davidov, E., Schmidt, P., Billiet, J., Meulemann, B. (Eds.), Cross-cultural Analysis: Methods and Applications, second ed. Routledge, New York, pp. 245–279.
van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M.G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., Yau, C., 2021. Bayesian statistics and modelling. Nat. Rev. Methods Primers 1 (1), 1–26.
van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J.B., Neyer, F.J., van Aken, M.A.G., 2014. A gentle introduction to Bayesian anylsis: applications to developmental research. Child Dev. 85 (3), 842–860.
van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., Muthén, B.O., 2013. Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. Front. Psychol. 4, 770.
van de Schoot, R., Schmidt, P., De Beuckelaer, A. (Eds.), 2015. Measurement Invariance. Frontiers Media, Lausanne.
van de Vijver, F.J.R., 2013. Contributions of internationalization to psychology: toward a global and inclusive discipline. Am. Psychol. 68 (8), 761.
van de Vijver, F.J.R., 2018a. Capturing bias in structural equation modeling. In: Davidov, E., Schmidt, P., Billiet, J., Meuleman, B. (Eds.), Cross-cultural Analysis: Methods and Applications, second ed. Routledge, New York, pp. 3–44.
van de Vijver, F.J.R., 2018b. Towards an integrated framework of bias in noncognitive assessment in international large-scale studies: challenges and prospects. Educ. Meas. 37 (4), 49–56.
van de Vijver, F.J.R., Avvisati, F., Davidov, E., Eid, M., Fox, J.P., Le Donné, N., Lek, K., Meuleman, B., Paccagnella, M., van de Schoot, R., 2019. *Invariance Analyses in Large-Scale Studies* (OECD Education Working Papers No. 201). OECD Publishing, Paris.
van de Vijver, F.J.R., Leung, K., 2011. Equivalence and bias: a review of concepts, models, and data analytic procedures. In: Matsumoto, D., van de Vijver, F.J.R. (Eds.), Culture and Psychology. Cross-Cultural Research Methods in Psychology. Cambridge University Press, Cambridge, pp. 17–45.
van de Vijver, F.J.R., Leung, K., 2021. Methodological concepts in cross-cultural research. In: Fetvadjiev, V.H., He, J., Fontaine, J.R.J. (Eds.), Methods and Data Analysis for Cross-Cultural Research. Cambridge University Press, Cambridge, pp. 4–9.
van de Vijver, F.J.R., Poortinga, Y.H., 1997. Towards an integrated analysis of bias in cross-cultural assessment. Eur. J. Psychol. Assess. 13 (1), 29–37.
van Erp, S., Mulder, J., Oberski, D.L., 2018. Prior sensitivity analysis in default Bayesian structural equation modeling. Psychol. Methods 23 (2), 363–388.
van Kesteren, E.-J., Oberski, D.L., 2022. Flexible extensions to structural equation models using computation graphs. Struct. Equ. Model. 29 (2), 233–247.
Veen, D., Egberts, M.R., Van Loey, N.E., van de Schoot, R., 2020. Expert elicitation for latent growth curve models: the case of posttraumatic stress symptoms development in children with burn injuries. Front. Psychol. 11, 1197.
Verdam, M.G.E., Oort, F.J., 2014. Measurement bias detection with Kronecker product restricted models for multivariate longitudinal data: an illustration with health-related quality of life data from thirteen measurement occasions. Front. Psychol. 5, 1022.
Verdam, M.G.E., Oort, F.J., 2019. The analysis of multivariate longitudinal data: an instructive application of the longitudinal three-mode model. Multivariate Behav. Res. 54 (4), 457–474.
Verhagen, A.J., Fox, J.P., 2013. Bayesian tests of measurement invariance. Br. J. Math. Stat. Psychol. 66 (3), 383–401.
Vermunt, J.K., 2010. Latent class modeling with covariates: two improved three-step approaches. Polit. Anal. 18 (4), 450–469.
Vermunt, J.K., Magidson, J., 2013. Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax. Statistical Innovations Inc, Belmont.
Vermunt, J.K., Magidson, J., 2016. Upgrade Manual for Latent GOLD, 5.1. Statistical Innovations, Belmont.
Weber, W., 2011. Testing for measurement equivalence of individuals' left-right orientation. Surv. Res. Methods 5 (1), 1–10.
Welkenhuysen-Gybels, J., Billiet, J., 2002. A comparison of techniques for detecting cross-cultural inequivalence at the item level. Qual. Quantity 36 (3), 197–218.
Welzel, C., Brunkert, L., Kruse, S., Inglehart, R.F., 2021. Non-invariance? An overstated problem with misconceived causes. Socio. Methods Res. (Advance online publication).
Welzel, C., Kruse, S., Brunkert, L., 2022. Against the mainstream: on the limitations of non-invariance diagnostics. Response to Fischer et al. and Meuleman et al. Socio. Methods Res. Advance online publication.
Weng, L.-J., 2004. Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. Educ. Psychol. Meas. 64 (6), 956–972.
West, S.G., Taylor, A.B., Wu, W., 2012. Model fit and model selection in structural equation modeling. In: Hoyle, R.H. (Ed.), Handbook of Structural Equation Modeling. Guilford Press, New York, pp. 209–231.
Wiley, D.E., Schmidt, W.H., Bramble, W.J., 1973. Studies of a class of covariance structure models. J. Am. Stat. Assoc. 86 (342), 317–321.
Willis, G.B., 2005. Cognitive Interviewing. A Tool for Improving Questionnaire Design. Sage, Thousand Oaks.
Willis, G.B., 2015. Research synthesis: the practice of cross-cultural cognitive interviewing. Publ. Opin. Q. 79 (S1), 359–395.
Willis, G.B., Miller, K., 2011. Cross-cultural cognitive interviewing: seeking comparability and enhancing understanding. Field Methods 23 (4), 331–341.
Woods, C.M., Grimm, K.J., 2011. Testing for nonuniform differential item functioning with multiple indicator multiple cause models. Appl. Psychol. Meas. 35 (5), 339–361.
Wu, A.D., Li, Z., Zumbo, B.D., 2007. Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: a demonstration with TIMSS Data. Practical Assess. Res. Eval. 12 (1), 1–26.
Yoon, M., Millsap, R.E., 2007. Detecting violations of factorial invariance using data-based specification searches: a Monte Carlo study. Struct. Equ. Model. 14 (3), 435–463.
Zavala-Rojas, D., Saris, W.E., Gallhofer, I.N., 2019. Preventing differences in translated survey items using the survey quality predictor. In: Johnson, T.P., Pennell, B.-E., Stoop, I.A.L., Dorer, B. (Eds.), Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC). Wiley, Hoboken, pp. 357–379.
Zercher, F., Schmidt, P., Cieciuch, J., Davidov, E., 2015. The comparability of the universalism value over time and across countries in the European Social Survey: exact vs. approximate measurement invariance. Front. Psychol. 6, 733.
Zhang, Y., Lai, M.H.C., Palardy, G.J., 2022. A Bayesian region of measurement equivalence (ROME) approach for establishing measurement invariance. Psychol. Methods. Advance online publication.