Journal of Applied Psychology

Investigating Measurement Invariance for Multiple Covariates in Organizational Research Using Exploratory Factor Analysis and Confirmatory Factor Analysis Trees

David Goretzko, Matt C. Howard, and Philipp Sterner

Online First Publication, February 19, 2026. https://dx.doi.org/10.1037/apl0001368

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Applied Psychology

ISSN: 0021-9010

https://doi.org/10.1037/apl0001368

Investigating Measurement Invariance for Multiple Covariates in Organizational Research Using Exploratory Factor Analysis and Confirmatory Factor Analysis Trees

David Goretzko[1, 2], Matt C. Howard[3], and Philipp Sterner[4, 5]

[1] Department of Psychological Methods (With a Focus on Methods for Psychotherapy Research), Goethe University Frankfurt
[2] Department of Methodology and Statistics, Utrecht University
[3] Department of Marketing and Quantative Methods, University of South Alabama
[4] Center for International Student Assessment, Technical University Munich
[5] Department of Psychology, Ludwig Maximilians University Munich

Organizational research often deals with unobservable (latent) variables such as, for example, job satisfaction or leadership styles. When comparing these latent variables across groups, a comparability of the measurements is important—so-called measurement invariance (MI) considered a prerequisite. Common methodology to test whether MI holds or to explore noninvariance can only be used with established measurement models and specific hypotheses about potential violations of MI in mind. Therefore, exploratory factor analysis trees and confirmatory factor analysis trees have recently been developed. They promise to be an effective tool for early investigations of MI during the development of measurement models (e.g., scale development) and with many (continuous) covariates defining countless groups for which MI may be violated.

Keywords: measurement invariance, measurement equivalence, questionnaire development, exploratory factor analysis, confirmatory factor analysis

Organizational researchers and practitioners are often interested in the analysis of constructs between groups (e.g., race and gender; Cho et al., 2023; Jebb & Tay, 2017; Lu, 2023; Podsakoff et al., 2019). However, inferences from these analyses can be misleading, if applied measures do not function equivalently across groups or measurement occasions, with these differences arising due to a variety of sources (e.g., culture, organization, assessment, and time; Robert et al., 2006; Somaraju et al., 2022). This differential functioning of measures is known as measurement nonequivalence (MNE) or noninvariance, which can cause hypothesis tests to result in false positives (i.e., Type I error) or false negatives (i.e., Type II error) that ultimately produce misleading inferences (Davidov et al., 2014; Raju et al., 2002; Vandenberg & Lance, 2000). Due to the consequential biasing influence of MNE, it is widely considered essential for researchers to confirm the equivalence (or invariance) of their measures before proceeding to hypothesis testing (Nye et al., 2019; Somaraju et al., 2022; Vandenberg, 2002).

A number of researchers have developed and tested hypotheses regarding the sources of measurement noninvariance, studying these sources as substantive effects rather than biasing influences alone (e.g., Bynum et al., 2013; Laguna et al., 2017; Somaraju et al., 2022). For instance, researchers have advanced measurement theory by treating measurement invariance (MI) across raters as explainable sources of variance (Bynum et al., 2013; Merritt, 2012; Morelli et al., 2014), and researchers have added temporal and cultural nuance to extend organizational theory by assessing MI over time and across cultures as explainable changes in the nature and/or meaning of constructs (Harari et al., 2019; Laguna et al., 2017; Nye et al., 2010). From this shift in perspective, recent authors have argued that typical sources investigated in tests of MI (e.g., race, gender, and time) are only proxies for more substantive influences (e.g., cultural frames, stereotype threat, and maturation). These authors recommend that researchers should instead create and test hypotheses regarding these substantive influences to further scholarship on both measurement and the substantive influence itself (Burlew et al., 2019; Ruglass et al., 2020; Somaraju et al., 2022).

Whether treated as a biasing influence or substantive effect, MI and associated analyses are widely popular in organizational research, particularly since the work of Vandenberg and Lance (2000) that reviewed and clarified the still predominant approach for testing MI, multiple-group confirmatory factor analysis (MG-CFA). Despite being the most popular method to test MI, MG-CFA is not always the best approach due to three central limitations: (a) groups to test MI for need to be specified a priori by the researcher, (b) it is not possible to assess continuous influences on MI, and (c) a fully specified CFA is required, that is, it cannot be used in exploratory settings

Scott B. Morris served as action editor.

David Goretzko https://orcid.org/0000-0002-2730-6347

David Goretzko played a lead role in conceptualization, project administration, and writing–original draft and an equal role in methodology and writing–review and editing. Matt C. Howard played a supporting role in conceptualization, writing–original draft, and writing–review and editing. Philipp Sterner played a lead role in formal analysis, a supporting role in conceptualization, and an equal role in methodology, writing–original draft, and writing–review and editing.

Correspondence concerning this article should be addressed to David Goretzko, Department of Psychological Methods (With a Focus on Methods for Psychotherapy Research), Goethe University Frankfurt, Theodor-W.-Adorno-Platz 6, 60323 Frankfurt, Germany. Email: goretzko@psych.uni-frankfurt.de

(e.g., early stages of scale development). Hence, new methods have been developed that address these issues. Among many others, exploratory factor analysis (EFA) trees (Sterner & Goretzko, 2023) as extensions or special cases of structural equation modeling (SEM) trees (Brandmaier et al., 2013b) utilize a decision-tree–based method (common in machine learning applications) to recursively assess EFA models in a data-driven manner to investigate MI.

The use of EFA comes with the advantage that no (potentially too restrictive) measurement model has to be specified. This is especially useful in earlier stages of questionnaire development, where measurement models are still "under construction" (Sterner & Goretzko, 2023). If the model is slightly misspecified, a CFA-based approach might not accurately recover the true structure of a model (Nájera et al., 2023). However, organizational researchers often already have a theoretical model in mind and may want to investigate MI on the basis of these theoretical considerations, highlighting the need for a CFA-based approach. Thus, the goals of this article are twofold: On the one hand, we present EFA trees to organizational researchers and explain how they can aid in the theoretical development of constructs and associated measures. On the other hand, we extend EFA trees by introducing CFA trees that allow the incorporation of initially gained theoretical insights into the investigation of MI. We illustrate how both approaches—EFA and CFA trees—can be integrated into the comprehensive workflow suggested by Somaraju et al. (2022) for an extensive and robust assessment of MI. In this, we aim to provide guidance on when to use which version of the trees.

Both EFA and CFA trees can uncover a lack of MI across multiple covariates simultaneously, such as across genders, age groups, or even possible interaction of age and gender. In doing so, EFA and CFA trees provide several benefits beyond extant approaches to test MI (Sterner & Goretzko, 2023): First, they can assess both categorical and continuous sources of MI simultaneously. While the most popular approach to test MI, MG-CFA, struggles with the assessment of categorical sources with many groups or continuous sources, EFA and CFA trees do not suffer from this concern. While there are other approaches that can deal with continuous covariates—especially moderated nonlinear factor analysis (e.g., Bauer, 2017; Bauer et al., 2020)—these approaches require the researcher to select covariates a priori and cannot identify nonlinear patterns and interactions on their own. Therefore, EFA and CFA trees hold promise in revising extant theory and extending future theory by offering new capabilities to test novel hypotheses and research questions associated with many groups or continuous sources.

Second, the study of MNE arising from multiple simultaneous sources and their interactions allows for a more nuanced MI assessments. This benefit enables researchers to assess a larger number of potential biasing influences, but it also allows researchers to test more diverse substantive effects on MI and advance their theoretical, conceptual understanding of a latent concept of interest (see also Sterner, Pargent, et al., 2024). In doing so, more specific tests of MI can be conducted, which hold promise to significantly advance theory associated with studied sources of MI.

Third, most approaches to analyzing MI are conducted in a fully confirmatory manner. Due to the integration of EFA or CFA with decision trees to produce a data-driven approach, the trees can—and are recommended to—be applied in an exploratory manner during the earlier phases of investigation or scale development (Sterner & Goretzko, 2023). By doing so, problematic sources of MI can be addressed before progressing with further studies or the scale development process. As we describe in more detail below, the choice of EFA or CFA tree depends on the application context as well as the level of MI one wants to investigate, enabling our discussed analyses to satisfy a broad range of research needs.

The article is structured as follows: First, we discuss the rationale behind EFA and CFA trees by describing how the trees function and how the two versions differ. Second, we demonstrate their application using publicly available data and provide a step-by-step example on how to apply EFA and CFA trees. Third, we conclude the current article by further detailing the implications of EFA and CFA trees, integrating them in a broader MI assessment workflow building on Somaraju et al. (2022). In doing so, we highlight valuable directions for future research and theoretical development, drawing attention to specific applications of EFA and CFA trees that may be particularly likely to result in meaningful revisions of extant theory or development of novel theory.

Why We Must Think About Measurement Invariance

For valid group comparisons of latent variables, MI has to be established. An invariant scale (or questionnaire) measures a construct with the same factorial structure across all groups that are of interest or all different measurement occasions in a study (Putnick & Bornstein, 2016; Van de Schoot et al., 2012). Accordingly, MI is a core assumption when conducting statistical tests that involve factor scores of latent concepts (or the respective latent means in a SEM)—for example, job satisfaction, commitment, or agreeableness. If MI is violated and the measurement model of a latent concept differs among groups, latent mean comparisons across these groups are often considered to be no longer meaningful[1] (Vandenberg & Lance, 2000). It becomes difficult or even impossible to disentangle true differences on the latent variable from differences in its measurement. In extreme cases, where the noninvariance is particularly pronounced, it may also be questionable if the scale even measures the same construct across the different groups. Hence, researchers need to test MI to ensure that they actually compare quantities that have the same meaning and are measured on the same scale (Somaraju et al., 2022; Vandenberg & Lance, 2000; Van de Schoot et al., 2012).

While establishing MI between two (or more) groups that should be compared with regard to a latent variable is often considered a prerequisite for obtaining unbiased estimates of the respective latent mean differences (e.g., Van De Schoot et al., 2015), it seems also meaningful to assess MI more broadly by exploring various covariates to find out for which subpopulations it does not hold. Such a broad assessment of MI can have multiple benefits:

1. Conducting the respective analyses during scale development can help to define the application context for this scale (e.g., whether it allows to measure some latent variable independently of a person's education).

2. Identifying the actual cause of noninvariance can inform the statistical modeling strategy (for a more detailed

[1] This is especially the case, if configural invariance is violated and factor structures differ across groups. Often, in the presence of small violations of metric invariance, for example, when only very few indicators are affected, researchers consider the measurements still comparable across groups. In this context, the concepts of partial and approximate invariance (e.g., Van De Schoot et al., 2013) are of interest. In this article focusing on the exploration of MI, we will only cover the concept of so-called full invariance, though.

discussion, see Sterner, Pargent, et al., 2024) and helps to disentangle biasing covariates from the grouping variable whose influence on the latent variable should be investigated. For example, there might be an age-related violation of MI and we aim at comparing job satisfaction among countries with different age distributions. In this case, we would falsely attribute the noninvariance to cultural aspects, if we only incorporate the covariate country in our analysis, even though the true cause of noninvariance is simply the age differences between the populations.

3. Besides informing the data analysis or modeling strategy, identifying variables that moderate the measurement model can also be used to learn something substantially about the measurement and the underlying latent concept itself. Therefore, a thorough MI analysis exploring a large set of (categorical and continuous) covariates as well as their interactions advances the conceptual understanding of a construct and supports researchers in enhancing their theories.

How Measurement Invariance is Currently Tested

When new questionnaires for the assessment of psychological constructs are designed, MI should be investigated to ensure that latent means can be compared among groups or different measurement occasions (Vandenberg & Lance, 2000). However, Maassen et al. (2023) showed that the assumption of MI is often not thoroughly tested and found many published scales to lack invariance for different common subgroups (e.g., genders). When researchers decide to test the different steps of MI, though, they usually rely on MG-CFA. MG-CFA uses different constraints to test whether the assumptions of configural, metric, scalar, or residual invariance are met. Since the different constraints yield nested models, models fulfilling subsequent steps of MI can be tested against each other by a likelihood ratio test (Van de Schoot et al., 2012). As the power of such a significance test is dependent on the sample size and the level of noninvariance that should be detected, some researchers argue to use fit index differences when comparing the nested models instead (e.g., Cheung & Rensvold, 2002). As fit indices are also dependent on so-called nuisance parameters such as model complexity, sample size or the amount of missing data (Goretzko et al., 2024), varying recommendations exist which differences between nested models should be considered "significant" in the context of MI testing (Chen, 2007; Meade et al., 2008; Putnick & Bornstein, 2016; Rutkowski & Svetina, 2014).

Independent of whether a significance test or fit index differences are used to determine whether a constrained model has the same fit as an unconstrained model, increasing the number of groups makes the approach more challenging. This is because the number of comparisons will increase drastically and therefore random differences based on sampling error will become more likely. This is also known as α-error inflation. In addition, MG-CFA requires the researcher to define the respective groups for which MI should be tested. That is, it cannot be used with continuous covariates (e.g., age) which would need to be discretized to form testable groups (see also Kim et al., 2017). Furthermore, violations of MI with regard to a different covariate that has not explicitly been used to define the groups that are tested, will also not be detected.

To counter these downsides and to address the issue of MI already during the earliest stages of the questionnaire development process, Sterner and Goretzko (2023) developed EFA trees that extend SEM trees suggested by Brandmaier et al. (2013b). Depending on the application context, a more confirmatory CFA-based method may be applicable, which is why we also introduce CFA trees and discuss both methods and illustrate how and when to use them to address MI more thoroughly. Although there are several other methods that have recently been developed and tackle some of these issues (see also, e.g., Somaraju et al., 2022; Sterner, De Roover, & Goretzko, 2024), such as multilevel factor mixture modeling (Kim et al., 2016), alignment optimization (Asparouhov & Muthén, 2014), or mixture-multigroup EFA (De Roover, 2021; De Roover et al., 2022), these methods are not able to fully explore a set of covariates and detect subgroups that differ with regard to the measurement model (i.e., for which MI is violated). Hence, we focus on EFA and CFA trees in this article, as they excel in this regard and ultimately provide greater benefits to tests of MI than these other approaches.

EFA and CFA Trees

Sterner and Goretzko (2023) proposed EFA trees as a new data-driven approach to explore MI for various categorical and continuous covariates (i.e., numerous groups that do not have to be specified by the user). The underlying idea of EFA trees is to combine EFA with model-based recursive partitioning (MOB) which has already been used in combination with item response models to detect differential item functioning in Rasch models (Strobl et al., 2015), 2PL- and 3PL models (Debelak & Strobl, 2019), their polytomous extensions (Komboz et al., 2018), and other item response theory models (Schneider et al., 2021). It has also been applied to the classical test theory framework and structural equation models (Brandmaier et al., 2013a, 2013b, 2016, 2018).

Accordingly, a more constrained factor model can also be used in a semiconfirmatory approach that we call CFA trees. EFA trees are suitable for MI exploration at the earliest stages of questionnaire development and therefore allow researchers to fully exploit the exploratory potential of the MOB approach (Goretzko & Bühner, 2022). CFA trees are also exploratory in nature to an extent, as they do not require the user to specify for which groups or even covariates MI is potentially violated. Contrary to EFA trees, specific constraints can be incorporated in the model, such as following an independent clusters assumption allowing each indicator to only load on one factor. In turn, CFA trees can be used to reassess a model that has been initially supported with alternative samples and/or analyses. In addition, as we will describe in more detail below, the two versions focus on different levels of MI: EFA trees are especially suitable for an investigation of metric MI (i.e., invariance of main- and cross-loadings), whereas CFA trees can be used (subsequently) to assess scalar MI (i.e., invariance of intercepts).

The Common Factor Model—EFA and CFA

The statistical model at the core of EFA and CFA trees is a common factor(s) model that represents the relations among unobservable, latent factors (e.g., job satisfaction) and observed indicators. The central idea is that the values of the measured variables (e.g., the answers to questionnaire items) can be explained by one or more latent variables and that the underlying latent

variables are causing variations in these measured variables. A linear relationship between each item (x_i) and latent factor (ξ_j) is assumed:

$$x_i = \tau_i + \lambda_{i1}\xi_1 + \lambda_{i2}\xi_2 + \cdots + \lambda_{ik}\xi_k + \epsilon_i, \quad (1)$$

with τ_i being the item intercepts and ϵ_i being an error term because the latent factors do not explain all the variance in the observed variables—in other words, there is measurement error, if we consider the items as measures of a latent concept.[2] The loadings (λ_{ij}) can be compared to regression slopes and reflect the strength of the relationship between a latent variable ξ_j and an indicator x_i. In EFA, researchers try to identify a loading pattern, that is, determine which indicators belong to which latent variable, which is why all loading parameters λ_{ij} are freely estimated (except from parameters that need to be constrained for estimation purposes). The intercepts are usually not considered in an EFA model (due to standardization of the data). In CFA, often based on results of previous EFAs, several loading parameters λ_{ij} are fixed to zero and a hypothesized measurement model is tested. In addition, the intercepts can be included in the model. This already highlights an important difference between the two versions of the trees: EFA trees can uncover noninvariance due to differences in primary and cross-loadings, whereas CFA-trees are limited to primary loadings (and selected cross-loadings when the independent clusters assumption is softened), but can uncover noninvariance due to intercept differences.

Independent of the modeling strategy, all loadings need to be equal across groups if metric MI holds. In addition, the intercepts need to be equivalent across groups for scalar MI to hold. If that is not the case, one or several indicators do not measure the same latent factors or do not reflect them in the same way for different subpopulations (e.g., different cultures or age groups) which makes it difficult to compare the latent variable of interest (e.g., job satisfaction) among these groups. Therefore, EFA and CFA trees aim at finding subsamples in the data for which these parameters differ which in turn means finding subsamples for which the current item set is noninvariant. By combining the two methods, the data can be scrutinized for all forms of metric MI (EFA trees) and scalar MI (CFA trees).

MOB

To find these subsets, MOB is applied (Zeileis & Hornik, 2007; Zeileis et al., 2008). The central idea of MOB is to build a decision tree that divides a sample into subsamples that differ with regard to their model parameters. This general idea can be combined with any parametric model, such as item response theory models (e.g., Strobl et al., 2015) or SEM (e.g., Brandmaier et al., 2013b), but also with statistical models that do not rely on latent variables, for example, generalized linear mixed models (Fokkema et al., 2018). In all cases, the tree structure is used to uncover differences in model parameters across various groups that are defined by different covariates (categorical or continuous). The basic steps of MOB are:

1. The model of interest (in our case, an EFA or CFA) is fitted on the full data set.
2. So-called structural change tests are performed to see whether parameter instabilities given the covariates occur. That is, the algorithm evaluates whether a covariate from the set of tested covariates is associated with model parameter differences and whether splitting the data set according to this variable improves model fit. If the tests of more than one covariate are deemed significant, the covariate with the smallest p value (below a specified level of significance) is chosen as a "split variable" that is used to divide the data set into subsets.
3. Then, the optimal split point for the selected covariate is found using an exhaustive search trying out all potential splits and choosing the one that separates the resulting subsets best (in terms of model fit) with regard to the respective model parameters.
4. These steps are repeated until no significant tests are observed or a stopping criterion (e.g., the maximum tree depth or the minimum node size in the leaf nodes, i.e., the minimum subsample size) is reached.

To avoid inflated Type I error rates ("α-error inflation"), a Bonferroni correction is applied at tree level which means that the family-wise error rate is limited for the whole EFA or CFA tree at the prespecified α-level. In simulation studies, Sterner and Goretzko (2023) also showed that for extremely large sample sizes (n = 10,000), the Type I error rate does not exceed the selected α. Readers who want further details about the mechanisms of MOB are referred to the works by Merkle et al. (2014), Zeileis and Hornik (2007), and Zeileis et al. (2008) as well as the article introducing EFA trees by Sterner and Goretzko.

Combining EFA or CFA and MOB

As described above, using MOB and the resulting tree structures to detect subsamples with varying model parameters can be applied to various different models. Accordingly, it has already been integrated within the context of latent variable modeling (e.g., Brandmaier et al., 2013b; Strobl et al., 2015). Hence, using the MOB framework in combination with EFA or CFA to detect violations of MI can be seen as a straightforward extension of the previous approaches (Goretzko & Bühner, 2022; Sterner & Goretzko, 2023). The idea of the EFA and CFA trees approach, therefore, can be described as follows:

1. A factor model (EFA or CFA, depending on the respective stage of scale development or research question) is fitted to the full sample. Sterner and Goretzko (2023) provided template code of a fitting function for EFA trees that is used internally in the "tree growing" function. We complement this template code by changing the model that is used within the MOB function to enable the fitting of CFA trees (available at https://osf.io/acy7x).[3]
2. Then, structural change tests explore various covariates of interest (e.g., age, gender, education, cultural background, organization size, industrial sector, etc.) to determine whether

[2] That is the conceptual difference between EFA and the related principal component analysis which does not consider measurement error and, therefore, in a narrow sense, is not a tool for the development of measurement models but for dimensionality reduction (e.g., Fabrigar et al., 1999; Goretzko et al., 2021; Howard, 2023).

[3] For readers familiar with *lavaan* language: both versions of the trees can be fitted using the *cfa()* function. For CFA trees, we set *meanstructure* = *TRUE* to incorporate intercepts.

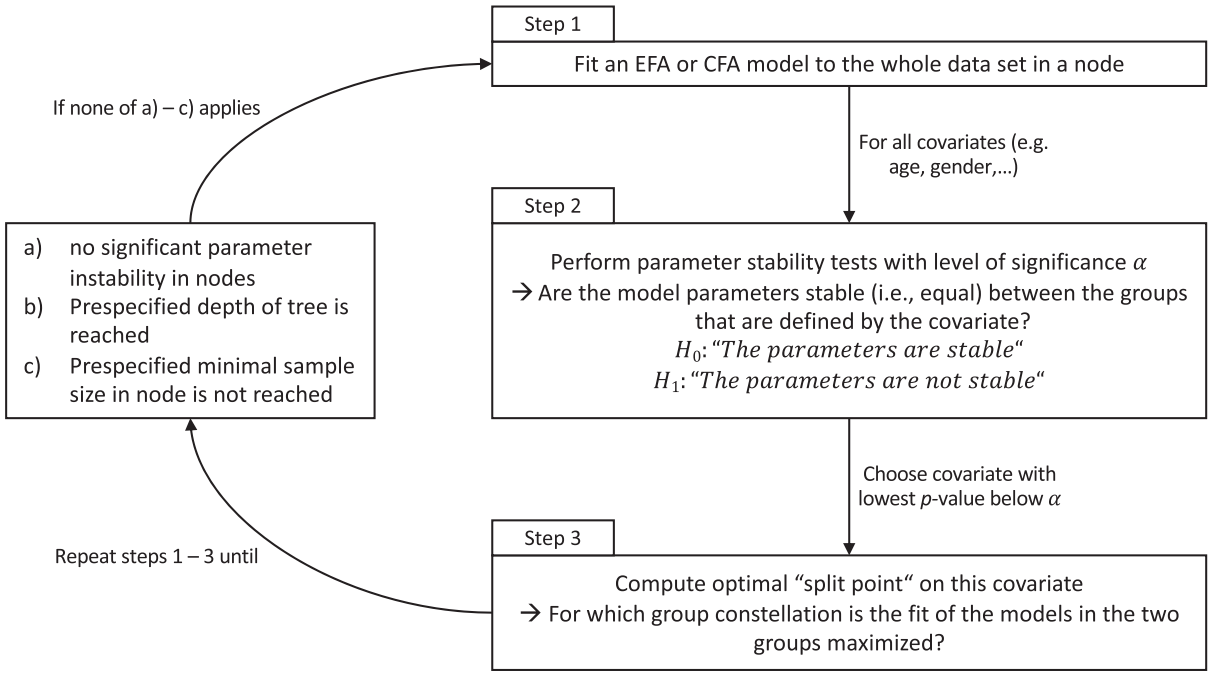
the model parameters—especially the loading parameters—are unstable (i.e., differ among potential subgroups) which would indicate noninvariance. Within the MOB framework implemented in the partykit package (Hothorn & Zeileis, 2015), generalized M-fluctuation tests (see, Zeileis & Hornik, 2007) are used to test the structural change of model parameters. The central idea of this test is that the scores (i.e., the values of the partial derivative of the log-likelihood function) of each observation evaluated at the respective parameter estimates should randomly fluctuate around zero as the likelihood is maximal for the parameter estimates ("maximum likelihood estimation") and its derivative becomes zero. If a covariate is associated with parameter instabilities (i.e., can be used to define noninvariant subpopulations), these individual scores do not fluctuate randomly around zero any more and the cumulative scores (cumulative sum of scores for individuals sorted by the respective covariate values) reveal patterns of parameter instability (e.g., only negative scores for individuals who are younger than 30 years old). A more thorough description of the structural change tests in MOB can be found in Zeileis et al. (2008), while Arnold et al. (2021) nicely illustrated the idea of score-based tests in the context of SEM trees and Strobl et al. (2015) do so for item response theory (Rasch) models.

3. If a splitting variable is selected, the optimal split point for this variable is detected by maximizing the partitioned likelihood (i.e., the sum of the two likelihoods on both sides of the split point) which requires the algorithm to go through the sorted covariate and try out all potential split points. Then, the data set is partitioned into two subsets according to the respective split point.
4. The process is repeated in each resulting subset until no more significant parameter instabilities are found or other stopping criteria are met (e.g., the maximal tree depth or minimum number of observations that need to remain in a subset can be specified to prevent the tree from growing too deep).
5. In each leaf node (i.e., in each subsample) an EFA or CFA is fitted separately, and the different loading patterns (or other model parameters) can be explored to find out which items are the drivers of a noninvariant measurement model.

A flowchart of these steps is presented in Figure 1. Accordingly, if MI is violated, EFA and CFA trees will provide a tree that divides the sample into subsamples that differ regarding their measurement models according to the chosen split variables. For each group, a separate factor model is estimated, so that a further inspection of these models will reveal the type of noninvariance, the severity of it (i.e., how large the differences in model parameters are), and especially which items are affected and therefore may need to be revised in the scale construction process. It should be kept in mind that the sample sizes in the leaf nodes must be large enough to allow for accurate model estimation if researchers want to interpret the parameter estimates. Nonetheless, EFA and CFA trees can also be useful for smaller sample sizes if the focus lies on simply detecting noninvariance. The simulation study by Sterner and Goretzko (2023) showed that the power to detect noninvariance is also high in small

Figure 1

Flowchart of the Steps Performed by the Model-Based Recursive Partitioning Algorithm in Combination With EFA



Note. EFA = exploratory factor analysis; CFA = confirmatory factor analysis.

sample cases. However, for a detailed evaluation of model parameter estimates beyond simple tests of noninvariance, we would recommend to set the minimum sample size in the leaf nodes to ≥250 observations, if the total sample size allows it.

One concern with exploratory methods like EFA or CFA trees is the possibility to capitalize on chance or overfit the data at hand. Despite its similarity to the tree-growing algorithms that are used in machine learning contexts—such as the classical Classification and Regression Trees algorithm (Breiman et al., 1984), MOB is not concerned with maximizing predictive accuracy and rather takes a significance testing perspective. Hence, approaches such as correcting the *p* values of the internal structural change tests (see above) are used to ensure valid inference (i.e., limiting the Type I error rate at tree level) and are preferred over cross-validation or comparable resampling strategies.

Rotational Indeterminacy

When relying on an EFA-based approach, the challenge of rotational indeterminacy, that is that factor solutions are only determined up to an admissible rotation (e.g., Browne, 2001), emerges. While several rotation methods can be chosen to obtain an interpretable EFA solution (e.g., Browne, 2001; Goretzko et al., 2021), the most appropriate rotation technique may differ among the subgroups detected by an EFA tree which hampers the comparability of the obtained factor solutions. Accordingly, selecting one rotation method for all terminal nodes of the EFA tree may not always be feasible.

Sterner and Goretzko (2023), therefore, suggested to use regularization to obtain sparse and interpretable factor solutions. Instead of using a two-step approach—estimating a primary factor solution and subsequently rotating it for interpretability—regularized EFA incorporates a penalty in the estimation function that shrinks small loading parameters toward zero (e.g., Goretzko, 2023; Scharf & Nestler, 2019). In doing so, negligible cross-loadings will become zero (or at least close to zero) which serves the same purpose as factor rotation[4]. To increase comparability between groups, alignment methods can be used (Asparouhov & Muthén, 2014, 2023) which work similarly as rotation-to-target methods (e.g., Browne, 2001). As an alternative, De Roover and Vermunt (2019) proposed multigroup factor rotation that combines classical factor rotation with alignment between groups. Instead of rotating the factor solution to solely maximize interpretability by approaching a simple structure, multigroup factor rotation also tries to maximize comparability at the same time. To achieve this, a joint rotation and agreement criterion is optimized, where a weighting of the subgoals can be determined by the researcher (per default both criteria are weighted equally; for further details, see De Roover & Vermunt, 2019; Sterner, De Roover, & Goretzko, 2024).

In the exemplary data analysis (see below), we simplified our analyses and applied Geomin rotation in all terminal nodes of our EFA tree, as it produces interpretable factor solutions for all subsets. We also demonstrate how alignment optimization (Asparouhov & Muthén, 2014) can be used to make the resulting loading patterns comparable. However, as rotational indeterminacy remains a challenge in EFA-based analyses, different analysis decisions, especially selecting a different rotation method, may be justifiable as well.

Data Example

To facilitate the use of EFA and CFA trees for applied researchers, we now demonstrate their application on exemplary data from organizational psychology. For this illustration, we chose data from the Grit Scale (Duckworth et al., 2007). In the additional online supplemental material (https://osf.io/acy7x), we present a second analysis using data from the Utrecht Work Engagement Scale (Schaufeli et al., 2006), where the tree does not split the data, that is, where measurement invariance is not violated.

Transparency and Openness

We adhered to the Journal of Applied Psychology methodological checklist, and we fully describe our sampling plan, all data exclusions (if any), all manipulations, and all measures in the study. All analysis code, and research materials are available at https://osf.io/acy7x. Data are openly available in online repositories (links are presented below). Data were analyzed using *R* (R Core Team, 2021) and the packages *lavaan* (Rosseel, 2012), *partykit* (Hothorn & Zeileis, 2015), *psych* (Revelle, 2024), as well as *semTools* (Jorgensen et al., 2022). The article was written in R markdown using the package *papaja* (Aust & Barth, 2020). This study's design and its analysis were not preregistered.

Grit Scale

To demonstrate how to proceed once EFA trees split the data and to show how they can be followed up by CFA trees, we applied these analyses to data collected by administering the Grit Scale. This twelve-item scale measures grit—a noncognitive trait defined as perseverance and passion for long-term goals—on two subscales, Consistency of Interests and Perseverance of Effort (Duckworth et al., 2007). For the analysis, we used publicly available data of $N = 4{,}270$ participants (retrieved from https://openpsychometrics.org/_rawdata/). These data contain (among other variables) answers to the Grit Scale as well as the covariates (nominal) gender, (continuous) age, and (ordinal) education. We only included participants who identified themselves as "male" or "female" on the covariate gender because the other two options both had less than 30 observations. In addition, we excluded participants who indicated their age to be under 18 (to only include adults) or over 78 (two participants indicated the implausible ages of 228 and 350). This resulted in a final sample size of $N = 3{,}151$. Because this sample size is quite large, we set the minimum sample size required in each node of the trees to 500 to allow for accurate parameter estimation in each subset. As level of significance we used $\alpha = .05$. We limited the tree depth to a maximum of two splits to keep the results interpretable.

EFA Tree

First, we grew an EFA tree to test for metric MI, specifically for the invariance of both main- and cross-loadings. In the EFA tree, typical constraints for estimating exploratory factor models were imposed—all indicators were standardized, all cross-loadings were estimated, and between-factor correlations were set to zero (initial and unrotated solution). Table 1 shows the results of the hypothesis tests in the parent node of an EFA tree applied to the empirical data. Two covariates, gender and age, showed a *p* value below the level of

[4] In multidimensional models, factor correlations and cross-loadings cannot be disentangled in a data-driven way, which is why factor correlations need to be considered as well when assessing MI to avoid a dependence on the chosen factor rotation or regularization technique (for a thorough discussion, see Thissen, 2024).

Table 1

Hypothesis Test Result in the Parent Node of the Exploratory Factor Analysis Tree Applied to the Grit Scale

Statistic	Gender	Education	Age
Test statistic	55.593	175.427	155.804
p	.044	.067	.000

Note. Test statistics were a χ^2 statistic for categorical and a supLM statistic for continuous covariates. A *p* value of 0 indicates that it is <.001.

significance. The EFA tree thus chose the covariate with the smaller *p* value (i.e., age) for splitting the data. The optimal split point for this node was at the age of 26, resulting in a node with participants who are 26 or younger ($n_{age \leq 26} = 1{,}905$) and a node with participants who are older than 26 years. In this "older" node, the EFA tree split the data again, this time at age 38 ($n_{26<age \leq 38} = 642$ and $n_{age>38} = 604$).[5] The EFA tree does identify which parameters differ across groups. For this, we must extract the models from the leaf nodes (i.e., the subsamples) and investigate their parameter estimates, for example, the loadings. Before doing so, however, we must test whether there are the same number of latent factors in all groups (i.e., whether the most basic form of configural invariance holds). Sterner and Goretzko (2023) recommended to conduct a parallel analysis (Horn, 1965) in each leaf node. Simply put, a parallel analysis compares the eigenvalues generated from our data set to eigenvalues generated from randomly simulated data. The number of eigenvalues larger than the 95th-percentile of their randomly generated counterparts is the number of factors that should be extracted. Here, in each leaf node, the parallel analysis suggested two latent factors, which is also in line with the theory behind the Grit Scale.

Table 2 shows the loading matrices of the three models in the leaf nodes. As mentioned above, different strategies to achieve interpretable loading matrices are possible, for example, regularization, multigroup factor rotation, and simple structure rotation (e.g., varimax, oblimin, or geomin, see also Browne, 2001). Due to its simplicity in interpretation and availability in common statistical software, we chose to apply geomin rotation for this exemplary analysis. Upon visual inspection, we can identify differences that probably led the tree to splitting the data: Group 2, compared to Groups 1 and 3, shows higher cross-loadings on Item 4 (Setbacks do not discourage me) and Item 8 (I have difficulty maintaining my focus on projects that take more than a few months to complete). That is, for people above 26 but below 38 years of age, both factors consistency of interests and perseverance of effort contribute more to variation in answers to these items (unlike in Groups 1 and 3, where the item answers are driven more strongly by one factor). In addition, in Group 3, compared to Groups 1 and 2, both the main and the cross-loading of Item 12 ("I am diligent") are lower. Thus, the scores on both factors are less related to the answers on Item 12 for people over the age of 38, compared to younger participants. As can be seen, investigating the results of an EFA tree is not much different than when investigating only one loading matrix of a single EFA. It is more informative, though, because information is provided on how measurements of a construct differ between various groups. The theory behind the concept of grit could now be used to help rephrase items, such that the loadings become invariant between groups. Alternatively, one could aim for partial invariance (i.e., achieve at least some invariant loadings) or remove the noninvariant items (given that this does not alter the meaning of the measured concept; Somaraju et al., 2022).

Table 2

Factor Solution of the Grit Scale (Standardized Loading Patterns) Based on an Exploratory Factor Analysis Tree With Three Terminal Nodes

	Group 1 (age ≤ 26)		Group 2 (26 < age ≤ 38)		Group 3 (age > 38)	
Item	CI	PE	CI	PE	CI	PE
Item1	0.13	0.62	0.10	0.62	0.10	0.56
Item2	0.05	0.47	0.01	0.62	0.11	0.51
Item3	−0.03	0.73	−0.02	0.57	0.00	0.51
Item4	−0.30	0.59	−0.51	0.52	−0.33	0.47
Item5	−0.22	0.71	−0.21	0.71	−0.18	0.51
Item6	0.00	0.71	−0.11	0.60	−0.11	0.59
Item7	0.65	−0.01	0.74	−0.05	0.71	−0.01
Item8	0.85	0.15	0.94	0.30	0.68	0.13
Item9	0.89	−0.08	1.00	−0.02	0.83	0.02
Item10	0.94	−0.02	0.92	−0.02	0.82	−0.07
Item11	0.79	−0.28	0.90	−0.22	0.74	−0.29
Item12	0.72	0.32	0.83	0.35	0.58	0.17

Note. CI and PE denote the latent factors consistency of interests and perseverance of effort, respectively. The factor solutions were obtained after a geomin rotation.

CFA Tree

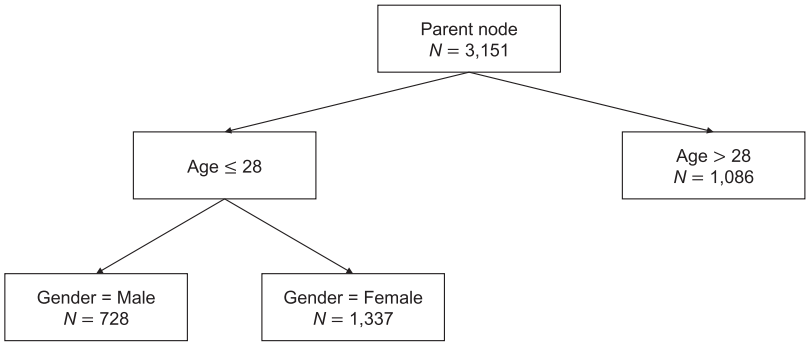
If researchers have a theoretically or empirically justified measurement model available, these considerations can be incorporated by growing a CFA tree instead of an EFA tree. Alternatively, an EFA tree can be followed up by a CFA tree, if the goal is to evaluate a newly uncovered measurement model (ideally on a new data set). As mentioned, while EFA trees allow for a more detailed investigation of metric MI, CFA trees can be applied to assess scalar MI (i.e., the invariance of intercepts). To grow a CFA tree, the mean structure is included in the model so that the intercepts are estimated and considered by the tree when assessing parameter instability. In addition, researchers can fix all parameters except the intercepts to the estimates of a single-group CFA and specify the parm argument in the *mob_control*() function to only consider intercepts when assessing parameter instability. This way, when then growing a CFA tree, only the intercepts are estimated freely and evaluated for splitting. For identification purposes, the latent means are set to zero and latent variances are constrained to one. As grouping variables are not known in advance (compared to a MG-CFA, for example) latent means and variances cannot be fixed in a reference group and freely estimated in other groups. Accordingly, CFA tree results need to be followed by a completely confirmatory approach (e.g., MG-CFA) to disentangle violations of MI from true differences in latent means or variances. In the following, we illustrate the CFA tree approach using the Grit Scale data, although the EFA tree analysis already suggested that metric MI is violated (see above).

Figure 2 shows the resulting CFA tree structure when assuming a theoretical measurement model without cross-loadings. The tree

[5] Note that it is a coincidence that the tree split twice on the covariate age. The second split could have also been on the covariates gender and education. This would have indicated an interaction between the two split covariates associated with a violation of MI.

Figure 2

Resulting Partition After Applying a Confirmatory Factor Analysis Tree to the Grit Scale Data



first split the data into two different age groups, a group which is 28 years or younger, and a group which is older than 28 (which is a first split point similar to the one of the EFA trees which split the data at age 26). In the "older" node (where age > 28, $n = 1{,}086$), the CFA tree did not split the data further. In the "younger" node, we can now see an interaction between two covariates that is associated with a violation of MI: The CFA tree splits the data on the covariate gender, resulting in a "younger male" ($n = 728$) and a "younger female" ($n = 1{,}337$) node. Again, we must inspect the models in the nodes to learn more about why the tree might have split the data. We already investigated the loading matrices of the EFA tree (which paint a more detailed picture due to the consideration of cross-loadings). In addition, in the CFA tree, all loadings were constrained to be equal across groups to the estimates of a single-group CFA. Thus, we do not present the loading matrices of the CFA tree.[6]

Instead, we focus on the investigation of scalar MI, that is, the invariance of intercepts. The intercepts (the parameters we want to compare for invariance) are intertwined with the latent means (the parameters of interest in our substantive analysis). That is, even if the intercepts were equal across groups but the groups differ in their latent means (here: in their latent grit), these differences would show up as intercept differences in the model and the CFA tree would still split the data due to the aforementioned identification constraints. To assess whether the CFA tree split the data due to noninvariance of item intercepts, we recommend to conduct a MG-CFA with the resulting nodes as groups. This lets us disentangle whether the split occurred due to noninvariance (i.e., measurement differences) or due to true differences.

Table 3 shows the results of this MG-CFA when using the nodes of the CFA tree as groups, setting the latent means to zero for the group of younger males, and selecting the first indicator per scale as an anchor item for scaling. The χ^2 difference tests indicate that both metric and scalar MI are violated. The differences in fit indices suggest that metric MI is supported (which was also assumend when growing the CFA tree), whereas the comparative fit index suggests that scalar MI is violated (due to a decrease in fit from the metric to the scalar model of 0.014). Thus, we can conclude that the CFA tree most likely split the data due to differences in intercepts and not only due to true differences in latent grit. However, researchers need to be cautious when interpreting the intercept parameters across nodes (e.g., in Table 4) as differences may also be conflated with latent mean differences. Accordingly, we advise them to also investigate the intercept differences in a subsequent MG-CFA.

Table 4 shows the estimated intercepts in each node. Upon visual inspection of the differences between nodes, we can see that mostly the intercepts of the older node differ from the two younger nodes. For items of the factor perseverance of effort, intercepts of the older node are systematically lower, while for items of the factor consistency of interests, they are systematically higher. In comparison, the intercepts between the two younger nodes do not seem to differ drastically. As with EFA trees, substantive knowledge about the theory behind the Grit Scale could now be used to reason about why intercepts differ between younger and older participants. Items could be rephrased or dropped accordingly, in order to establish MI between nodes.

So far, we have mainly demonstrated that MI is violated between some groups (e.g., younger male and female, and older participants in the CFA tree example). While the reported fit indices can be seen as effect sizes, they contain only little information about the influence of noninvariance on substantive comparisons of grit between groups. In addition, they are sensitive to sample size and other parameters, like model or loading size (e.g., Goretzko et al., 2024). To assess the influence of noninvariance on our analyses of interest, other effect size measures can be calculated. Nye and Drasgow (2011) introduced d_{MACS} (MACS: mean and covariance structure) as a sample size-independent effect size measure. d_{MACS} quantifies the contribution of noninvariance to expected score differences for each item. Similar to other effect sizes, the metric of d_{MACS} is pooled standard deviations between a reference and a focal group. In addition, a combined measure can be calculated that quantifies the influence of noninvariance on the mean of the total scale (consistency of interests and perseverance of effort, in our case).

Table 5 shows the d_{MACS} values for each item, where the younger male group is the reference group, and the younger female and the older group are the focal groups. When comparing younger male and younger female participants, effect sizes of noninvariance for the factor perseverance of effort range from negligible (0.02) to small (0.25), and for the factor consistency of interests from 0.06 to

[6] When running a CFA tree focusing on metric noninvariance of primary loadings using standardized variables and not including the mean structure, data splits were similar to that of an EFA tree with all cross-loadings (data were split at age 28 instead of ages 26 and 38, though). The full analysis is presented in the additional online supplementalry material (https://osf.io/acy7x).

Table 3

Results of Multigroup Confirmatory Factor Analysis Applied to the Grit Scale With Nodes of the Confirmatory Factor Analysis Tree as Groups

Model	$\Delta\chi^2$	Δdf	p	CFI	RMSEA	ΔCFI	ΔRMSEA
Configural	NA	NA	NA	0.89	0.09	NA	NA
Metric	44.64	20	.00	0.89	0.08	0.00	0.00
Scalar	186.49	20	.00	0.88	0.08	−0.01	0.00

Note. $\Delta\chi^2$ = difference in test statistics between two consecutive models; Δdf = difference in degrees of freedom between two consecutive models; ΔCFI = difference in comparative fit index between two consecutive models; ΔRMSEA = difference in root-mean-square error of approximation between two consecutive models; NA = not Applicable. A p value of 0 indicates that it is <.001.

0.10, which might be considered negligible, too. The influence of noninvariance on the mean of the total scale is −0.24 and 0.23 for perseverance of effort and consistency of interests, respectively, indicating a small effect of noninvariance. When comparing younger male and older participants, effect sizes of noninvariance for the factor perseverance of effort range from negligible (0.05) to medium (0.58), and for the factor consistency of interests from small (0.23) to medium (0.42). The influence of noninvariance on the mean of the total scale (i.e., the estimated bias in raw scores) is −2.10 and 2.21 for perseverance of effort and consistency of interests, respectively, which could be considered a very large effect of noninvariance given that the pooled standard deviations of the raw scores are around 4.5 and 5.3, respectively. Consequently, comparisons between younger male and younger female participants could potentially be made with caution, whereas comparisons with older participants should be avoided.

In their MI-workflow, Somaraju et al. (2022) also suggested using the alignment approach in confirmatory analyses. Alignment (Asparouhov & Muthén, 2014) aims at achieving comparable latent means when full MI is not supported; that is, when there are some small differences in parameter values. As indicated by the d_{MACS} values, the noninvariance of some items might distort our inference with latent means. Alignment helps us with the identification of noninvariant items as it does not require the selection of invariant anchor items contrary to MG-CFA and, therefore, enables researchers to untie the knot of intertwined latent means and scalar noninvariance. To highlight the connectivity of EFA and CFA trees to existing methods (see also Sterner, De Roover, & Goretzko, 2024), we applied alignment to the three nodes of the presented CFA tree. The full Mplus output of the alignment is available at https://osf.io/acy7x.

Table 4

Intercepts of the Models in the Leaf Nodes of the Confirmatory Factor Analysis Tree for the Grit Scale

Factor	Item	Younger male	Younger female	Older
PE	1	2.11	2.14	1.87
	4	2.61	2.90	2.60
	6	2.12	1.88	1.58
	9	2.51	2.47	2.29
	10	2.65	2.60	2.05
	12	2.39	2.17	1.90
CI	2	2.38	2.46	2.64
	3	2.88	2.81	3.19
	5	2.64	2.75	3.15
	7	2.93	2.98	3.28
	8	2.94	2.90	3.34
	11	2.67	2.77	3.06

Note. CI and PE denote the latent factors consistency of interests and perseverance of effort, respectively.

Most notably, intercepts were noninvariant on Items 4 and 6 in the younger male group (compared to younger female and older participants), on Item 9 in the older group, and on Item 12 in the younger female group (all items belong to the factor perseverance of effort). Regarding loadings, only the loading of Item 6 was noninvariant in the older group (compared to the two younger groups). The aligned factor intercepts showed that on both factors, the older group differed significantly from the two younger groups in their latent means, with significantly higher values in consistency of interests (older: 0.442; younger female: 0.037; younger male as reference group) and significantly lower values in perseverance of effort (older: −0.579; younger female: −0.061; younger male as reference group). The average overall invariance index was $R^2 = 0.631$ (ranging from 0, completely noninvariant, to 1, completely invariant)[7]. To summarize, results of the alignment procedure show that most parameters are invariant between nodes after optimizing for alignment. However, full invariance is not given (due to the rather low R^2). This was to be expected because the groups that entered the alignment analysis resulted from a CFA tree which is designed to identify noninvariant groups and the presented effect size measures suggest a substantial amount of noninvariance that alignment optimization cannot (fully) resolve. Nonetheless, if we are interested in comparing latent means between the groups that result from the CFA tree, alignment is a helpful follow-up analysis to account for smaller noninvariances between the groups in the nodes making them more comparable (cf. Sterner, De Roover, & Goretzko, 2024).

We hope that this empirical example provides a starting point for organizational researchers in determining when to use which version of the trees. As we have shown, for an even more thorough investigation of MI, the trees can be combined by growing a CFA tree after an EFA tree. Both EFA and CFA trees can be seamlessly integrated into the workflow by Somaraju et al. (2022; see Figure 3) by acting as a precursor of MG-CFA or alignment optimization, as demonstrated above. Paired with subject matter expertise, this allows for a detailed assessment of how measures of a construct function across groups that do not have to be defined in advance but are found in a data-driven way.

Discussion

Organizational researchers regularly perform analyses that assess the relations of constructs across groups, which has become more popular in recent years with the widespread application of more sophisticated research designs to assess concepts like multilevel emergence or causality (Cho et al., 2023; Jebb & Tay, 2017;

[7] According to Asparouhov and Muthén (2014), this metric has to be seen as a "rough measure" of the level of noninvariance. Calculating this value for intercept parameters, for example, requires the use of loading parameter estimates, which is why noninvariance of loading parameters can influence its value also for intercept parameters. However, it provides a proxy for which items are noninvariant across groups and allows for evaluating the aligned factor solution. Interested readers can find more information on how it is calculated in Asparouhov and Muthén.

Table 5
d_{macs} Values for All Items With Younger Male Participants as Reference Group in the Confirmatory Factor Analysis Model

	Younger female		Older	
Item	PE	CI	PE	CI
GS1	0.02	NA	0.26	NA
GS4	0.25	NA	0.05	NA
GS6	0.23	NA	0.58	NA
GS9	0.07	NA	0.20	NA
GS10	0.07	NA	0.48	NA
GS12	0.22	NA	0.49	NA
GS2	NA	0.08	NA	0.23
GS3	NA	0.07	NA	0.33
GS5	NA	0.08	NA	0.42
GS7	NA	0.04	NA	0.35
GS8	NA	0.06	NA	0.32
GS11	NA	0.10	NA	0.35

Note. CI and PE denote the latent factors consistency of interests and perseverance of effort, respectively. GS = Grid Scale; NA = not applicable.

Lu, 2023; Podsakoff et al., 2019). To ensure that inferences derived from the applied measures are appropriate across the studied groups and/or measurement occasions, researchers have likewise increasingly applied analyses to assess MI and developed theoretical arguments for the sources of different types of MI. EFA and CFA trees address several statistical limitations of common approaches enabling the assessment of MI in the earlier phases of an investigation—such that researchers can correct any emergent issues earlier in the research process. Due to their exploratory nature, EFA and CFA trees should not be understood as strict tests that tell researchers whether certain group comparisons are warranted or not, but rather as tools to learn more about the measurement process, how the scales and indicators function, and potentially about the latent construct itself.

That said, such exploratory and data-driven approaches will not replace confirmatory analyses like MG-CFA but rather serve as additional tools to gain a deeper understanding of mechanisms affecting the measurement of a variable of interest. As Fischer and Rudnev (2024) argue, investigating sources of measurement noninvariance can be a stand-alone research project. We believe that exploratory tools such as the presented EFA and CFA trees can support organizational researchers with such endeavors. In that regard, we emphasize that the usage and interpretation of these methods clearly depend on the research goal and state of the scale development process. Not every violation of MI requires an immediate reaction such as removing items from the item set, rephrasing items, or even exchanging the full scale. The thorough exploration of the measurement model and potential moderating or biasing effects of covariates rather enables the researcher to make adequate modeling decisions (see also Sterner, De Roover, & Goretzko, 2024) and gain a deeper understanding of how different (sub-)populations can be assessed (Fischer & Rudnev, 2024). However, when researchers want to formally test whether measures are invariant between specific groups, other methods (e.g., MG-CFA) still need to be used.

The empirical example illustrates how EFA and CFA trees can be applied and combined in a thorough MI workflow, especially as an initial screening tool. Our results showed that EFA and CFA trees can be effective at identifying violations of MI in a data-driven way. While confirmatory approaches to formally test MI are still necessary to fully ensure MI across groups, we suggest that EFA and CFA trees are a valuable addition to the organizational researcher's methodological toolbox (see also Figure 3).

Integration of EFA and CFA Trees in MI Testing Workflows

As Somaraju et al. (2022) pointed out, there is no MI methodology that can be used for all purposes and under every circumstance. Researchers must combine different approaches to properly test for MI, explore violations of it, and identify its sources as well as the (theoretical) reasons behind the nonequivalence of measurements. Somaraju et al. suggested a detailed workflow to guide researchers through the MI testing procedure and recommended next steps depending on their overall aim. This concerns whether researchers want to detect noninvariance or identify sources of noninvariance as well as the outcome of the respective analyses (i.e., the steps within the workflow). As MG-CFA plays a central role in their workflow, it is a confirmatory approach and requires established/hypothesized measurement models and a definition of groups for which MI is tested.

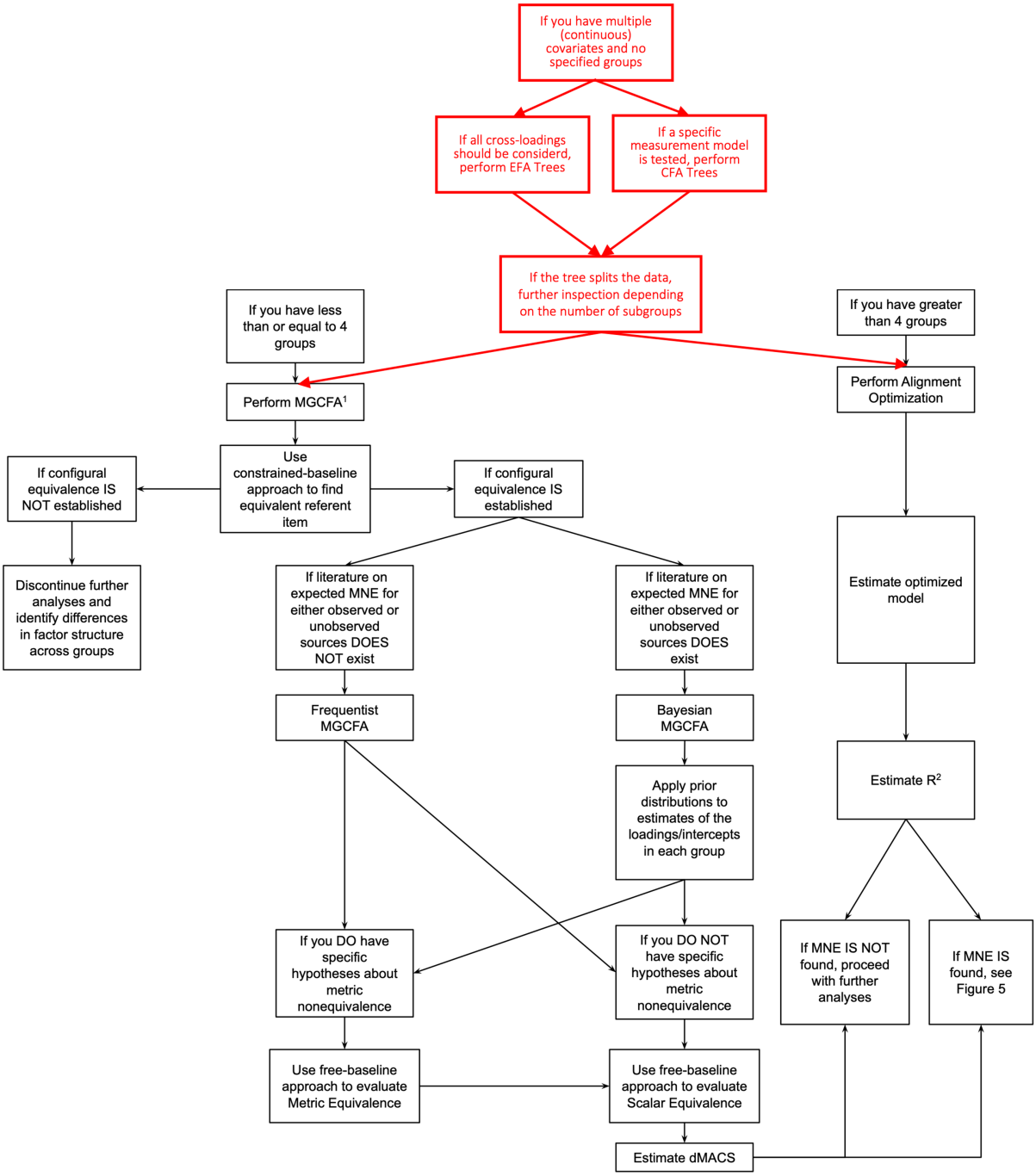
EFA and CFA trees could be used as an alternative first step in such a workflow, if researchers want to apply it in a more exploratory way. For example, EFA and CFA trees should be applied first, if researchers are interested in identifying subgroups with varying measurement models in a purely data-driven manner, or if they are at the early stages of scale development and the measurement model is still "under construction." EFA and CFA trees may also be used as an initial screening tool before applying MG-CFA (or alignment optimization if many subgroups with varying measurement models have been identified) to validate their exploratory results (ideally on a separate data sample). That is, EFA or CFA trees provide the respective groups for which MI is potentially violated (see also the empirical example above). MG-CFA and alignment optimization are then used to confirm this nonequivalence and help to identify the sources of it. Subsequently, in case of metric noninvariance, for example, Multiple Indicators and Multiple Causes models (Muthén, 1989) are used to further explore the specific sources and degree of this violation of metric MI. In an alternative scenario with numerous subgroups, alignment optimization (Asparouhov & Muthén, 2014) is applied after an EFA or CFA tree split the data to evaluate which items are actually invariant as well as which and how many items are problematic (Somaraju et al., 2022). Together, Figure 3 extends the workflow of Somaraju et al. (2022) by integrating EFA and CFA trees, which reflects our suggestions above and advances current practices on testing for MI. It is important to note, however, that applying CFA trees with assumed metric MI after an EFA tree analysis should be limited to situations where the preceding EFA trees did not split the data, where violations of metric invariance are considered negligible, or the analysis of subgroups that were identified as metrically invariant.

When to Use EFA and When to Use CFA (Trees)

Due to the purely data-driven MOB algorithm, both EFA and CFA trees serve a clear exploratory purpose. Violations of MI can be assessed for undefined groups, numerous covariates, and their interactions; however, researchers still must decide whether to use an EFA or a more constrained CFA model within the MOB framework. As stated above, EFA trees are a tool that can be used even without specifying a measurement model (i.e., defining which

Figure 3

Adjusted Measurement Invariance Detection Workflow From Somaraju et al. (2022)—Integration of EFA and CFA Trees to Account for Continuous Covariates and Explore Multiple Potential Causes of Noninvariance



Note. EFA = exploratory factor analysis; CFA = confirmatory factor analysis; MGCFA = multiple-group confirmatory factor analysis; MNE = measurement nonequivalence; d_{MACS} = mean and covariance structure. From "A Review of Measurement Equivalence in Organizational Research: What's Old, What's New, What's Next?" by A. V. Somaraju, C. D. Nye, and J. Olenick, 2022, *Organizational Research Methods*, *25*(4), p. 771 (https://doi.org/10.1177/10944281211056524). Copyright 2022 by SAGE Publications. Reprinted with permission. See the online article for the color version of this figure.

indicators are related to which latent factors). Hence, EFA trees can be seen as the first MI tool that already addresses MI at the earliest stages of scale construction (Goretzko & Bühner, 2022; Sterner & Goretzko, 2023). When investigating MI in established scales, researchers may prefer fitting confirmatory models and therefore may tend to favor CFA trees in these settings.

The choice between EFA and CFA is not easy and will always depend on the aim of the study. The very flexible and less restricted EFA model and a CFA with independent clusters assumption (i.e., each indicator is only allowed to load on one factor) can be seen as the two endpoints of a continuum of factor analytic approaches that incorporate different degrees of theoretical constraints.

When developing measurement models, researchers typically start with an EFA and subsequently fix loading parameters step-by-step to obtain a sparser, more interpretable, and generalizable solution. Nájera et al. (2023) illustrate and compare the different approaches along this continuum. Their simulations show that rather strict confirmatory models perform comparably poorly if they are not correctly specified. Given that CFA with an independent clusters assumption will not fit the data from most psychological scales well (Hopwood & Donnellan, 2010; Sellbom & Tellegen, 2019) and many published studies report unsatisfactory model fit for CFA models (Goretzko et al., 2024), less constrained factor models may be preferable when exploring MI with a tree-based method. CFA trees will also likely split the data in case of noninvariance when the measurement model is misspecified and a relevant cross-loading is omitted, but the resulting parameter estimates will not be accurate, and researchers may misinterpret the cause of the noninvariance. Alternatively, researchers could constrain all parameters except the intercepts to be equal to the estimates of a single-group CFA. This way, only intercepts are considered for splitting when investigating parameter instability (i.e., scalar noninvariance). An EFA, where all loading parameters are freely estimated (apart from those fixed for identification purposes), does not require the user to explicitly specify the model, which is why relevant parameters will (almost certainly) be in the model. Since CFA models can be easily misspecified, some researchers (e.g., Crede & Harms, 2019) suggest always estimating an EFA as well to inform the interpretation of the hypothesized CFA model.

In our opinion, with MOB being a fully exploratory algorithm, it appears to be preferable to rely on EFA trees instead of CFA trees as a default (when analyzing configural and metric invariance). CFA trees should be used when (a) the overall structure of the measurement model has been well established[8], (b) specific covariances among residuals should be considered (which is usually not done in EFA and requires adding additional constraints to identify the model), (c) scalar invariance is investigated (see our empirical example), or (d) more complex model structures such as bi-factor models (e.g., Khojasteh & Lo, 2015) are analyzed.

There is a major caveat when applying EFA—namely rotational indeterminacy (e.g., Browne, 2001). The factor solution is not uniquely identifiable in EFA[9] and various rotations can be applied to foster interpretability without changing the model fit (Browne, 2001; Fabrigar et al., 1999; Goretzko et al., 2021). Accordingly, it remains difficult (and to some extent arbitrary) to choose a suitable rotation method when performing an EFA. In combination with MOB, this issue becomes even more challenging as different EFA fitted to different subsets may require different rotations to obtain interpretable solutions (Sterner & Goretzko, 2023). De Roover and Vermunt (2019) developed multigroup factor rotation to solve this problem, while Sterner and Goretzko (2023) relied on regularized EFA (Goretzko, 2023; Scharf & Nestler, 2019). When integrating EFA trees in the workflow of Somaraju et al. (2022; Figure 3), researchers can also directly use alignment optimization (Asparouhov & Muthén, 2023) instead of common factor rotation. Hence, despite the challenge of handling rotational indeterminacy (which most users know from simple EFA applications), EFA trees are usually more promising than CFA trees when exploring (configural and metric) MI broadly among various covariates.

Detecting the Cause(s) of Noninvariance

While EFA and CFA trees do not replace common approaches to MI testing such as MG-CFA, they extend the methodological toolbox for researchers to investigate nonequivalences and potentially biasing influences on measurement models. The assessment of MI should not just solely focus on covariates that define the groups of interest (e.g., different genders, age groups, or countries). Sterner, Pargent, et al. (2024) illustrated how a second covariate that is not the grouping variable of interest can distort the results of MI testing if not properly accounted for. They also advocate for taking a causal perspective on the measurement of latent variables, and, therefore, MI as well. To broadly explore the influence of various covariates on the measurement of a specific latent variable and to inform their conceptual and theoretical work, researchers need data-driven approaches such as the EFA or CFA trees. Based on the results of such exploratory analyses, a more comprehensive conceptualization of the measurement model can be developed that includes all (potential) causes of MNE. These hypothesized influences can then be thoroughly tested and/or accounted for with an appropriate modeling strategy (e.g., moderated nonlinear factor analyses, Bauer, 2017).

Controlling for a covariate for which MI is violated, even though it is not the grouping variable that is actually of interest, can be compared to the inclusion of control variables in a regression analysis which is common practice in organizational research (Carlson & Wu, 2012). To obtain unbiased estimates of the effect of interest, potential confounders need to be controlled for. However, selecting covariates to enter the model should not be done carelessly (e.g., Carlson & Wu, 2012; Mändli & Rönkkö, 2023; Spector & Brannick, 2011; Wysocki et al., 2022). Here, a thoroughly thought-out conceptual model comes into play. While such models can be used to derive appropriate models for the statistical analysis of a phenomenon from theoretical assumptions (e.g., Cinelli et al., 2022; Huenermund et al., 2022; Mändli & Rönkkö, 2023), they can also be used to improve the development of measurement models and MI testing (Sterner, Pargent, et al., 2024). Accordingly, we believe that exploratory methods such as EFA and CFA trees, that help us to identify potential confounders, are invaluable when it comes to inform measurement theories and related conceptual work.

[8] In case a well-established scale is analyzed, researchers may want to use CFA trees to avoid focusing on minor differences in negligible cross-loadings that may mislead an EFA-based approach. However, this is only advisable in cases where the respective scale has been thoroughly tested with diverse samples, so that substantive differences in the overall pattern of cross-loadings can be ruled out.

[9] Rotational indeterminacy arises for multivariate normal data, while factors that are nonnormally distributed may be (partially) identifiable (Rohe & Zeng, 2023). However, uniquely identifying latent factors in data stemming from questionnaire research typical to psychology and organizational studies remains an issue (Pargent et al., 2023).

Theoretical and Pracitical Implications

Identifying Sources of Nonequivalence With EFA and CFA Trees

The study of MI has advanced beyond simply testing whether MI holds, but a steady stream of research has also investigated the sources of any observed nonequivalence (Fischer & Rudnev, 2024; Somaraju et al., 2022). One predominant stream is testing sources of measurement noninvariance in multisource performance ratings. Authors often test whether certain characteristics of the rater and ratee impact MI, but they were previously limited to testing categorical variables with few groups (e.g., gender and race, see Somaraju et al., 2022). EFA and CFA trees enable authors to assess a wider array of sources, such as those that are continuous or with many groups, and future researchers can broaden theory on multisource performance ratings. For instance, researchers can now assess the extent that personality or cognitive processes influence the invariance of rater assessments, integrating calls in the study of rater source effects into research on MI (Jackson et al., 2020).

Furthermore, Somaraju et al. (2022) argued that theory testing with MI has been stifled because, "there is a dearth of research on the potential causes of nonequivalence" (p. 756). We believe that this dearth is due to the difficulty of exploring these causes with MG-CFA, which is a barrier that is significantly reduced with exploratory tools such as EFA and CFA trees. Specifically, Somaraju et al. continued by stating that MNE "is often attributed to differences in group classification (e.g., gender, racial or ethnic group), when in reality it may be due to underlying psychological differences" (p. 756). These psychological differences are difficult to investigate through MG-CFA, as they are continuous by nature. Authors would be forced to artificially dichotomize standings on these psychological differences, for which they rarely have sound justification. EFA and CFA trees can study these influences in their continuous forms, eliminating any need to dichotomize and enabling researchers to directly study these psychological differences. Therefore, we foresee EFA and CFA trees beginning a novel stream on the causes of MNE, as they are the analyses that aptly fit the needs of calls produced by prior authors.

Organizational researchers should embrace MI exploration as a way to learn more about not only their measurement models but also broader theoretical concepts (e.g., Laguna et al., 2017; Somaraju et al., 2022). EFA and CFA trees, as new and more exploratory approaches to the investigation of measurement equivalence, can contribute strongly to theory building and this learning process by motivating research questions that would otherwise have remained untouched. Notably, these analyses can uncover interactions between covariates that are associated with noninvariance and thus with how a construct is perceived across different groups. MNEs that will be undetected with MG-CFA can be explored and hypotheses for subsequent confirmatory analyses can be generated (see adjusted workflow in Figure 3). In particular, the ability of EFA and CFA trees to handle continuous covariates and detect complex interactions among them may be used to find violations of MI and advance theory in ways that would not be possible with other methods. For example, Harari et al. (2019) utilized MG-CFA to discover that, "some forms of response bias … changed across time" regarding rater evaluations of ratee job performance, but the limitations of the analysis did not allow the authors to probe this finding further to more specifically evaluate the nature and sources of these changes. EFA and CFA trees can now be applied to more precisely investigate how these response biases changed, and the inspection of multiple covariates at once could provide insights into potential causes of this change—significantly advancing theory on these associated topics. Similarly, a tree could find out that an assessment center task is only properly reflecting the ability of interest (e.g., a leadership competency) for candidates with an IQ below 130 and is not measuring this ability in the same way when it comes to candidates with an IQ above this threshold. Researchers could then probe this finding in a more directed manner, further testing why an IQ of 130 was the identified splitting point. Especially in diagnostic settings like this, fairness based on measurement equivalence is of high importance, and researchers could derive more complete models and theories for the sources of MI.

EFA and CFA Trees in Practice

EFA and CFA trees can likewise produce immediate impacts on modern-day business practices. Legislation often drives the decision of which variables to assess in tests of MI[10]. In the United States, for example, specific demographic groups (e.g., race, gender, religion, and age) are protected against discrimination by the Civil Rights Act, the Americans with Disabilities Act, and the Age Discriminant Act. Present applications guided by this legislation have largely applied MG-CFA to ensure that measures perform similarly across protected groups, and we believe that many immediate applications of EFA and CFA trees will be applied in a similar manner. In doing so, practitioners can provide deeper insights into the nature of these protected classes. For example, they could investigate whether the differential functioning of measures depends on multiple-class membership, which would advance the importance of studying intersectionality at work (Rosette et al., 2018; Salter et al., 2021). While only a suggestion, this is indeed a possibility that can now be studied with EFA and CFA trees.

Also, organizations are increasingly multinational, representing employees across multiple countries and even continents (Hines, 2021). Within these companies, it is problematic to assume that developed selection tests or employee surveys would function identically across all locations. For this reason, it is often recommended to perform tests of MI before broadly applying these tests and surveys. Thus far, MG-CFA has performed well for this purpose, but EFA and CFA trees may be more ideal for many organizations. Some organizations may want to ensure the invariance of their measures across multiple groups at the same time that may have interactive effects, such as location and gender as the dynamics of the latter may depend on the former. Likewise, larger organizations may presently use MG-CFA to determine whether MI is violated across an array of locations at the same time, but this approach may pose difficulties in determine why measurement may not hold when studying these constructs. These organizations may instead want to treat certain geographically driven variables (e.g., culture) as continuous and conduct EFA and CFA trees, as this analysis could provide more directed insights into why measurement invariance may not hold (e.g., tree split at higher power distance).

[10] We thank an anonymous reviewer for pointing out this consideration.

Future Research

As our goal was not to provide a comprehensive analysis of all methods to assess MI, we did not provide detailed comparisons between the methods, which is a clear opportunity for future research to advance our understanding of MI and the methods available to investigate it (similar to, e.g., Kim et al., 2017; Sterner, De Roover, & Goretzko, 2024). Future research should also include specific methodological investigations of EFA and CFA trees.

Simulation studies can help to identify the contexts for which the analyses provide accurate results. Previous studies have shown that EFA trees and MOB in general perform well in many scenarios frequently seen in the organizational sciences (e.g., Brandmaier et al., 2013b; Sterner & Goretzko, 2023), but authors should investigate more atypical conditions, such as data that are severely nonnormal. Furthermore, data conditions with many covariates that share a lot of variance (i.e., highly multicollinear variables that are considered as potential split variables) must be thoroughly investigated in future research; previous studies (e.g., Debelak & Strobl, 2019; Sterner & Goretzko, 2023; Strobl et al., 2015; Zeileis et al., 2008) have not considered these scenarios, while Brandmaier et al. (2013b) noted that the Bonferroni correction within the trees could become overly conservative when covariates are highly correlated.

Researchers should produce a focused analysis on the potential role of EFA trees in the scale development process. We propose that EFA trees can be used in the earlier phases of measurement assessment, as it can be utilized in any context that EFA would regularly be applied. This approach would require some rethinking about appropriate sample(s) to obtain and measure(s) to use in these earlier phases, as discussed in our introduction of EFA trees. Therefore, it may be beneficial to develop explicit guidelines of how EFA trees can and should be used in conjunction with other analyses to ensure the psychometric properties and validity of measures when administered across multiple potential populations—perhaps even unknowingly. In doing so, EFA and CFA trees can be further developed implementing effect-size–based stopping criteria that would allow researchers to specify which violations of MI they deem meaningful. This way, not only statistical but also practical relevance could be considered when growing the tree. In the context of item response modeling trees, similar approaches have recently been developed to quantify differential item functioning (e.g., Henninger et al., 2023).

Conclusion

Of course, EFA and CFA trees will not solve all the issues of contemporary MI methods such as MG-CFA (Putnick & Bornstein, 2016; Somaraju et al., 2022; Sterner & Goretzko, 2023). However, they extend the toolbox that organizational researchers can utilize to ensure MI holds when comparing latent means across groups and to investigate sources of MNE to advance their (causal) theories. We understand EFA and CFA trees as another step in the process outlined by Somaraju et al. (2022), both with their own strengths and capabilities regarding specific levels of MI. In this, we deem them especially useful when multiple (continuous) covariates and their interactions should be investigated, when little is known about potential violations of MI, and/or when the measurement model development is not finished yet. When applied in this workflow, EFA and CFA trees promise to be a great addition to the organizational research methodology and offer significant capabilities to advance relevant theory.

References

Arnold, M., Voelkle, M. C., & Brandmaier, A. M. (2021). Score-guided structural equation model trees. *Frontiers in Psychology*, *11*, Article 564403. https://doi.org/10.3389/fpsyg.2020.564403

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

Asparouhov, T., & Muthén, B. (2023). Multiple group alignment for exploratory and structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(2), 169–191. https://doi.org/10.1080/10705511.2022.2127100

Aust, F., & Barth, M. (2020). *Papaja: Create APA manuscripts with R Markdown*. https://github.com/crsh/papaja

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*(3), 507–526. https://doi.org/10.1037/met0000077

Bauer, D. J., Belzak, W. C., & Cole, V. T. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 43–55. https://doi.org/10.1080/10705511.2019.1642754

Brandmaier, A. M., Driver, C. C., & Voelkle, M. C. (2018). Recursive partitioning in continuous time analysis. In K. van Montfort, J. H. L. Oud, & M. C. Voelkle (Eds.), *Continuous time modeling in the behavioral and related sciences* (pp. 259–282). Springer. https://doi.org/10.1007/978-3-319-77219-6_11

Brandmaier, A. M., Prindle, J. J., McArdle, J. J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. *Psychological Methods*, *21*(4), 566–582. https://doi.org/10.1037/met0000090

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013a). Exploratory data mining with structural equation model trees. In J. J. McArdle & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences*. Routledge. https://www.researchgate.net/publication/258027588_Exploratory_data_mining_with_structural_equation_model_trees

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013b). Structural equation model trees. *Psychological Methods*, *18*(1), 71–86. https://doi.org/10.1037/a0030001

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150. https://doi.org/10.1207/S15327906MBR3601_05

Burlew, A. K., Peteet, B. J., McCuistian, C., & Miller-Roenigk, B. D. (2019). Best practices for researching diverse groups. *American Journal of Orthopsychiatry*, *89*(3), 354–368. https://doi.org/10.1037/ort0000350

Bynum, B. H., Hoffman, B. J., Meade, A. W., & Gentry, W. A. (2013). Reconsidering the equivalence of multisource performance ratings: Evidence for the importance and meaning of rater factors. *Journal of Business and Psychology*, *28*, 203–219. https://doi.org/10.1007/s10869-012-9272-7

Carlson, K. D., & Wu, J. (2012). The illusion of statistical control: Control variable practice in management research. *Organizational Research Methods*, *15*(3), 413–435. https://doi.org/10.1177/1094428111428817

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

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://www.tandfonline.com/doi/abs/10.1207/S15328007sem0902_5

Cho, I., Hu, B., & Berry, C. M. (2023). A matter of when, not whether: A meta-analysis of modesty bias in east Asian self-ratings of job performance. *Journal of Applied Psychology*, *108*(2), 291–306. https://doi.org/10.1037/apl0001046

Cinelli, C., Forney, A., & Pearl, J. (2022). A crash course in good and bad controls. *Sociological Methods & Research*, *53*(3), 1071–1104. https://doi.org/10.1177/00491241221099552

Crede, M., & Harms, P. (2019). Questionable research practices when using confirmatory factor analysis. *Journal of Managerial Psychology*, *34*(1), 18–30. https://doi.org/10.1108/JMP-06-2018-0272

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*, 55–75. https://doi.org/10.1146/annurev-soc-071913-043137

De Roover, K. (2021). Finding clusters of groups with measurement invariance: Unraveling intercept non-invariance with mixture multigroup factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(5), 663–683. https://doi.org/10.1080/10705511.2020.1866577

De Roover, K., & Vermunt, J. K. (2019). On the exploratory road to unraveling factor loading non-invariance: A new multigroup rotation approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(6), 905–923. https://doi.org/10.1080/10705511.2019.1590778

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. https://doi.org/10.1037/met0000355

Debelak, R., & Strobl, C. (2019). Investigating measurement invariance by means of parameter instability tests for 2PL and 3PL models. *Educational and Psychological Measurement*, *79*(2), 385–398. https://doi.org/10.1177/0013164418777784

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087–1101. https://doi.org/10.1037/0022-3514.92.6.1087

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Fischer, R., & Rudnev, M. (2024). From MIsgivings to MIse-en-scène: The role of invariance in personality science. *European Journal of Personality*, *39*(4), 662–673. https://doi.org/10.1177/08902070241283081

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods*, *50*, 2016–2034. https://doi.org/10.3758/s13428-017-0971-x

Goretzko, D. (2023). Regularized exploratory factor analysis as an alternative to factor rotation. *European Journal of Psychological Assessment*, *41*(4), 264–276. https://doi.org/10.1027/1015-5759/a000792

Goretzko, D., & Bühner, M. (2022). Note: Machine learning modeling and optimization techniques in *psychological assessment*. *Psychological Test and Assessment Modeling*, *64*(1), 3–21.

Goretzko, D., Pham, T. T. H., & Bühner, M. (2021). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology*, *40*(7), 3510–3521. https://doi.org/10.1007/s12144-019-00300-2

Goretzko, D., Siemund, K., & Sterner, P. (2024). Evaluating model fit of measurement models in confirmatory factor analysis. *Educational and Psychological Measurement*, *84*(1), 123–144. https://doi.org/10.1177/00131644231163813

Harari, M. B., Naemi, B., & Viswesvaran, C. (2019). Is the validity of conscientiousness stable across time? Testing the role of trait bandwidth. *Journal of Occupational and Organizational Psychology*, *92*(1), 212–220. https://doi.org/10.1111/joop.12241

Henninger, M., Debelak, R., & Strobl, C. (2023). A new stopping criterion for rasch trees based on the mantel–haenszel effect size measure for differential item functioning. *Educational and Psychological Measurement*, *83*(1), 181–212. https://doi.org/10.1177/00131644221077135

Hines, J. R. (2021). *Global goliaths: Multinational corporations in the 21st century economy*. Brookings Institution Press.

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, *14*(3), 332–346. https://doi.org/10.1177/1088868310361240

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447

Hothorn, T., & Zeileis, A. (2015). Partykit: A modular toolkit for recursive partytioning in R. *The Journal of Machine Learning Research*, *16*(1), 3905–3909. https://jmlr.org/papers/v16/hothorn15a.html

Howard, M. C. (2023). A systematic literature review of exploratory factor analyses in management. *Journal of Business Research*, *164*, Article 113969. https://doi.org/10.1016/j.jbusres.2023.113969

Huenermund, P., Louw, B., & Rönkkö, M. (2022). The choice of control variables: How causal graphs can inform the decision. *Academy of Management Proceedings*, *2022*(1), 15534. https://doi.org/10.5465/AMBPP.2022.294

Jackson, D. J., Michaelides, G., Dewberry, C., Schwencke, B., & Toms, S. (2020). The implications of unconfounding multisource performance ratings. *Journal of Applied Psychology*, *105*(3), 312–329. https://doi.org/10.1037/apl0000434

Jebb, A. T., & Tay, L. (2017). Introduction to time series analysis for organizational research: Methods for longitudinal analyses. *Organizational Research Methods*, *20*(1), 61–94. https://doi.org/10.1177/1094428116668035

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). *semTools: Useful tools for structural equation modeling*. https://CRAN.R-project.org/package=semTools

Khojasteh, J., & Lo, W.-J. (2015). Investigating the sensitivity of goodness-of-fit indices to detect measurement invariance in a bifactor model. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(4), 531–541. https://doi.org/10.1080/10705511.2014.937791

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(4), 524–544. https://doi.org/10.1080/10705511.2017.1304822

Kim, E. S., Joo, S.-H., Lee, P., Wang, Y., & Stark, S. (2016). Measurement invariance testing across between-level latent classes using multilevel factor mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(6), 870–887. https://doi.org/10.1080/10705511.2016.1196108

Komboz, B., Strobl, C., & Zeileis, A. (2018). Tree-based global model tests for polytomous Rasch models. *Educational and Psychological Measurement*, *78*(1), 128–166. https://doi.org/10.1177/0013164416664394

Laguna, M., Mielniczuk, E., Razmus, W., Moriano, J. A., & Gorgievski, J. M. (2017). Cross-culture and gender invariance of the Warr (1990) job-related well-being measure. *Journal of Occupational and Organizational Psychology*, *90*(1), 117–125. https://doi.org/10.1111/joop.12166

Lu, J. G. (2023). A creativity stereotype perspective on the bamboo ceiling: Low perceived creativity explains the underrepresentation of East Asian leaders in the United States. *Journal of Applied Psychology*, *109*(2), 238–256. https://doi.org/10.1037/apl0001135

Maassen, E., D'Urso, E. D., van Assen, M. A. L. M., Nuijten, M. B., De Roover, K., & Wicherts, J. M. (2023). The dire disregard of measurement invariance testing in psychological science. *Psychological Methods*, *30*(5), 966–979 https://doi.org/10.1037/met0000624

Mändli, F., & Rönkkö, M. (2023). To omit or to include? Integrating the frugal and prolific perspectives on control variable use. *Organizational Research Methods*, *28*(1), 114–137. https://doi.org/10.1177/10944281231221703

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance.

Journal of Applied Psychology, *93*(3), 568–592. https://doi.org/10.1037/0021-9010.93.3.568

Merkle, E. C., Fan, J., & Zeileis, A. (2014). Testing for measurement invariance with respect to an ordinal variable. *Psychometrika*, *79*(4), 569–584. https://doi.org/10.1007/s11336-013-9376-7

Merritt, S. M. (2012). The two-factor solution to Allen and Meyer's (1990) affective commitment scale: Effects of negatively worded items. *Journal of Business and Psychology*, *27*, 421–436. https://doi.org/10.1007/s10869-011-9252-3

Morelli, N. A., Mahan, R. P., & Illingworth, A. J. (2014). Establishing the measurement equivalence of online selection assessments delivered on mobile versus nonmobile devices. *International Journal of Selection and Assessment*, *22*(2), 124–138. https://doi.org/10.1111/ijsa.12063

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585. https://doi.org/10.1007/BF02296397

Nájera, P., Abad, F. J., & Sorrel, M. A. (2023). Is exploratory factor analysis always to be preferred? A systematic comparison of factor analytic techniques throughout the confirmatory–exploratory continuum. *Psychological Methods*, *30*(1), 16–39. https://doi.org/10.1037/met0000579

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. *Organizational Research Methods*, *22*(3), 678–709. https://doi.org/10.1177/1094428118761122

Nye, C. D., Brummel, B. J., & Drasgow, F. (2010). Too good to be true? Understanding change in organizational outcomes. *Journal of Management*, *36*(6), 1555–1577. https://doi.org/10.1177/1094428118761122

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, *96*(5), 966–980. https://doi.org/10.1037/a0022955

Pargent, F., Goretzko, D., & von Oertzen, T. (2023). New insights into PCA+ varimax for psychological researchers: A short commentary on Rohe & Zeng (2023). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*(4), 1087–1088. https://github.com/FlorianPargent/pca_varimax_commentary

Podsakoff, N. P., Spoelma, T. M., Chawla, N., & Gabriel, A. S. (2019). What predicts within-person variance in applied psychology constructs? An empirical examination. *Journal of Applied Psychology*, *104*(6), 727–754. https://doi.org/10.1037/apl0000374

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*(3), 517–529. https://doi.org/10.1037/0021-9010.87.3.517

Revelle, W. (2024). *Psych: Procedures for psychological, psychometric, and personality research* (R package version 2.5.6). Northwestern University. https://CRAN.R-project.org/package=psych

Robert, C., Lee, W. C., & Chan, K.-Y. (2006). An empirical analysis of measurement equivalence with the indcol measure of individualism and collectivism: Implications for valid cross-cultural inference. *Personnel Psychology*, *59*(1), 65–99. https://doi.org/10.1111/j.1744-6570.2006.00804.x

Rohe, K., & Zeng, M. (2023). Vintage factor analysis with Varimax performs statistical inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*(4), 1037–1060. https://doi.org/10.1093/jrsssb/qkad029

Rosette, A. S., deLeon, R. P., Koval, C. Z., & Harrison, D. A. (2018). Intersectionality: Connecting experiences of gender with race at work. *Research in Organizational Behavior*, *38*, 1–22. https://doi.org/10.1016/j.riob.2018.12.002

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02

Ruglass, L. M., Morgan-López, A. A., Saavedra, L. M., Hien, D. A., Fitzpatrick, S., Killeen, T. K., Back, S. E., & López-Castro, T. (2020). Measurement nonequivalence of the clinician-administered PTSD scale by race/ethnicity: Implications for quantifying posttraumatic stress disorder severity. *Psychological Assessment*, *32*(11), 1015–1027. https://doi.org/10.1037/pas0000943

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

Salter, N. P., Sawyer, K., & Gebhardt, S. T. (2021). How does intersectionality impact work attitudes? The effect of layered group memberships in a field sample. *Journal of Business and Psychology*, *36*(6), 1035–1052. https://doi.org/10.1007/s10869-020-09718-z

Scharf, F., & Nestler, S. (2019). Should regularization replace simple structure rotation in exploratory factor analysis? *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 576–590. https://doi.org/10.1080/10705511.2018.1558060

Schaufeli, W. B., Bakker, A. B., & Salanova, M. (2006). the measurement of work engagement with a short questionnaire: A cross-national study. *Educational and Psychological Measurement*, *66*(4), 701–716. https://doi.org/10.1177/0013164405282471

Schneider, L., Strobl, C., Zeileis, A., & Debelak, R. (2021). An R toolbox for score-based measurement invariance tests in IRT models. *Behavior Research Methods*, *54*, 2101–2113 https://doi.org/10.3758/s13428-021-01689-0

Sellbom, M., & Tellegen, A. (2019). Factor analysis in *psychological assessment* research: Common pitfalls and recommendations. *Psychological Assessment*, *31*(12), 1428–1441. https://doi.org/10.1037/pas0000623

Somaraju, A. V., Nye, C. D., & Olenick, J. (2022). A review of measurement equivalence in organizational research: What's old, what's new, what's next? *Organizational Research Methods*, *25*(4), 741–785. https://doi.org/10.1177/10944281211056524

Spector, P. E., & Brannick, M. T. (2011). Methodological urban legends: The misuse of statistical control variables. *Organizational Research Methods*, *14*(2), 287–305. https://doi.org/10.1177/1094428110369842

Sterner, P., De Roover, K., & Goretzko, D. (2024). New developments in measurement invariance testing: An overview and comparison of EFA-based approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *32*(1), 1–19. https://doi.org/10.1080/10705511.2024.2393647

Sterner, P., & Goretzko, D. (2023). Exploratory factor analysis trees: Evaluating measurement invariance between multiple covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(6), 871–886. https://doi.org/10.1080/10705511.2023.2188573

Sterner, P., Pargent, F., Deffner, D., & Goretzko, D. (2024). A causal framework for the comparability of latent variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*(5), 747–758. https://doi.org/10.1080/10705511.2024.2339396

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*(2), 289–316. https://doi.org/10.1007/s11336-013-9388-3

Thissen, D. (2024). A review of some of the history of factorial invariance and differential item functioning. *Multivariate Behavioral Research*, *60*(2), 211–235. https://doi.org/10.1080/00273171.2024.2396148

Van De Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement

invariance. *Frontiers in Psychology*, *4*, Article 770. https://doi.org/10.3389/fpsyg.2013.00770

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Measurement invariance. *Frontiers in Psychology*, *6*, Article 1064. https://doi.org/10.3389/fpsyg.2015.01064

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods*, *5*(2), 139–158. https://doi.org/10.1177/109442810205002001

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Wysocki, A. C., Lawson, K. M., & Rhemtulla, M. (2022). Statistical control requires causal justification. *Advances in Methods and Practices in Psychological Science*, *5*(2). https://doi.org/10.1177/25152459221095823

Zeileis, A., & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, *61*(4), 488–508. https://doi.org/10.1111/j.1467-9574.2007.00371.x

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514. https://doi.org/10.1198/106186008X319331

Received April 30, 2024
Revision received November 4, 2025
Accepted December 18, 2025 ■