METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Tutorial

Metric Invariance in Exploratory Graph Analysis via Permutation Testing

Laura Jamison [1], Alexander P. Christensen [2], Hudson F. Golino [1]

[1] *Department of Psychology, University of Virginia, Charlottesville, VA, USA.* **[2]** *Department of Psychology and Human Development, Vanderbilt University, Nashville, TN, USA.*

Methodology, 2024, Vol. 20(2), 144–186, https://doi.org/10.5964/meth.12877

Received: 2023-09-26 • **Accepted:** 2024-05-21 • **Published (VoR):** 2024-06-28

Handling Editor: Isabel Benítez, University of Granada, Granada, Spain

Corresponding Author: Laura Jamison, 485 McCormick Rd, Charlottesville, VA 22904, USA. E-mail: lj5yn@virginia.edu

Supplementary Materials: Code, Data [see Index of Supplementary Materials]

Abstract

Establishing measurement invariance (MI) is crucial for the validity and comparability of psychological measurements across different groups. If MI is violated, mean differences among groups could be due to the measurement rather than differences in the latent variable. Recent research has highlighted the prevalence of inaccurate MI models in studies, often influenced by the software used. Additionally, unequal group sample sizes, noninvariant referent indicators, and reliance on data-driven methods reduce the power of traditional SEM methods. Network psychometrics lacks methods comparing network structures conceptually similar to MI. We propose a more conceptually consistent method within the Exploratory Graph Analysis (EGA) framework using network loadings, analogous to factor loadings. Our simulation study demonstrates that this method offers comparable or improved power, especially in scenarios with smaller or unequal sample sizes and lower noninvariance effect sizes, compared to SEM MI testing.

Keywords

measurement invariance, permutation testing, metric invariance, network psychometrics, exploratory graph analysis

Measurement invariance assesses whether a latent variable is measured equivalently across groups. This equivalence indicates that a measure quantitatively has the same meaning to each group and is therefore measuring the same construct in the same way across groups. Demonstrating measurement invariance is vital for the generalizabil-

ity of psychometric measurement. For any measure where validity and reliability are necessary, tests for measurement invariance prior to administration across qualitatively distinct groups or time points should be conducted (Vandenberg & Lance, 2000). Many researchers claim that comparisons between cultures, administration modes, language versions, or sociodemographic groups cannot be credibly interpreted unless a measure demonstrates invariance (Borsboom, 2006). If measurement invariance is violated, score differences between groups can be the result of measurement rather than the true latent variable (Chen, 2007).

Traditionally, measurement invariance is tested using either Item Response Theory (IRT) or Structural Equation Modeling (SEM; Stark et al., 2006). Because SEM is more prevalent across psychological domains (Putnick & Bornstein, 2016), we narrow our focus to this framework. Within SEM, four consecutive tests are used to establish measurement invariance: configural (equivalence of factor structure), metric (equivalence of factor loadings), scalar (equivalence of item intercepts), and strict (equivalence of item residuals; Widaman & Reise, 1997). For a measure to be considered fully invariant, it must pass each of these tests.

The current paper presents a method to test metric invariance using network psychometrics in the Exploratory Graph Analysis (EGA) framework (Golino et al., 2020; Golino & Epskamp, 2017). First, a brief overview of the limitations associated with testing for measurement invariance in traditional psychometrics, focusing on metric invariance, is provided. Afterward, EGA is introduced and the proposed method to test metric invariance is discussed.

Measurement Invariance in Traditional Psychometrics

Factorial Invariance

Testing for factorial (measurement) invariance is conducted by comparing a more constrained model to the previous stage's less constrained model, from a weaker to stronger level of invariance (e.g., configural model to metric model), using a likelihood ratio test. The constrained model sets relevant parameters (e.g., loadings) to be equal across groups and model fit is compared to the unconstrained model where the same parameters are estimated freely. If the constrained model has a better fit, then invariance at that level is established. This process starts by testing configural invariance.

Configural invariance (factor structure equivalence) is established by assessing the fit of a Multi-Group Confirmatory Factor Analysis model. Following R. E. Millsap (2011), let the common factor model be defined as:

$$X_j = \gamma_{jk} + \sum_{m=1}^{M} \lambda_{jmk} W_m + U_j, \tag{1}$$

where γ_{jk} is the latent intercept for variable j in population k, λ_{jmk} is the factor pattern loadings for variable j corresponding to the M common factors ($m = 1, ..., M$) in population k, W_m represents the common factor scores for factor m, and U_j is the unique factor score for variable j. When $\mathbf{W}' = (W_1, W_2, ..., W_m)$ and $\mathbf{U}' = (U_1, U_2, ..., U_p)$ it is assumed that $E_k(\mathbf{U}) = \mathbf{0}$ and that it is uncorrelated with $\mathbf{W}$.

The unconditional mean and covariance structure for the measured variables $\mathbf{X}$ can then be expressed as:

$$Cov_k(\mathbf{X}) = \sum_{X_k} = \Lambda_k \Phi \Lambda_k' + \theta_k \tag{2}$$

and

$$E_k(\mathbf{X}) = \mu_{X_k} = \tau_k + \Lambda_k \kappa_k, \tag{3}$$

where $\Phi_k = Cov_k(\mathbf{W})$ and $\theta_\mathbf{k} = Cov_k(\mathbf{U})$. Finally, the test for configural invariance can be defined as,

$$\sum_{X_k} = \Lambda_{kc} \Phi_k \Lambda_{kc}' + \Theta_k \tag{4}$$

and

$$\mu_{X_k} = \tau_k + \Lambda_{kc} \kappa_k, \tag{5}$$

for $k = 1, ..., K$ where Λ_{kc} denotes the pattern loadings matrices that have the factor structure with configural invariance.

This formula implies that each population has the same number of factors containing the same distribution of variables. If the model fits satisfactorily on all groups, then the organization of items into constructs is appropriate for all groups (Putnick & Bornstein, 2016). In other words, configural invariance is established such that the pattern of zero and nonzero loadings (fixed and free loadings) exists in all groups (Widaman & Reise, 1997). Configural invariance only demonstrates that similar, not equivalent, latent factors exist in all groups. Similar latent factors contain the same items across groups but do not necessarily imply that the groups have equivalent loadings, intercepts, or error terms. Testing for equivalent loadings is the next step.

Metric invariance (loading equivalence) can be defined as

$$\sum_{X_k} = \Lambda \Phi_k \Lambda' + \Theta_k \tag{6}$$

and

$$\mu_{X_k} = \tau_k + \Lambda\kappa_k \tag{7}$$

for $k = 1, \ldots, K$. This model constrains loadings to be equivalent across groups and is then compared to the configural model (unconstrained model). If the metric invariance model has a better fit, then each item contributes to their respective latent factors (and the overall latent construct) similarly across all groups (Putnick & Bornstein, 2016). If metric invariance is not established, then comparisons of factor variances and covariances (and subsequently scaled correlations) across groups cannot be made (Widaman & Reise, 1997). Without metric invariance, testing for scalar and strict invariance should not be conducted; however, testing for partial invariance of loadings is often appropriate.

Partial Invariance

Partial invariance occurs when only a portion of a parameter set lacks invariance. For metric invariance, the goal is to determine how many loadings lack invariance in each latent factor. Opinions vary on what level or proportion of partial invariance is permissible (Putnick & Bornstein, 2016). Testing for partial invariance can be useful to provide a more fine-grained perspective on which specific item parameters are noninvariant. In the case of metric invariance, individual constraints can be selectively introduced to Λ (loadings) and tested. It's possible that metric invariance is not found because only one item's loading is noninvariant across groups.

If partial invariance is found, at any level, then the researcher must determine (based on substantive reasoning or empirical criteria) how to handle instances of noninvariance. Arguably, identifying specific instances of noninvariance provides more useful information than an omnibus test for invariance, which only indicates that invariance exists across the parameters but not any one parameter specifically. Testing for partial invariance provides the same level of information as an omnibus test (whether noninvariance is present) but also where, if any, noninvariance exists.

It's possible that partial invariance testing could identify noninvariance not identified by an omnibus test. This problem is well documented for omnibus tests (Raykov et al., 2013), and the potential effects of misidentifying items as invariant can be consequential. Prior research has indicated that conducting individual local tests can lead to a more accurate evaluation of noninvariance (Jung & Yoon, 2016; Raykov et al., 2020; Stark et al., 2006). Therefore, examining local tests rather than relying on overall global testing provides more detailed information and lowers the risk of Type II errors.

There are several methods available to test partial invariance. Some methods use referent indicators or assume a specific indicator is invariant a priori, which presents many issues. These issues and some potential solutions are discussed in the next section; however, due to the issues associated with the selection of a referent, our study focuses on methods that test partial invariance that do not require the selection of a single referent indicator. Instead, we consider the following methods: factor-ratio test (Rensvold &

Cheung, 1998), a data-driven, sequential application of the modification index proposed by Yoon and Millsap (2007), and a method using a multiple testing procedure proposed by Raykov et al. (2013).

The factor-ratio test assesses partial invariance by comparing a fully unconstrained model to versions of a constrained model. Multiple constrained models are defined using all possible combinations of referent variables and choosing one of the remaining variables to test for invariance. A simulation study conducted by French and Finch (2008) found that this method works well to control false positive rates across data conditions and can successfully identify invariant items even when noninvariant items are present in the same factor. This procedure is computationally greedy as it investigates all possible combinations of referent indicators.

Yoon and Millsap (2007) proposed a data-driven method that sequentially evaluates modification indices. Within a fully constrained metric model, the factor variance of only one group is fixed to one. The factor variance of the other group is estimated freely. For both groups, factor loadings are constrained. Modification indexes are evaluated to estimate the change in χ^2 when the fixed parameters are freed. If an invariance constraint shows a significant modification index, then that parameter is relaxed. The process is continued until all modification indices are non-significant. Their simulation study found that this method controls false positive rates very well but primarily in "ideal" data conditions (large sample sizes, greater difference in loadings, low cross loadings). A limitation of this approach is that model misspecifications can lead to artificial inflation of Type I error rates (Kim & Yoon, 2011; Whittaker, 2012), especially as model modifications are made throughout the testing process (Yoon & Millsap, 2007).

Finally, Raykov et al. (2013) introduced a multiple comparison method which uses the Benjamini-Hochberg procedure (BH-procedure; Benjamini & Hochberg, 1995) to control the Type I error rate introduced by multiple comparisons. The method compares two models using χ^2 testing. One model (the baseline model) is a fully constrained model; the other model frees one set of parameters (e.g., loadings) across groups. The two models are compared and this process is repeated for all parameters. The number of tests conducted is equal to the number of variables. Zhang and Yang (2022) found that this method maintains high rates of power to detect noninvariance across varying data conditions (sample size, degree of noninvariance, proportion of noninvariance, and location of noninvariance). Although this method circumvents the choice of a referent indicator, the use of a fully constrained baseline model (i.e., including any model with constrained noninvariant items) could negatively impact accuracy (Benjamini & Hochberg, 1995). Given the cumbersome nature of the factor-ratio test and model misspecification limitations with the data-driven approach proposed by Yoon and Millsap (2007), we choose to focus on the multiple comparison method of Raykov et al. (2013) in this study.

Problems With Traditional Testing

χ^2 goodness of fit statistics are commonly used across all four measurement invariance tests, including tests of partial invariance. Putnick and Bornstein (2016) tested model fit alternatives such as Root Mean Square Error of Approximation (RMSEA), Standardized Root Mean-square Residual (SRMR), Comparative Fit Index (CFI), and Tucker-Lewis Index (TLI), finding that the choice of criterion could impact the discovery rate of invariant indicators. Importantly, model fit indices can further be impacted by disparate sample sizes across groups (Chen, 2007; Kaplan & George, 1995).

Another concern is that each stage requires certain decisions to be made about the model specification which, if incorrectly made, can have unanticipated consequences. A referent indicator used in partial metric invariance, for example, assumes that the chosen item is invariant which can adversely impact model interpretation (Johnson et al., 2009). Using these traditional methods, the assumption of noninvariance is not frequently tested, most likely due to the complicated nature of the methods available to test it (Finch & French, 2008). It is often unknown which items are invariant a priori. Procedures have been developed to identify which items are noninvariant prior to selecting a referent indicator (Cheung & Lau, 2012; Cheung & Rensvold, 1999; Rensvold & Cheung, 2001). These tests, however, can be quite complicated from both a conceptual and implementation standpoint with varying evidence of their statistical power (French & Finch, 2006; Jung & Yoon, 2016).

Lack of reporting or proper specification has been found in one out of four studies employing measurement invariance tests (Schroeders & Gnambs, 2020). After researchers analyzed the components of each study, they found that the most influential predictor of model misspecification was the software used, concluding a dearth in the statistical training of psychologists as the cause. R. Millsap and Olivera-Aguilar (2012) similarly pointed out that both the skill and experience levels of researchers have strong impacts on the effectiveness of testing measurement invariance.

In light of these findings, our study proposes a method that does not require intensive model specifications and model comparisons, with all model parameters tested without the need to introduce further testing or adjustments. Additionally, this method is straightforward to implement in the popular R statistical software (R Core Team, 2020). The primary goal of the current work is to provide a method to test metric invariance in the EGA framework.

Exploratory Graph Analysis

Network psychometric methods are an alternative to latent variable modeling. Networks represent variables as nodes (circles) and their relationships (e.g., partial correlations) as edges (lines). Because the relationships between variables are not known a priori, they must be estimated. There are many methods to estimate a network with the graph-

ical least absolute shrinkage and selection operator (GLASSO; Epskamp & Fried, 2018; Friedman et al., 2008) being one of the most common.

A key feature of network models is that each node is (usually) not connected to all other nodes (known as *sparsity*). Often, some nodes are more densely connected to each other relative to other nodes in the network. These sets of connected nodes are often referred to as *communities*, which are consistent with latent factors when data are generated from a factor model (Golino & Epskamp, 2017). Community detection algorithms are a common, data-driven way to identify communities in networks (Fortunato, 2010). The combination of the GLASSO with the Walktrap community detection algorithm (Pons & Latapy, 2006) has been labeled as Exploratory Graph Analysis (Golino et al., 2020; EGA; Golino & Epskamp, 2017) in the network psychometrics literature.

Across the broader field of network psychometrics, several methods have been developed to identify differences in network structure (Van Borkulo et al., 2022; Williams et al., 2020) and sub-groups (Danaher et al., 2014; Haslbeck & Bork, 2022; Jones et al., 2020). Although these methods aim to identify differences between networks, they all tend to treat the networks as unidimensional—that is, these methods do not account for the community structure of the network. Therefore, unless the construct is assumed to be unidimensional, the detected differences are unlikely to parallel traditional measurement invariance procedures. Establishing community structure and a metric consistent with factor loadings is key for developing such a comparison method.

Recent work has demonstrated that a node's strength or absolute sum of a node's connections to other nodes is related to confirmatory factor analysis (CFA) loadings (Hallquist et al., 2021). Hallquist et al. (2021) found, however, that the strength of a node is comprised of both dominant and cross loadings. To circumvent this issue, Christensen and Golino (2021b) proposed a measure called *network loadings* that splits a node's strength based on the dimensions identified by EGA. In their simulation, they found that this measure was consistent with factor loadings when data were generated by a factor model. The development of network loadings opened the door for broader measurement evaluation within network psychometrics such as item selection, weighted between-person scores, and hierarchical dimensionality assessment (Christensen & Golino, 2021b; Jiménez et al., 2023).

The goal of this study is to leverage these network loadings to establish a method within the network psychometric framework to test for measurement invariance. The extent of measurement invariance within the network psychometrics framework only includes configural and metric invariance because latent variables are not estimated using networks. Consequently, intercepts and residual variances are not feasible because there are no latent factors created. Therefore, measurement invariance using network psychometrics, like network loadings, is a heuristic for configural and metric invariance in latent factors rather than a direct equivalent.

Present Research

Configural Invariance

Before introducing the proposed method to test metric invariance, configural invariance must be established first. Configural invariance in the EGA framework exists when the same nodes have been partitioned into the same communities for all groups. This task can be initially tested in a cursory way by estimating EGA separately for each group and comparing their structures. Even if the initial structure as defined by EGA indicates configural invariance, further testing should be conducted to minimize any effects of sampling variability. In other words, additional testing should be conducted to test if items are consistently organized into the same communities or if the number of communities and their structure fluctuates.

Bootstrap EGA (bootEGA; Christensen & Golino, 2021a) produces a sampling distribution of EGA results that can be used to evaluate the stability of the identified structure. One statistic called *structural consistency* assesses the proportion of bootstraps in which the same *exact* structure as the initial EGA was recovered. If the groups are pooled together into one sample, higher *structural consistency* indicates that it is more likely for this structure to be representative of the population structure for all groups. Lower *structural consistency* indicates that configural noninvariance may be present. Additionally, if varying the number of samples drawn in bootEGA (or even which specific samples are drawn) shows structural variation, configural noninvariance may be present.

Structural consistency can be further broken down to assess the stability of items (proportion of bootstraps in which an item was assigned to the same dimension as the original EGA). Items showing a stability of $< .70$ are considered to be unstable (Christensen & Golino, 2021a). If there are distinct groups in a sample, then it is expected that each resample will have a different proportion of cases from each group. Therefore, if each group has a different configuration of assignment of nodes to communities, this lack of configural invariance will appear as items showing instability in community assignment. To reach configural invariance, items displaying instability ($< .70$) should be removed.

To test for configural invariance in the EGA framework, a straightforward approach is to conduct bootEGA on the entire sample and remove items with $< .70$ stability. Without these items, bootEGA can be re-applied to identify any further items contributing to instability. This process should be repeated until a consistent common structure (i.e., all item stabilities > 0.70) can be identified across all groups within a sample. Importantly, this approach does not allow for partial configural invariance.

Metric Invariance in the EGA Framework

Once configural invariance is established, metric invariance can be tested. The proposed method tests the equivalence of network loadings across groups via permutation testing.

Permutation testing has many advantages over traditional hypothesis testing approaches. Permutation tests make no parametric assumptions about populations, making them more flexible and robust to parametric deviations (Chihara & Hesterberg, 2022). Further, permutation tests can be applied to any test statistic, providing flexibility to adapt the model to any hypothesis or statistic (Chihara & Hesterberg, 2022; Ludbrook & Dudley, 1998). To elaborate on our procedure, we first must define network loadings.

Let $\mathbf{W}$ represent a symmetric $v \times v$ network made up of edge weights (e.g., partial correlations) where v is the number of variables. Node strength is then defined as

$$S_i = \sum_{j=1}^{n} |w_{ij}|, \tag{8}$$

where $|w_{ij}|$ is the absolute weight between node i and j and S_i is node i's strength or the sum of the absolute weights between node i and all n other nodes. Node strength can then be split between the communities identified by EGA:

$$\ell_{ic} = \sum_{j \in c}^{C} |w_{ij}|, \tag{9}$$

where ℓ_{ic} is the sum of the edge weights in community c that are connected to node i (i.e., node i's loading for community c), and C is the number of estimated communities.

This formulation computes the absolute sum of a node's connections to each community resulting in within (assigned) and between (non-assigned) community strengths. In other words, a node's strength is divided into its connections to each community in the network. Equation (9) can be standardized using the following formula:

$$\aleph_{ic} = \frac{\ell_{ic}}{\sqrt{\sum \ell_c}}, \tag{10}$$

where $\sqrt{\sum \ell_c}$ is equal to the square root of the sum of all the weights for the nodes in community c.

Standardized loadings, $\aleph$, are absolute weights and, as is done in factor analysis, the signs are added after the loadings are computed (Comrey & Lee, 2013). However, unlike factor analysis, the number of communities is extracted from the network's structure before computing network loadings. Additionally, variables have already been assigned to a community rather than being assigned to the community where to which they have the highest loading (as is done in factor analysis). Due to a network's sparsity (i.e., lack of edges between some nodes), it is possible for a node to have a network loading of zero to some communities because it does not have any connections to nodes in those communities.

To test the equivalence of network loadings across groups, we propose applying a permutation test which works as follows. The original $o \times k$ data, D (where o is sample size and k is number of variables), is split by grouping variable G into two groups, G_1 and G_2, to form two new datasets, D_1 and D_2, respectively. EGA is performed separately on D_1 and D_2. In order for further testing to occur, the community structure as identified by EGA must be identical for both D_1 and D_2 (i.e., configural invariance). Once established, corresponding $k \times c$ network loading matrices $\aleph_1$ and $\aleph_2$ are computed where c is the number of communities. The difference between the two matrices is then computed,

$$\mathcal{T} = \aleph_1 - \aleph_2, \tag{11}$$

to form an $k \times c$ matrix $\mathcal{T}$ which contains the difference for each network loading. Only the differences between assigned community loadings are retained, representing a vector of assigned loadings, τ. To form a null distribution for each loading difference to be compared to, the grouping variable G is permutated and becomes G_R and the original data D is split by G_R to form two new datasets, D_{R1} and D_{R2}, thereby removing the original relationship between item responses and group membership. This process is done repeatedly a P number of times, $p = 1, ..., P$, creating P new datasets, D_{R1_p} and D_{R2_p}.

EGA is performed on each permuted dataset D_{R1_p} and D_{R2_p}, network loadings are computed, and the difference between the assigned network loadings for each item is calculated to create a vector τ_{R_p} representing the null distribution for each item as follows:

$$\mathcal{T}_{R_p} = \aleph_{R1_p} - \aleph_{R2_p}, \tag{12}$$

where $\mathcal{T}_{R_p}$ represents a $v \times c$ matrix of differences between loading matrices, $\aleph_{R1_p}$ and $\aleph_{R2_p}$. From $\mathcal{T}_{R_p}$, only the assigned community loadings are retained, forming τ_{R_p}. Within each variable, v, these differences are put in ascending order, $\tau_{R_{1k}} \leq ... \leq \tau_{R_{pk}}$, forming a null distribution of the difference in network loadings if there was a random relationship between group assignment and network loading. The final step is to compare each test statistic to their respective null distributions at $\alpha = .05$. p-values for item invariance we calculated as follows:

$$p_{value} = \frac{1}{P} \sum_{p=1}^{P} s_p, \text{ where } s_p = \begin{cases} 1, & \text{if } |\tau_{R_p}| \geq \tau \\ 0, & \text{otherwise} \end{cases}. \tag{13}$$

This formulation of p_{value} is a vector whose elements are a two-sided p-value for each respective variable. If any p-value is less than .05, then metric invariance was violated. If not all p-values are less than .05, then partial metric invariance has been found; however, as previously mentioned, there is no agreement in the literature, to our knowledge, as to what constitutes an acceptable level of partial invariance.

The method described above is specifically outlined for the comparison of two groups. Conveniently, this model can be easily extended to three or more groups without sacrificing computational efficiency. Similar to logic used when conducting multiple comparisons after an omnibus test (Maxwell et al., 2018), it stands to reason that if noninvariance were to be found using this method, then it would be found between the groups with the largest difference in loadings. Therefore, for each variable we need only identify the groups with the minimum and maximum network loadings. If these two groups are significantly different from one another, then invariance cannot be supported. In this way, this method runs the same number of tests regardless of how many groups are being assessed. If noninvariance is found for an item, should the researcher wish, follow up tests can continue to be conducted to identify which groups specifically are different from one another. For each variable, the minimum loading would be compared to the second highest loading. If noninvariance is again found, the minimum loading would be then compared to the third highest loading, so on and so forth, until no significant differences are found.

Method

The following section outlines the methods used for each portion of the simulation study. As a benchmark for our proposed method, we compare it to the procedure presented by Raykov et al. (2013), which is outlined first. After, we discuss the multiple comparison procedure (aforementioned BH-procedure) and how it is applied in the current study. The data generation, conditions, and evaluation metrics are also described.

SEM Procedure

To test metric and partial metric invariance using SEM, we estimated two models: a configural, unconstrained model, see Equation (4), and a model with loadings constrained to equality across k populations, constrained metric model, see Equation (6). In order to directly compare to our proposed method, we tested for partial metric invariance. Testing for partial metric invariance was conducted using three methods: Free, Fixed, or Wald. The Free method follows the method proposed by Raykov et al. (2013). Using the {semTools} package (Version 0.5-6; Jorgensen et al., 2022), the Fixed and Wald methods are run simultaneously with Free. Since it is a common software method for researchers to use in practice, we evaluated the results of all three approaches. In all methods, an original model was chosen to be either the constrained or unconstrained model. After, the loadings were either fixed or freed iteratively to create a new model which was then compared to the original model. Using these methods circumvented a common problem in many approaches to invariance testing: we did not exclude the testing of any variables

by fixing the loading of one variable per factor to 1. In this way, we could make direct comparisons between the SEM and proposed methods.

The Free method uses the constrained model as the original model. Iteratively, each variable j is freed in the matrix Λ to create $\mathcal{J}$ models. Each model is then compared to the original model using a likelihood ratio test and an assessment of CFI for a total of $\mathcal{J}$ tests. The Fixed method uses the unconstrained model as the original model. Iteratively, each variable j is constrained to be equal across populations k to create $\mathcal{J}$ models. Each model is then compared to the original model using a likelihood ratio test and an assessment of CFI for a total of $\mathcal{J}$ tests. The Wald method is similar to Free. It uses the constrained model as the original model, but rather than iteratively freeing each variable j and conducting likelihood ratio tests, it uses a multivariate Wald test. Nonetheless, multiple hypotheses are being tested. These methods do not adjust for Type I error rate and so it is often necessary to apply a multiple comparison test (Raykov et al., 2013).

Multiple Comparison Problem

Within both partial invariance frameworks (EGA and SEM), multiple hypotheses are tested which can artificially inflate the Type I error rate. To adjust for this inflation, a multiple comparison procedure (MCP) can be applied (Raykov et al., 2013; Steinberg, 2001). To select which MCP to apply, it's important to consider consequences in the trade off of identifying (non)invariant items. Most MCPs focus on controlling the Family Wise Error Rate (FWER). FWER attempts to avoid making any Type I error and is inclined toward the notion that a Type I error is a serious issue.

In the context of partial invariance, a Type I error would suggest that a variable is noninvariant when it is truly invariant. In most research contexts, the cost of falsely identifying an item as invariant is greater than falsely identifying an item as noninvariant, particularly if the construct will be used to compare across groups (Shi et al., 2019). Therefore, FWER may suggest that more variables are invariant than there truly are, potentially leading to more costly consequences than using an uncorrected p-value. An alternative and less conservative MCP is the Benjamini-Hochberg procedure (BH-procedure; Benjamini & Hochberg, 1995) which controls the False Discovery Rate (FDR). FDR takes a more balanced approach to the multiple comparison problem by using the expected number of falsely rejected null hypotheses if any null hypotheses are rejected to determine its correction. Formally, FDR works to control ϕ:

$$\phi = \frac{V}{V + S} \tag{14}$$

where V is the number of falsely rejected null hypotheses and S is the number of correctly rejected null hypotheses out of the set of all hypotheses tested. The BH-procedure provides adequate control over false positives while showing marked improvements in

power above and beyond traditional MCP methods (e.g., Tukey, Bonferroni, Scheffe; Benjamini & Hochberg, 1995).

The BH-procedure works by sorting individual *p*-values in ascending order and assigning them a rank. The adjusted *p*-value is computed using:

$$p_{adjusted_i} = \min(1, \min_{j \geq i} \frac{mr_j}{j}), \tag{15}$$

where *mr* represents the total number of *p*-values and *r* represents to rank for the corresponding *p*-value *j*.

Raykov et al. (2013) first proposed the use of the BH-procedure to test partial invariance due to its more liberal approach of lowering the risk of false positive noninvariant variables relative to FWER's focus on lowering the risk of false positive noninvariant variables entirely. Given that the consequences are usually more dire when *not* correctly detecting noninvariant variables, the BH-procedure was preferred over FWER as our MCP. All *p*-values calculated using the BH-procedure, hereafter, are referred to as *corrected p*-values.

Data Generation

Data was generated following a common factor model, as was done by Golino et al. (2020). We begin by computing a population correlation matrix for each group, $\mathbf{R}_{\mathbf{R}_G}$, with communalities in the diagonal,

$$\mathbf{R}_{\mathbf{R}_G} = \Lambda_G \Phi \Lambda_G', \tag{16}$$

where $\mathbf{R}_{\mathbf{R}_G}$ is the reproduced population correlation matrix for each group *G*, Λ_G is a $k \times m$ factor loading matrix for *k* variables and *m* factors for each group *G*, and *Phi* (Φ) is the structure matrix of the latent variables (i.e., a $m \times m$ matrix of correlations among factors). The population does not contain any correlated residuals and therefore no minor factors.

Then, by inserting unities in the diagonal of $\mathrm{R}_{\mathrm{R}_G}$ it becomes a full rank matrix and is now population correlation matrix $\mathrm{R}_{\mathrm{P}_G}$. Each group in *G* is assigned a $\mathrm{R}_{\mathrm{P}_G}$ matrix. A Cholesky decomposition is performed on each $\mathrm{R}_{\mathrm{P}_G}$:

$$\mathbf{R}_{\mathbf{P}_G} = \mathbf{U}'\mathbf{U}. \tag{17}$$

If any $\mathbf{R}_{\mathbf{P}_G}$ is not semi-positive definite or an item's communality is greater than 0.90, then a new $\mathbf{R}_{\mathbf{P}_G}$ matrix is constructed. From this, the sample data matrix (continuous variables) can be computed as:

$$\mathbf{X}_G = \mathbf{Z}_G\mathbf{U}_G, \tag{18}$$

where $\mathbf{Z}_G$ is a matrix of random standard normal deviates with rows equal to the sample size and columns equal to the number of variables.

Design

The overall design of the simulation study followed closely that of Kim and Yoon (2011) with a few modifications. A two-factor model was simulated, each factor containing six variables, similar to Kim and Yoon (2011) and Yoon and Millsap (2007). Typically, simulation studies investigating invariance methods use unidimensional models; however, we decided to simulate two factors. This approach allowed us to manipulate interfactor correlation and investigate whether it impacted the power of the proposed method. Only one variable in one factor was simulated to have unequal dominant loadings across group. Since our main goal was to assess each method's ability to identify noninvariant items correctly, having only one noninvariant item allows for a direct estimate of the true positive rate. Additionally, it allows us to compare the ability of each method to detect invariant items within factors both with and without noninvariant items.

For simplicity, we only simulated two groups. Factor loadings were set to be the same across factors for each respective variable (0.80, 0.70, 0.60, 0.80, 0.70, 0.60). Keeping high, static factor loadings allowed us to make sure configural invariance was not negatively impacted, particularly for data conditions with a high difference in loadings and/or a high interfactor correlation. Similar to Golino and Epskamp (2017), the correlation between factors was set to be low (0.30), medium (0.50), or high (0.70).

The loading of Variable 5 in Factor 1 (0.70) was decreased in G_1 by either 0.20 (small difference) or 0.40 (large difference) as was done in Kim and Yoon (2011). Static factor loadings ensured that the magnitude of loading differences will have the same interpretation across data conditions (Yoon & Millsap, 2007). Per group, there was either the same sample size per group (500 or 1000 in both G_1 and G_2) or different sample sizes per group (500 in G_1 and 1000 in G_2). This design allowed us to compare the new method's ability to detect noninvariant items in conditions that traditional methods currently usually struggle (i.e., disparate and/or small sample). This resulted in 18 separate conditions. For each condition, 500 datasets were simulated.

Measurement invariance was tested on each simulated dataset using both EGA in the {EGAnet} package (Version 1.1.1; Golino & Christensen, 2022) and SEM using the {lavaan} (Version 0.6.17; Rosseel, 2012) and {semTools} in R. All analyses were conducted in R and full code can be found in Jamison, Golino, and Christensen (2024).

Data Analysis

To assess the accuracy of each model's (non)invariance detection, we used confusion matrix metrics. Because loadings were only changed for one variable (Variable 5 in

Factor 1) and all other variables had equivalent loadings in the population, noninvariance should only be detected for Variable 5. Therefore, a true positive (*TP*) occurs when the model identifies noninvariance in Variable 5, and a false positive (*FP*) occurs when any other variable is identified as noninvariant. A false negative (*FN*) occurs when the model identifies Variable 5 as invariant and a true negative (*TN*) occurs when any other variable is identified as invariant. An item is considered noninvariant if its significance is $p < 0.05$ and invariant if it's significance is $p \geq 0.05$.

The following confusion matrix metrics were used to identify more specific measures of accuracy: *Hit Rate*, *Sensitivity*, *Specificity*, and *F*1. The {caret} package (version 6.0.94; Kuhn, 2022) in R was used to calculate *Sensitivity*, *Specificity*, and *F*1. All metrics were calculated separately using both uncorrected and corrected (using the BH-procedure) *p*-values.

Hit Rate, or $\frac{TP+TN}{TP+FP+TN+FN}$, provides a straightforward, overall assessment of method accuracy of correctly identifying invariance or noninvariance. *Sensitivity* or $\frac{TP}{TP+FN}$ represents the proportion of true positives correctly identified by the method out of all the truly noninvariant items. *Specificity*, or $\frac{TN}{TN+FP}$, represents the proportion of true negatives correctly identified by the method out of all true null hypotheses. *F*1, or $\frac{2TP}{2TP+FP+FN}$, provides a similar metric to *Sensitivity* but places greater emphasis on identifying Variable 5 as noninvariant relative to identifying the other variables as invariant.

It's important to contextualize these measures in our current simulation. Because there is only one possible *TP* or *FN* (i.e., Variable 5), *Sensitivity* breaks down to *TP*. *Hit Rate* either breaks down to $\frac{TN}{FP+TN+FN}$ if Variable 5 is identified as invariant or $\frac{TP+TN}{TP+FP+TN}$ when Variable 5 is identified as noninvariant. Similarly, *F*1 either breaks down to zero if Variable 5 is identified as invariant or $\frac{2TP}{2TP+FP}$ when Variable 5 is identified as noninvariant. Therefore, *F*1 is weighted toward identifying the noninvariant variable while lowering the (relative) cost of a FP. Finally, *Specificity* is a pure measure of the extent to which all invariant variables are correctly identified as invariant. Within the context of our study, greater weight should be given to detecting noninvariance over invariance. Therefore, *Sensitivity* and *F*1 should be given greater weight over *Specificity* and *Hit Rate*.

Results

We assessed method accuracy overall as well as across simulation conditions. When indicating simulation conditions, we will use the following labels: "Correlation Between Factors" indicates variation in the correlation between factors (0.3, 0.5, 0.7), "Diff" indicates the level of noninvariance (0.2 or 0.4), "N" indicates sample size (500, 1000, Different), and "p-Value" indicates the significance level where "corrected" indicates the BH-procedure adjusted *p*-values and "uncorrected" indicates the standard, unadjusted *p*-values.

Effect of MCP on p-Values

Configural invariance was recovered in 99.73% of the simulated datasets using EGA and 100% using SEM. To provide a direct, full comparison between both methods (rather than removing items showing configural noninvariance), only those datasets where configural invariance was found for EGA were retained and the others were discarded. All datasets were retained for analysis using SEM methods. Within each method, the accuracy of metric invariance methods were assessed. Figures 1 and 2 show the mean and 95% confidence interval across all datasets of the *p*-values split by method, sample size, correlation between factors, and loading difference. A dashed line intercepts the *y*-axis at .05 representing the α level. The mean *p*-value is represented for each variable.

Figure 1

Mean Uncorrected p-Value by Variable

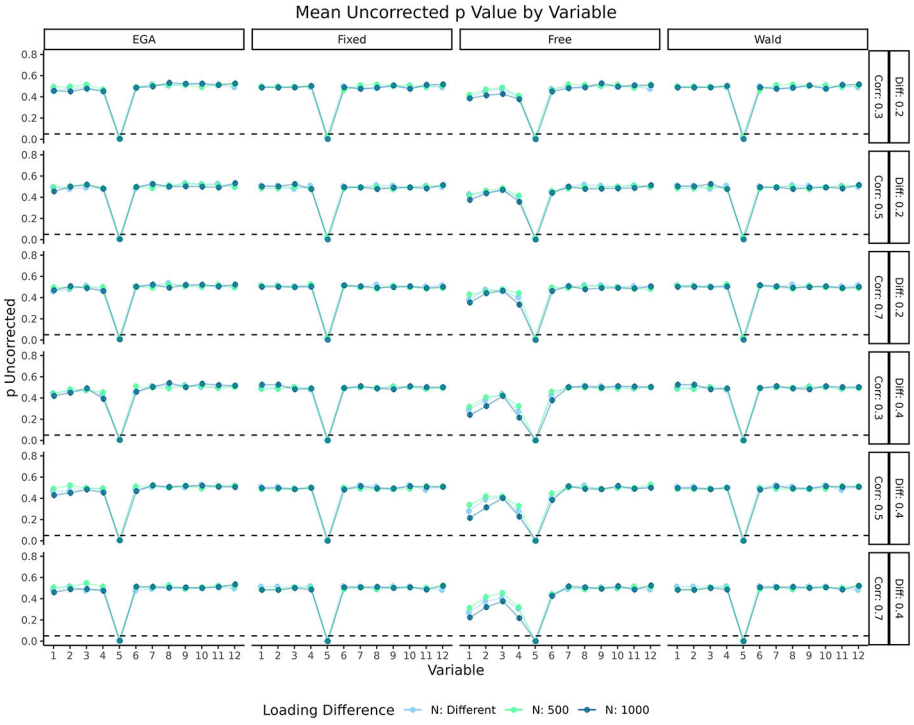
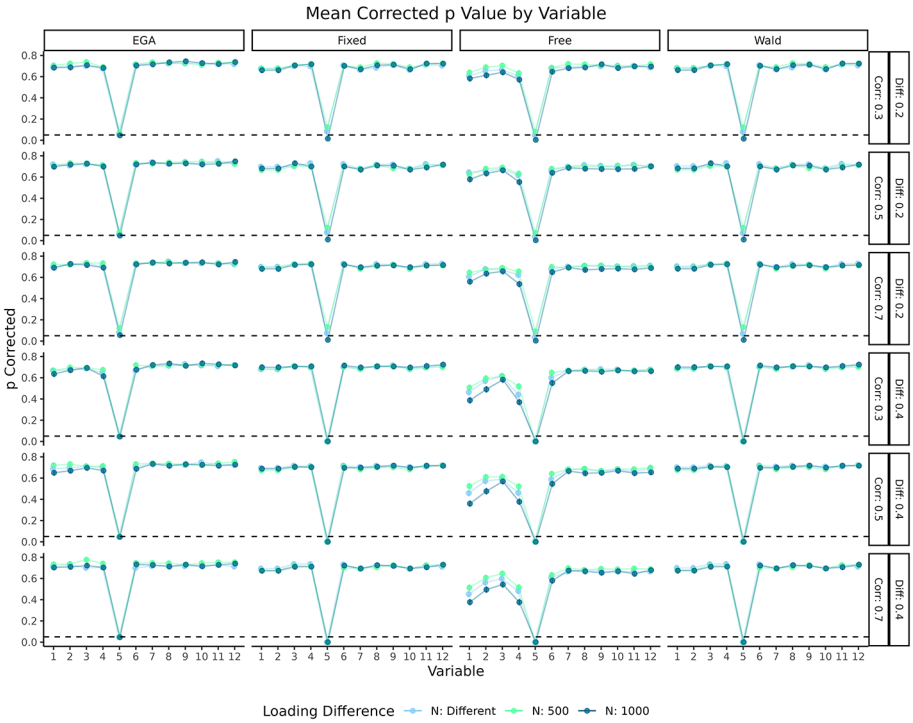


Figure 2

Mean Corrected p-Value by Variable



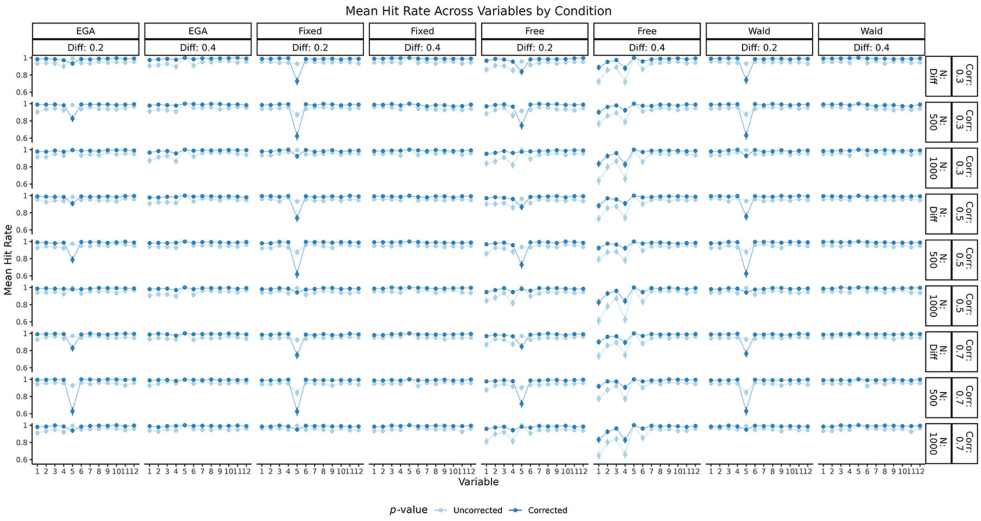
In both Figures 1 and 2, Variable 5 is the only variable which should be significant or should show a mean *p*-value consistently below .05. Across both uncorrected (Figure 1) and corrected (Figure 2) *p*-values for all four methods, regardless of condition, the lowest mean *p*-value across variables is indeed Variable 5. This is in line with the manipulation used, changing the loading between groups by either 0.2 or 0.4 for only Variable 5. When the *p*-value is not corrected, as in Figure 1, the Free method has a lower mean *p*-value for all variables in Factor 1 (where noninvariance was simulated to exist), but not in Factor 2 where no noninvariance was simulated. This pattern is not present for the other methods. When the *p*-value is corrected (see Figure 2), the average *p*-value is higher than when it is uncorrected regardless of whether an item is invariant or not. All 3 SEM methods have a more noticeable increase in average *p*-value for Variable 5 when the difference in loadings for Variable 5 is set to 0.2 and sample size is either different or 500. Under these same conditions, this same trend in EGA is only noticeable when the correlation between factors increases to 0.7.

Hit Rate

In almost all cases, corrected *p*-values produce a higher mean *Hit Rate* than uncorrected *p*-Values (Figure 3). When the difference in loadings is 0.4, EGA, Fixed, and Wald all have almost perfect *Hit Rate* across all variables. In this condition, the same trend arises in Free as was seen in Figures 1 and 2: mean *Hit Rate* is lower in general for items in Factor 1, however its level of mean *Hit Rate* for Factor 2 when noninvariance is not present, is more similar to that of the other three methods.

Figure 3

Hit Rate by Condition



When the difference in loading is set to 0.2, for Variable 5, all four methods experience lower mean *Hit Rate* when the *p*-value is corrected as compared to the uncorrected *p*-value. This trend is most notable for Fixed, Free, and Wald when sample size is "Different" or 500, but does not appear when sample size is 1000. EGA only shows this trend when sample size is 500 and gradually becomes more disparate as the correlation between factors increases from 0.3 to 0.7. The magnitude of this effect is the same for Fixed, Free, and Wald regardless of the correlation between factors. This indicates that EGA's ability to correctly identify noninvariant variables is not as heavily influenced by data structures as Fixed, Free, and Wald. The Free method is better able to accurately identify invariant variables when noninvariant items are not present in the same factor.

Overall Metrics

Looking at the accuracy methods overall (not split by simulation condition), when a correction is applied, an interesting pattern appears (Table 1). Both *F*1 and *Specificity* increase for all four methods, but *Sensitivity* decreases. Using uncorrected *p*-values, *Sensitivity* is nearly 1 for all four methods, EGA being the highest at 0.99 and Fixed the lowest at 0.96. Once the BH-procedure is applied, *Sensitivity* decreases for all four methods, most dramatically for Fixed and Wald, falling below 0.90. When using corrected *p*-values, EGA has the highest values for *F*1 (0.91) and is tied for the highest *Specificity* with Fixed and Wald at 0.99. Free has the lowest values (using corrected *p*-values) of all of four methods for both *F*1 (0.83) and Sensitivity (0.97). For all four methods, *Sensitivity* increased slightly by applying the BH-procedure, while *F*1 values dramatically increase by applying the BH-procedure, going up on average by 0.14.

Table 1

Overall Metrics by Method

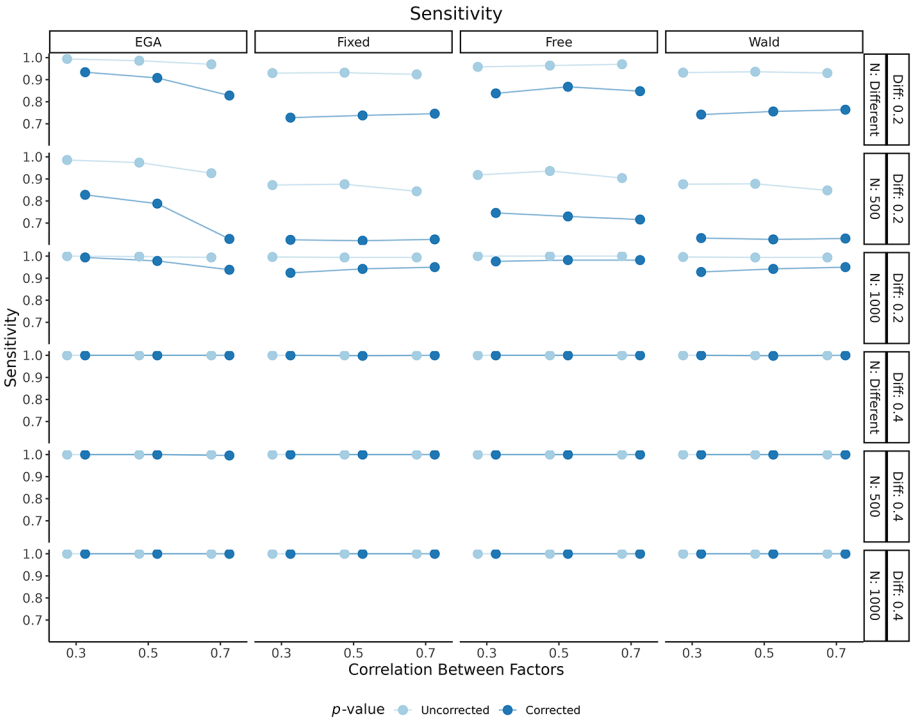
	Sensitivity		F1		Specificity	
Type	**Uncorrected**	**Corrected**	**Uncorrected**	**Corrected**	**Uncorrected**	**Corrected**
EGA	0.99	0.93	0.76	0.91	0.94	0.99
Fixed	0.96	0.88	0.76	0.88	0.95	0.99
Free	0.98	0.93	0.65	0.83	0.90	0.97
Wald	0.97	0.89	0.77	0.89	0.95	0.99

Sensitivity

When the difference in loadings is 0.4, all methods in all conditions have perfect *Sensitivity* regardless of whether or not the *p*-value is corrected (Figure 4). When the difference in loadings is 0.2, uncorrected *p*-values lead to a higher level of *Sensitivity*. In this condition, almost perfect *Sensitivity* is achieved using corrected *p*-values when sample size size is 1000 for all methods. When the difference in loadings is set to 0.2, corrected *p*-values are used, and sample size is either "Different" or 500, EGA and Free are performing better than Fixed and Wald. However, EGA is more heavily influenced by the increase in correlation between factors; when the correlation between factors reaches 0.7, EGA's performance falls below Free's to the same level as Fixed and Wald. Though when the correlation between factors is 0.3 or 0.5, EGA outperforms Free. All in all, setting the difference in loadings to 0.4 does not affect the ability of any of the methods to identify *TP*'s (noninvariant variables). However, when the difference is lower, correcting *p*-values lowers the *Sensitivity* for all the methods. EGA is, again, less affected by this difference and sample size in its ability to detect *TP*'s except when the correlation between factors is high.

Figure 4

Sensitivity by Condition

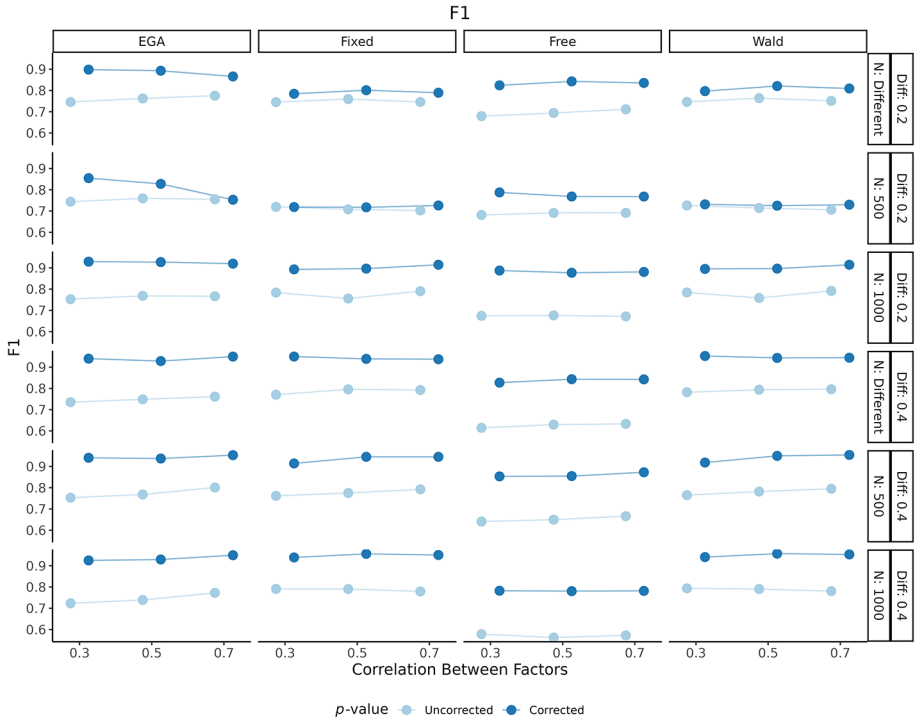


F1

In all conditions and across all four methods, corrected *p*-values produce higher *F*1 values than uncorrected *p*-values (Figure 5). When the difference between loadings is set to 0.4, EGA, Fixed, and Wald have similar (and nearly perfect) F1 values. Free, however, has lower F1 values in this condition than the other three methods, particularly when the sample size is increased to 1000. When the difference between loadings is set to 0.2 and F1 is calculated using corrected *p*-values, a similar pattern arises that was seen in *Sensitivity*. EGA outperforms the other three methods when sample size is "Different" or 500. However, EGA is more heavily influenced by the increase in correlation between factors; when the correlation between factors reached 0.7, EGA's performance falls below Free's to the same level as Fixed and Wald. When the correlation between factors is 0.3 or 0.5, EGA outperforms Free.

Figure 5

F1 by Condition

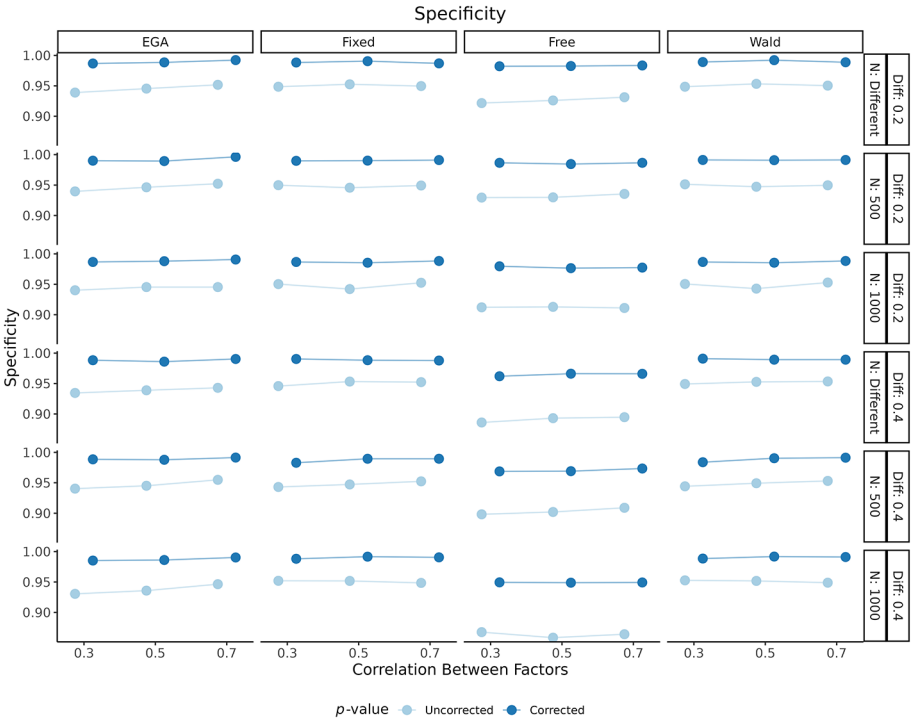


Specificity

Across all these conditions, *Specificity* calculated using corrected *p*-values is higher than uncorrected *p*-values (Figure 6). All methods have consistently high and comparable levels of *Specificity*, except for the same trend that has been appearing for Free. When the difference in loadings increases from 0.2 to 0.4, the *Specificity* for the Free method decreases. Altogether, this indicates that each method is able to comparably recover *TN*'s or invariant items (except for the Free method in one condition).

Figure 6

Specificity by Condition



Applying the Test for Metric Invariance to the BAPQ

To demonstrate a substantive application of this approach, we apply our proposed test for metric invariance to the Broad Autism Phenotype Questionnaire (BAPQ; Hurley et al., 2007). Appendix B contains the results from the application of the traditional partial invariance SEM method as implemented by the `partialInvariance()` function in {semTools}. Data was obtained from the Simons Foundation Powering Autism Research for Knowledge (SPARK) of the Simons Foundation Autism Research Initiative (SFARI), a large research initiative which has collected data from over 50,000 individuals with autism and their families (Feliciano et al., 2018). The BAPQ is a 36-item questionnaire designed to assess autism-related traits in adults. Participants are asked to rate the how often a statement applies to them on a 6-point Likert scale ranging from *(1) Very Rarely* to *(6) Very Often*. Items were intended to relate to one of three domains: aloofness, rigid personality, or pragmatic language.

This questionnaire was given to the parents (either mother or father) of an autistic child to assess their phenotypic level of autistic traits. We begin assessing measurement

invariance between mothers and fathers by establishing configural invariance. To do so we apply EGA separately to the data on mothers and the data on fathers and compare their community structures. For this example, we are using the {EGAnet} package (Version 2.0.6; Golino & Christensen, 2024).

```
# Load EGAnet Package
library(EGAnet)

# Load in the Data
load("../2. Data/bapq.all.RData")

# Set mother indices
mother <- bapq.all$Parent == "Mother"

# Extract items only
items <- bapq.all[,4:39]

## Mother
ega.mother <- EGA (data = items[mother,])

## Father
ega.father <- EGA(data = items[!mother,])
```

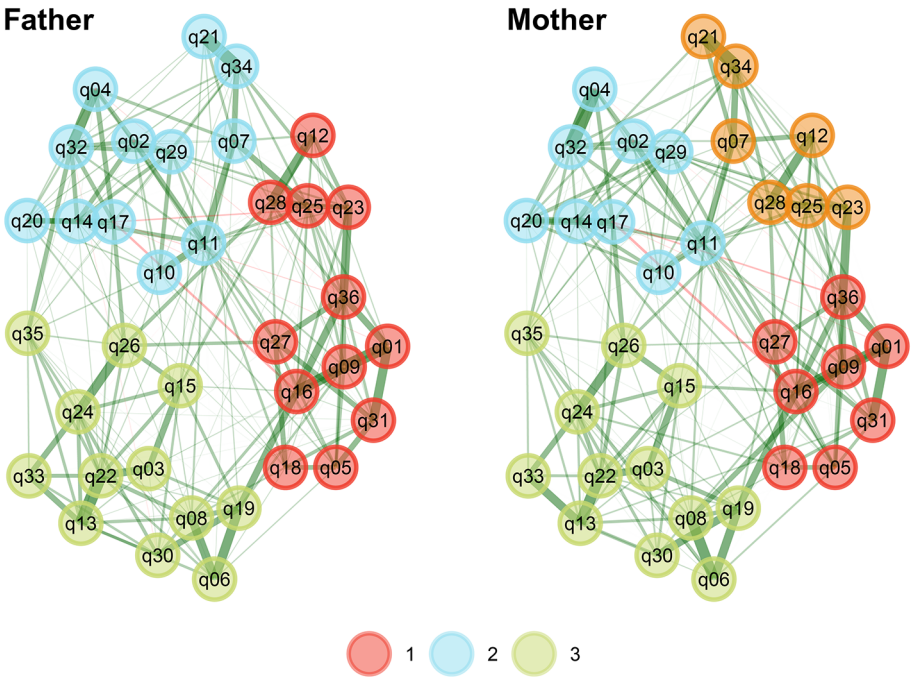
Visually we can see that the two graphs contain nonequivalent community structures (Figure 7). We can apply the `invariance()` function to the data, which will first identify a common structure that exists using `bootEGA()`, removing item stabilities less than 0.70, to establish configural invariance. After, establishing configural invariance, the procedure will continue to test metric invariance.[1]

```
# Perform invariance
bapq_invariance <- invariance(
  data = items, group = bapq.all$Parent,
  ncores = 8, seed = 1, loading.method = "experimental"
)
```

1) In this example, we are demonstrating the use of a recently developed revised form of network loadings. See Christensen et al. (2024) for more details. We demonstrate that the results of the simulation do not differ for these revised network loadings (see Supplemental Materials).

Figure 7

Comparison of the EGA Networks for the Mothers (Left) and Fathers (Right) With Color Denoting the Community for Each Node



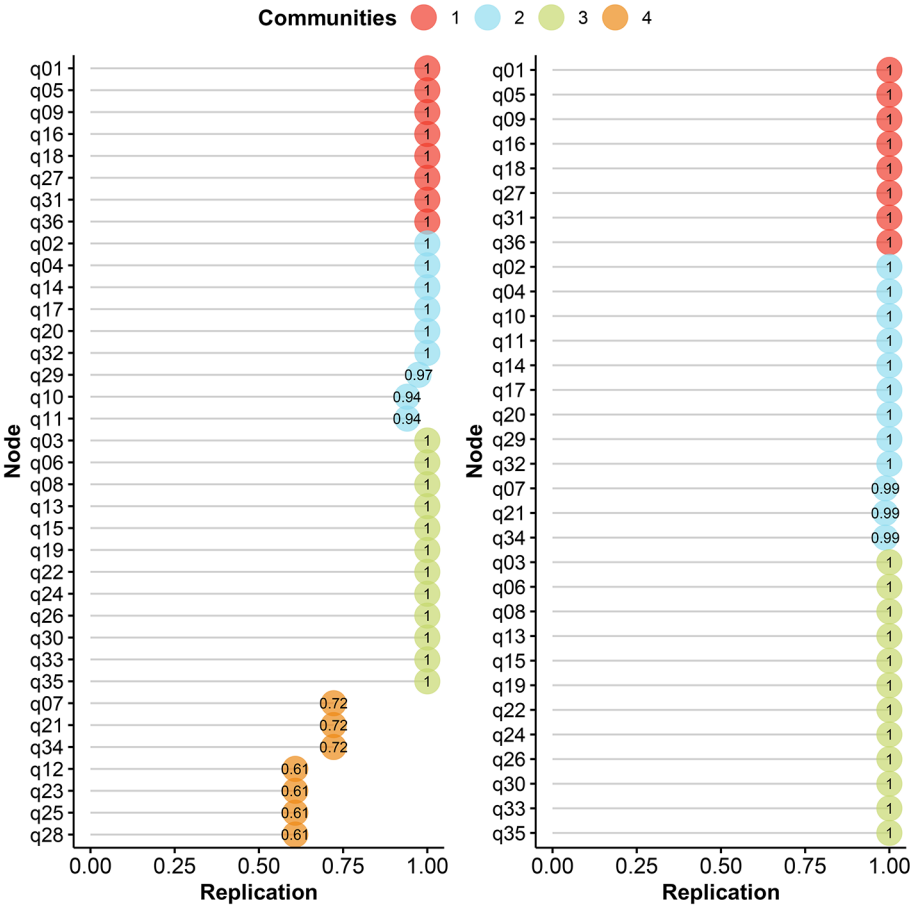
The function will print out how many items were identified for configural invariance, for example: `Configural invariance was found with 32 variables`. To view which items were removed from the the original 36, the following object can be accessed:

```
[1] "q12" "q23" "q25" "q28"
```

In Figure 8, the before and after item stability are plotted with the latter being accessed in the results using `plot(bapq_invariance$configural.results$item_stability)`.

Figure 8

Comparison of the BAPQ Item Stability Before (Left) and After (Right) Configural Invariance



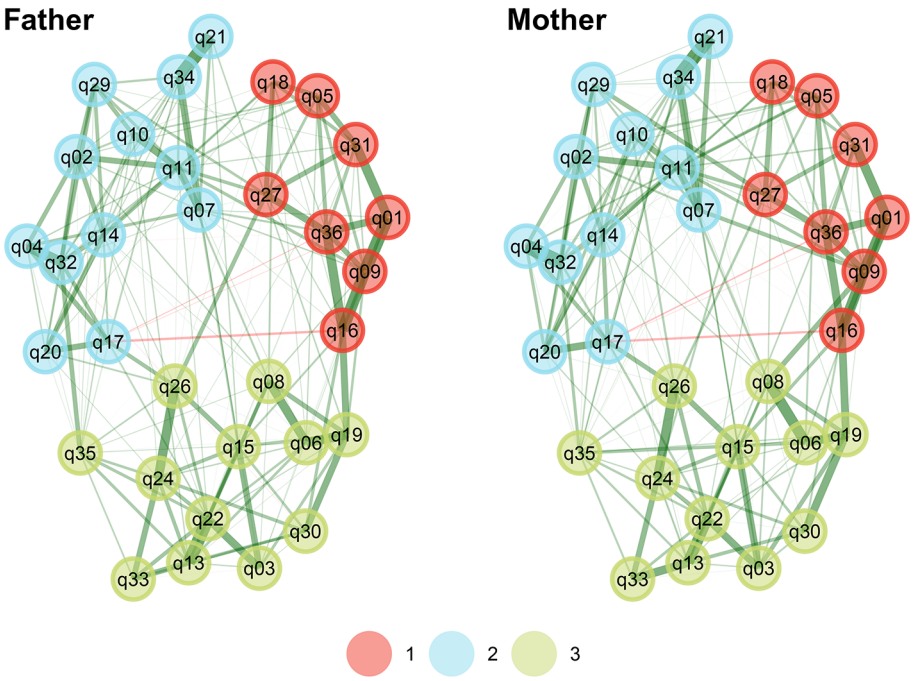
Evaluating each group separately, we can see that the both groups have equivalent structures (Figure 9):

```
## Father
ega.father <- EGA(data = items[!mother, stable_names])

## Mother
ega.mother <- EGA(data = items[mother, stable_names])
```

Figure 9

Comparison of the EGA Networks for the Fathers (Left) and Mothers (Right) With Color Denoting the Community for Each Node



Finally, we can print the results of invariance to see a table that breaks down the metric invariance for each item:

```
# Print summary
summary(bapq_invariance)
```

	Membership	Difference	p	p_BH	sig	Direction
q01	1	-0.006	0.662	0.850		
q05	1	0.028	0.038	0.174	*	Father > Mother
q09	1	0.022	0.208	0.428		
q16	1	0.027	0.054	0.192	.	
q18	1	-0.001	0.948	0.979		
q27	1	0.030	0.076	0.243	.	
q31	1	0.033	0.030	0.174	*	Father > Mother
q36	1	-0.018	0.354	0.539		
q02	2	0.017	0.296	0.515		
q04	2	0.003	0.870	0.960		
q07	2	-0.021	0.214	0.428		

```
q10       2          -0.017     0.332    0.531
q11       2           0.038     0.038    0.174    *   Father > Mother
q14       2          -0.004     0.810    0.926
q17       2          -0.016     0.306    0.515
q20       2          -0.059     0.004    0.064   **   Father < Mother
q21       2          -0.043     0.004    0.064   **   Father < Mother
q29       2           0.044     0.006    0.064   **   Father > Mother
q32       2           0.019     0.214    0.428
q34       2           0.024     0.152    0.428
q03       3          -0.032     0.044    0.176    *   Father < Mother
q06       3           0.001     0.922    0.979
q08       3          -0.004     0.798    0.926
q13       3           0.010     0.564    0.785
q15       3          -0.017     0.264    0.497
q19       3          -0.007     0.664    0.850
q22       3           0.039     0.030    0.174    *   Father > Mother
q24       3           0.022     0.182    0.428
q26       3          -0.019     0.186    0.428
q30       3           0.000     0.984    0.984
q33       3           0.005     0.708    0.871
q35       3          -0.011     0.422    0.614
---
Signif. code: 0 '***'  0.001 '**'  0.01 '*'  0.05 '.'  0.1 'n.s.' 1
```

The item text for the 6 items showing metric noninvariance using uncorrected p-values are displayed in Table 2. Using the corrected p-values, there were no items that were detected as noninvariant. The noninvariant items detected with the uncorrected p-values spanned each dimension of the BAPQ. All differences between mothers and fathers corresponded with deficits in fathers relative to mothers—that is, there was larger loadings for items related to deficits or behaviors contrary to norms in the general population or smaller loadings for items related to norms in the general population.

Table 2

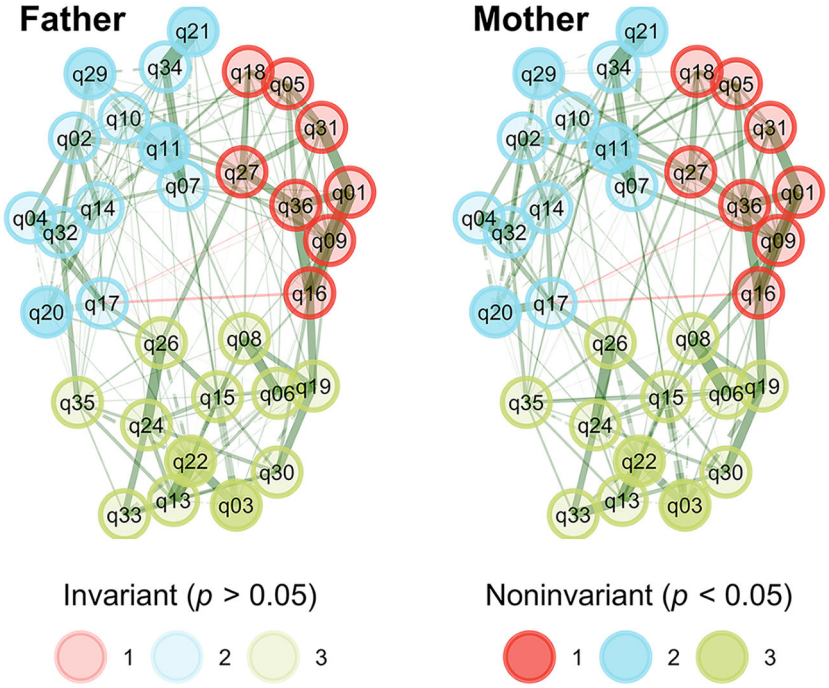
Metric Invariance Significant Results

Item Label	Item Description	p	p_{BH}	Direction
q11	I feel disconnected or 'out of sync' in conversations with others	.048	.256	Father > Mother
q20	I speak too loudly or softly	.002	.064	Father < Mother
q21	I can tell when someone is not interested in what I'm saying	.004	.064	Father < Mother
q29	I leave long pauses in conversation	.010	.107	Father > Mother
q03	I am comfortable with unexpected changes in plans	.040	.256	Father < Mother
q22	I have a hard time dealing with changes in my routine	.026	.208	Father > Mother

These results can further be visualized by using the `plot()` function on the output of `invariance` (Figure 10).

Figure 10

Comparison of the EGA Networks for the Fathers (Left) and Mothers (Right) With Color Denoting the Community for Each Node and Transparency Denoting (Non)invariance



Discussion

Establishing measurement invariance is crucial for the use of any measurement across groups in any clinical or research setting. Traditionally, SEM approaches are the most common methods for testing measurement invariance. Previous research within network psychometrics has established a handful of methods for comparing networks, but nothing comparable to SEM that accounts for multidimensionality. With the introduction of *network loadings* by Christensen and Golino (2021b), the space for further methodological development in network psychometrics has opened including the configural and metric invariance methods proposed in this study.

The simulation compared the proposed metric invariance method to existing SEM methods, manipulating sample size, loadings difference, and correlation between factors. Three methods in the SEM framework were used to test partial metric invariance: Fixed, Free, and Wald. In all four methods, we first tested for configural invariance, then tested for metric, and then partial metric invariance. A key addition to our comparisons was

the inclusion of a multiple comparison procedure. Most MCPs control Family Wise Error Rate (FWER) which is concerned with controlling the number of Type I errors made in general. Raykov et al. (2013) propose using the Benjamin-Hochberg procedure (BH-procedure) introduced by Benjamini and Hochberg (1995) to control the False Discovery Rate (FDR) when testing partial invariance. Since there is not a high level of risk in falsely identifying noninvariant items and there should be more emphasis on correctly identifying noninvariant items rather than invariant items.

The results of our simulation indicate that applying the BH-procedure provided a gain in the corrected identification of invariant items but not noninvariant items. This is in line with literature indicating that independent tests do not benefit from the application of an MCP (Rubin, 2024). Identifying noninvariant items was particularly challenging for the BH-procedure when the difference between loadings was small. Because smaller differences between groups will have larger *p*-values, there is a greater chance that detected differences will result in values at or near 0.05 which often end up with non-significant values after correction. For the corrected *p*-values in conditions with smaller differences, this poorer detection of noninvariant items was reflected in the *Hit Rate*.

When the difference in loadings was higher, all methods correctly identified the noninvariant item, regardless of *p*-value correction. But when the difference in loadings was lower, sample size and interfactor correlation differentially impacted the accuracy for the noninvariant item and the uncorrected *p*-value was more accurate in these particular cases. The EGA approach was less influenced by these cases than the three SEM methods. With a smaller sample size or different sample sizes, the EGA approach's accurate detection of noninvariant items was also better than the SEM methods. As the correlation between factors increased, however, the accuracy decreased when sample size was either "Different" or 500.

These results were further corroborated by *Sensitivity* or the ability to detect noninvariant items. All methods were better at identifying noninvariant items when the difference in loadings was larger. The EGA approach performed better than the SEM methods when sample size was "Different" or 500 but was negatively impacted by the increase in correlation between factors. Finally, the SEM Free method's detection of invariant items was different across the two factors: *Hit Rate* was lower for invariant items Factor 1, where noninvariance was present, and higher for the invariant items in Factor 2. *Specificity* indicated that all methods performed comparably at identifying invariant variables. Of the three SEM methods, the Free method showed the lowest accuracy across all metrics, except for *Sensitivity* where it showed a similar ability to correctly identify noninvariant variables.

Turning to the *p*-value correction, the results indicate that including a *p*-value correction provides a gain in the ability of each method to correctly identify invariant items, but in some instances may hinder their ability to correctly identify noninvariant items,

particularly for SEM. We believe that this latter consequence is problematic. The goal is often to detect whether noninvariance exists. In most cases, applied researchers would prefer to err on the side of caution when determining whether groups are equivalent. When *p*-values are corrected, there is a bias toward suggesting items are invariant.

This finding raises the question of whether correcting *p*-values is useful to apply in the context of metric invariance or if it hinders the ability of these methods to properly identify noninvariant items. The *Hit Rate* results by variable in particular indicate that uncorrected *p*-values are more accurate when the difference in loadings is lower and equally as accurate when the difference in loadings is higher. These results are paralleled by *Specificity* where there is little to no concerning effect of falsely identifying an item as noninvariant.

Our results, however, do not discount the utility of *p*-value correction in testing metric invariance. Instead, we recommend in practice that noninvariant variables identified both uncorrected and corrected *p*-values should be evaluated. Based on the results of both *p*-values, the researcher can determine the consequences associated with each result, leveraging their knowledge of the literature, research context, and research questions. Another alternative is to change the α level when applying the MCP. In Appendix A we have included the all results with an additional condition where the corrected *p*-values are assessed for significance at the $\alpha = 0.10$ level. The results indicate that this method slightly improves the accuracy of identifying noninvariant items for the EGA approach but makes no impact for the SEM methods.

The use of any latent variable measure across qualitatively distinct groups should necessitate the testing of measurement invariance. Current methodology for testing measurement invariance is problematic from a conceptual and software implementation standpoint. The proposed method is easier to implement in software than the existing methods. It also shows a stronger ability to correctly identify noninvariant items in several data conditions, namely differing sample sizes across groups or lower sample sizes within groups, especially when the correlation between factors is not very high.

Additionally, given that the communities estimated by EGA represent latent dimensions when the data generation mechanism is a latent variable model, EGA can still be applied even when this is not the case (Golino et al., 2022; Kjellström & Golino, 2019). Therefore, it is both important and intriguing to note that unlike existing measurement invariance methods using SEM, the proposed method does not necessitate a latent variable model as the data generation mechanism and could be used in other applications such as topic modeling.

Conclusion

Ensuring the equivalence of a measure across assessment groups is vital to the efficacy of group comparison. Although many methods have been proposed to improve measurement invariance in the SEM framework, many unresolved problems still remain. The

EGA approach proposed in this study performed comparably to existing SEM methods and, in several conditions, outperformed them with the aim of detecting noninvariant items. The method was then applied to a substantive dataset to demonstrate its assessment of metric invariance in a real-world dataset, finding important differences in the BAPQ inventory that exist between mothers and fathers of children with ASD.

Appendices

Appendix A

Overall Metrics

Table 3

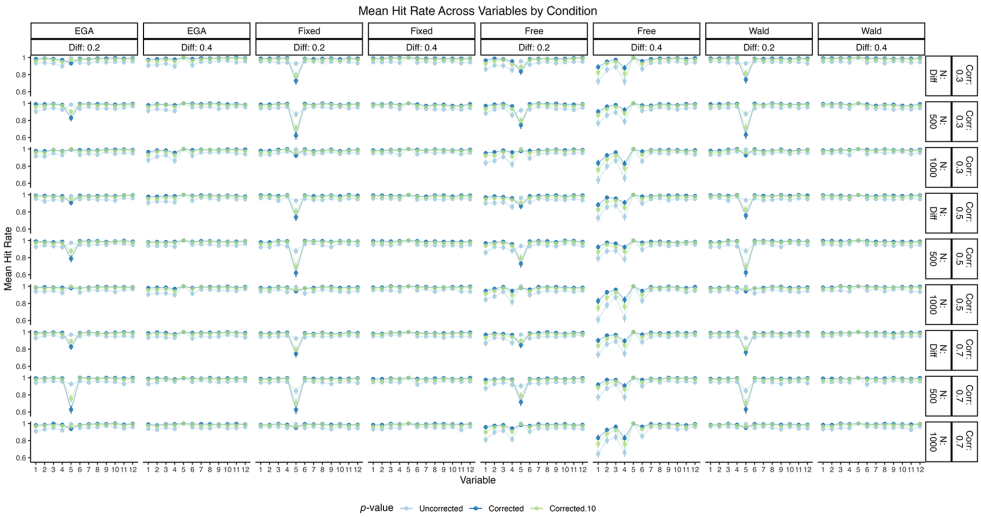
Overall Metrics by Method

	Sensitivity			F1			Specificity		
Type	Uncorrected a = .05	Corrected a = .05	Corrected a = .10	Uncorrected a = .05	Corrected a = .05	Corrected a = .10	Uncorrected a = .05	Corrected a = .05	Corrected a = .10
EGA	0.99	0.93	0.96	0.76	0.91	0.87	0.94	0.99	0.98
Fixed	0.96	0.88	0.91	0.76	0.88	0.84	0.95	0.99	0.98
Free	0.98	0.93	0.95	0.65	0.83	0.76	0.90	0.97	0.95
Wald	0.97	0.89	0.91	0.77	0.89	0.85	0.95	0.99	0.98

Hit Rate

Figure 11

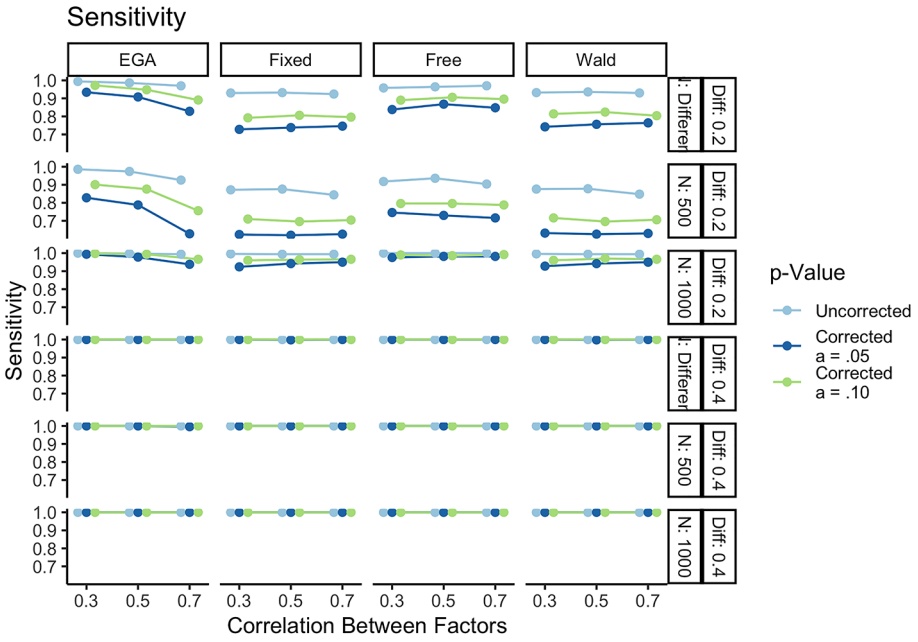
Mean Hit Rate by Condition



Sensitivity

Figure 12

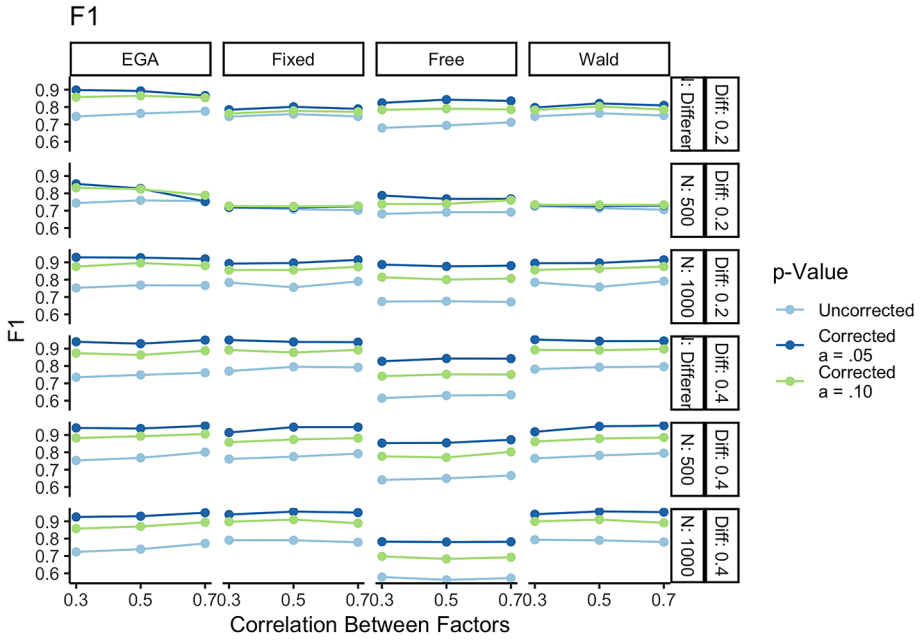
Sensitivity by Condition



F1

Figure 13

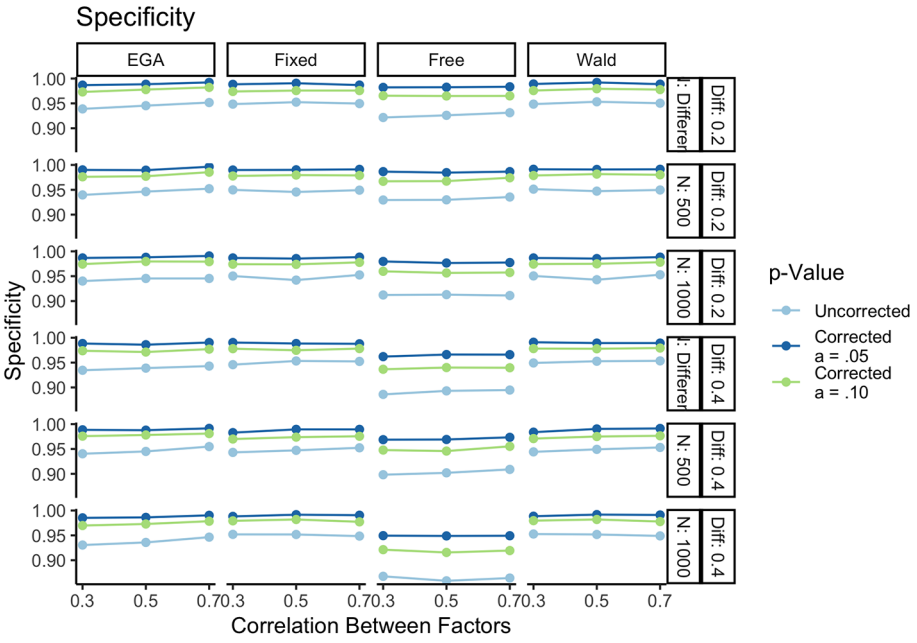
F1 by Condition



Specificity

Figure 14

Specificity by Condition



Appendix B

SEM Partial Invariance With the BAPQ

We applied a traditional SEM approach to test for partial metric invariance on the BAPQ dataset using the {lavaan} and {semTools} packages in R. First, we start with a three factor CFA model with the structure as outlined by Hurley et al. (2007) and Broderick et al. (2015). Table 4 shows the model fit statistics from this CFA model.

```
#Loading libraries
library (semTools)
library (lavaan)

# Confirmatory Three Factor Model
cfa.model <- "
aloof =~ q01 + q05 + q09 + q12 + q16 + q18 + q23 + q25 + q27 + q28 + q31 + q36
pragmatic =~ q02 + q04 + q07 + q10 + q11 + q14 + q17 + q20 + q21 + q29 + q32 + q34
rigid =~ q03 + q06 + q08 + q13 + q15 + q19 + q22 + q24 + q26 + q30 + q33 + q35"

cfa_fit <- cfa(cfa.model, data = bapq.all)
```

Table 4

Model Fit Metrics for CFA Model

Metric	Value
CFI	0.80
TLI	0.79
RMSEA	0.07

These model fit statistics do not meet the guidelines set up by Hu and Bentler (1999). To see if we can find a model that fits these data well, we next run an EFA, investigating the fit of models containing one, two, three, and four factors. Table 5 shows the model fit statistics from these EFA models.

```
# Obtaining item names
items <- bapq.all[,4:39]
var.names <- names(items)

# Assessing EFA from 1 to 4 factors
fit <- efa(data = bapq.all[,var.names], nfactors = 1:4)
```

Table 5

EFA Model Fit Statistics

Number of Factors	AIC	BIC	χ^2	*df*	*p*-value	CFI	RMSEA
1	592574.3	593051.3	27448.65	594	0	0.64	0.09
2	582739.1	583448.0	17543.47	559	0	0.77	0.07
3	575437.4	576371.4	10173.73	525	0	0.87	0.06
4	572586.0	573738.6	7256.29	492	0	0.91	0.05

These results indicate that a four factor model is the best for these data. Using this model, we assessed configural invariance. Table 6 shows the model fit statistics from this test.

```
# Four-Factor configural invariance model
conf <- "
f1 =~ q02 + q04 + q14 + q17 + q20 + q29 + q32
f2 =~ q03 + q06 + q08 + q13 + q15 + q19 + q22 + q24 + q26 + q30 + q33 + q35
f3 =~ q01 + q05 + q09 + q10 + q11 + q12 + q16 + q18 + q23 + q25 + q27 + q28 + q31 + q36
f4 =~ q07 + q21 + q34"

configural <- cfa(conf, data = bapq.all, std.lv = TRUE, group = "Parent")
```

Table 6

Model Fit Statistics From the Four-Factor Configural Invariance Model

Metric	Value
CFI	0.80
TLI	0.79
RMSEA	0.07

These model fit statistics also do not meet the guidelines set up by Hu and Bentler (1999). From this model, we iteratively pruned items with the lowest factor loadings, each time reassessing configural invariance. The best fitting model that increased the values of CFI and TLI while not drastically increasing RMSEA is a two-factor model. Table 7 shows the model fit statistics from this configural invariance model.

```
# Two-Factor configural invariance model
conf <- "
f2 =~ q03 + q08 + q13 + q19 + q22 + q24
f3 =~ q01 + q09 + q16 + q23 + q25 + q36"

configural  <- cfa (conf,  data =  bapq.all, std.lv = TRUE, group = "Parent")
```

Table 7

Model Fit Statistics from the Two-Factor Configural Invariance Model

Metric	Value
CFI	0.92
TLI	0.91
RMSEA	0.09

Note that, ideally, we would see CFI and TLI values above 0.95 and RMSEA below 0.05. However, we could not attain that fit using this modeling approach on these data. This is the best fitting configural model, therefore for demonstration purposes, we will continue with this factor structure to test for partial metric invariance.

Using this pruned two-factor model, we assessed partial metric invariance using the `partialInvariance()` function. We do this with both corrected (Benjamini-Hochberg) *p*-values and uncorrected *p*-values.

```
# Metric invariance model
weak <- "
f2 =~ q03 + q08 + q13 + q19 + q22 + q24
f3 =~ q01 + q09 + q16 + q23 + q25 + q36
f2 ~~ c(1, NA)*f2
f3 ~~ c(1, NA)*f3"
weak <- cfa(weak, data = bapq.all, group="Parent", group.equal="loadings"
models <- list(fit.configural = configural, fit.loadings = weak)

# Partial invariance models
pi_model <- partialInvariance(models, "metric")
pi_model.h <- partialInvariance(models, "metric", p.adjust = "hochberg")
```

Table 8 shows the *p*-values for each item split by the *p*-values were uncorrected or corrected and which method (Free, Fixed, or Wald) was used. The cells highlighted gray indicate an item that was identified as noninvariant. Note that there is not a great deal of difference between corrected and uncorrected *p*-values except for those calculated using the Wald method. In this table, within each testing method, at least half of the items are identified as metric noninvariant. However, these results should not be considered for any substantive interpretation because configural invariance was not reliably established.

Table 8

BAPQ Partial Metric Invariance p-Value by Item, Method, and p-Value Correction

Item		Uncorrected			Corrected		
Label	**Description**	**Free**	**Fixed**	**Wald**	**Free**	**Fixed**	**Wald**
q03	I am comfortable with unexpected changes in plans.	0.12	0.02	0.12	0.12	0.02	0.35
q08	I have to warm myself up to the idea of visiting an unfamiliar place.	0.06	0.02	0.12	0.06	0.02	0.35
q13	I feel a strong need for sameness from day to day.	0.03	0.07	0.00	0.03	0.07	0.00
q19	I look forward to trying new things.	0.06	0.02	0.10	0.06	0.02	0.35
q22	I have a hard time dealing with changes in my routine.	0.02	0.07	0.00	0.02	0.07	0.00
q24	I act very set in my ways.	0.01	0.00	0.43	0.01	0.00	0.43
q01	I like being around other people.	0.11	0.09	0.00	0.11	0.09	0.00
q09	I enjoy being in social situations.	0.00	0.08	0.00	0.00	0.08	0.00
q16	I look forward to situations where I can meet new people.	0.07	0.03	0.08	0.07	0.03	0.35
q23	I am good at making small talk.	0.00	0.04	0.01	0.00	0.04	0.09
q25	I feel like I am really connecting with other people.	0.05	0.01	0.28	0.05	0.01	0.43
q36	I enjoy chatting with people.	0.00	0.09	0.00	0.00	0.09	0.00

Funding: The authors have no funding to report.

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors have declared that no competing interests exist.

Supplementary Materials

For this article, the materials provided are additional graphs demonstrating that the results of the simulation do not differ for these revised network loadings (see Jamison, Christensen & Golino, 2024).

Index of Supplementary Materials

Jamison, L., Christensen, A. P., & Golino, H. F. (2024). Supplementary materials to *"Metric invariance in exploratory graph analysis via permutation testing"* [Graphs on network loadings]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.14700

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care, 44*(11), S176–S181. https://doi.org/10.1097/01.mlr.0000245143.08679.cc

Broderick, N., Wade, J. L., Meyer, J. P., Hull, M., & Reeve, R. E. (2015). Model invariance across genders of the broad autism phenotype questionnaire. *Journal of Autism and Developmental Disorders, 45*, 3133–3147. https://doi.org/10.1007/s10803-015-2472-z

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., & Lau, R. S. (2012). A direct comparison approach for testing measurement invariance. *Organizational Research Methods, 15*(2), 167–198. https://doi.org/10.1177/1094428111421987

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*(1), 1–27. https://doi.org/10.1177/014920639902500101

Chihara, L. M., & Hesterberg, T. C. (2022). *Mathematical statistics with resampling and R.* John Wiley & Sons.

Christensen, A. P., & Golino, H. (2021a). Estimating the stability of psychological dimensions via bootstrap exploratory graph analysis: A Monte Carlo simulation and tutorial. *Psych, 3*(3), 479–500. https://doi.org/10.3390/psych3030032

Christensen, A. P., & Golino, H. (2021b). On the equivalency of factor and network loadings. *Behavior Research Methods, 53*(4), 1563–1580. https://doi.org/10.3758/s13428-020-01500-6

Christensen, A. P., Golino, H., Abad, F. J., & Garrido, L. E. (2024). *Revised network loadings* [OSF project page containing simulation, analysis, and visualization data and R codes]. OSF. https://osf.io/dvwqs/

Comrey, A. L., & Lee, H. B. (2013). *A first course in factor analysis*. Psychology Press.

Danaher, P., Wang, P., & Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76*(2), 373–397. https://doi.org/10.1111/rssb.12033

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods, 23*(4), 617–634. https://doi.org/10.1037/met0000167

Feliciano, P., Daniels, A. M., Snyder, L. G., Beaumont, A., Camba, A., Esler, A., Gulsrud, A. G., Mason, A., Gutierrez, A., Nicholson, A., Paolicelli, A. M., McKenzie, A. P., Rachubinski, A. L., Stephens, A. N., Simon, A. R., Stedman, A., Shocklee, A. D., Swanson, A., Finucane, B.,... Chung, W. K. (2018). SPARK: A US cohort of 50,000 families to accelerate autism research. *Neuron, 97*(3), 488–493. https://doi.org/10.1016/j.neuron.2018.01.015

Finch, W. H., & French, B. F. (2008). Comparing factor loadings in exploratory factor analysis: A new randomization test. *Journal of Modern Applied Statistical Methods, 7*(2), Article 3. https://doi.org/10.22237/jmasm/1225512120

Fortunato, S. (2010). Community detection in graphs. *Physics Reports, 486*, 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

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

French, B. F., & Finch, W. H. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling: A Multidisciplinary Journal, 15*(1), 96–113. https://doi.org/10.1080/10705510701758349

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics, 9*(3), 432–441. https://doi.org/10.1093/biostatistics/kxm045

Golino, H., & Christensen, A. P. (2022). *EGAnet: Exploratory graph analysis – A framework for estimating the number of dimensions in multivariate data using network psychometrics* [Computer Software]. R Project for Statistical Computing. https://cran.r-project.org/web/packages/EGAnet/EGAnet.pdf

Golino, H., & Christensen, A. P. (2024). *EGAnet: Exploratory graph analysis – A framework for estimating the number of dimensions in multivariate data using network psychometrics* [Computer Software]. R Project for Statistical Computing. https://r-ega.net

Golino, H., Christensen, A. P., Moulder, R., Kim, S., & Boker, S. M. (2022). Modeling latent topics in social media using dynamic exploratory graph analysis:The case of the right-wing and left-wing trolls in the 2016 US elections. *Psychometrika, 9*, 156–187. https://doi.org/10.1007/s11336-021-09820-y

Golino, H., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PloS One, 12*(6), Article e0174035. https://doi.org/10.1371/journal.pone.0174035

Golino, H., Shi, D., Christensen, A. P., Garrido, L. E., Nieto, M. D., Sadana, R., Thiyagarajan, J. A., & Martinez-Molina, A. (2020). Investigating the performance of exploratory graph analysis and traditional techniques to identify the number of latent factors: A simulation and tutorial. *Psychological Methods, 25*(3), 292–320. https://doi.org/10.1037/met0000255

Hallquist, M. N., Wright, A. G., & Molenaar, P. C. (2021). Problems with centrality measures in psychopathology symptom networks: Why network psychometrics cannot escape psychometric theory. *Multivariate Behavioral Research, 56*(2), 199–223. https://doi.org/10.1080/00273171.2019.1640103

Haslbeck, J., & Bork, R. van. (2022). Estimating the number of factors in exploratory factor analysis via out-of-sample prediction errors. *Psychological Methods, 29*(1), 48–64. https://doi.org/10.1037/met0000528

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hurley, R. S., Losh, M., Parlier, M., Reznick, J. S., & Piven, J. (2007). The broad autism phenotype questionnaire. *Journal of Autism and Developmental Disorders, 37*(9), 1679–1690. https://doi.org/10.1007/s10803-006-0299-3

Jamison, L., Christensen, A. P., & Golino, H. F. (2024). *Metric invariance in exploratory graph analysis via permutation testing* [Code, data]. OSF. https://osf.io/4xyuc/?view_only=4ca1662a08ab4b3183217bbbc2f4e00c

Jiménez, M., Abad, F. J., Garcia-Garzon, E., Golino, H., Christensen, A. P., & Garrido, L. E. (2023). Dimensionality assessment in bifactor structures with multiple general factors: A network psychometrics approach. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000590>

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling, 16*(4), 642–657. https://doi.org/10.1080/10705510903206014

Jones, P. J., Mair, P., Simon, T., & Zeileis, A. (2020). Network trees: A method for recursively partitioning covariance structures. *Psychometrika, 85*(4), 926–945. https://doi.org/10.1007/s11336-020-09731-4

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

Jung, E., & Yoon, M. (2016). Comparisons of three empirical methods for partial factorial invariance: Forward, backward, and factor-ratio tests. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(4), 567–584. https://doi.org/10.1080/10705511.2015.1138092

Kaplan, D., & George, R. (1995). A study of the power associated with testing factor mean differences under violations of factorial invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 2*(2), 101–118. https://doi.org/10.1080/10705519509539999

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling, 18*(2), 212–228. https://doi.org/10.1080/10705511.2011.557337

Kjellström, S., & Golino, H. (2019). Mining concepts of health responsibility using text mining and exploratory graph analysis. *Scandinavian Journal of Occupational Therapy, 26*(6), 395–410. https://doi.org/10.1080/11038128.2018.1455896

Kuhn, M. (2022). *Caret: Classification and regression training* [Computer Software]. R Project for Statistical Computing. https://CRAN.R-project.org/package=caret

Ludbrook, J., & Dudley, H. (1998). Why permutation tests are superior to t and f tests in biomedical research. *American Statistician, 52*(2), 127–132. https://doi.org/10.2307/2685470

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective*. Routledge.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Millsap, R., & Olivera-Aguilar, M. (2012). Investigating measurement invariance using confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 380-392). Guilford Press.

Pons, P., & Latapy, M. (2006). Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications, 10*(2), 191–218. https://doi.org/10.7155/jgaa.00124

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. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raykov, T., Marcoulides, G. A., Harrison, M., & Zhang, M. (2020). On the dependability of a popular procedure for studying measurement invariance: A cause for concern? *Structural Equation Modeling: A Multidisciplinary Journal, 27*(4), 649–656. https://doi.org/10.1080/10705511.2019.1610409

Raykov, T., Marcoulides, G. A., & Millsap, R. E. (2013). Factorial invariance in multiple populations: A multiple testing procedure. *Educational and Psychological Measurement, 73*(4), 713–727. https://doi.org/10.1177/0013164412451978

Rensvold, R. B., & Cheung, G. W. (1998). Testing measurement models for factorial invariance: A systematic approach. *Educational and Psychological Measurement, 58*(6), 1017–1034. https://doi.org/10.1177/0013164498058006010

Rensvold, R. B., & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. In C. A. Schriesheim & L. L. Neider (Eds.), *Research in management* (Vol. 1, pp. 25–50). Information Age Publishing.

Rosseel, Y. (2012). Lavaan: An r package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rubin, M. (2024). Inconsistent multiple testing corrections: The fallacy of using family-based error rates to make inferences about individual hypotheses. *Methods in Psychology, 10*, Article 100140. https://doi.org/10.1016/j.metip.2024.100140

Schroeders, U., & Gnambs, T. (2020). Degrees of freedom in multigroup confirmatory factor analyses. *European Journal of Psychological Assessment, 36*(1), 105–113. https://doi.org/10.1027/1015-5759/a000500

Shi, D., Song, H., & Lewis, M. D. (2019). The impact of partial factorial invariance on cross-group comparisons. *Assessment, 26*(7), 1217–1233. https://doi.org/10.1177/1073191117711020

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292–1306. https://doi.org/10.1037/0021-9010.91.6.1292

Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology, 81*(2), 332–342. https://doi.org/10.1037/0022-3514.81.2.332

Van Borkulo, C. D., Bork, R. van, Boschloo, L., Kossakowski, J. J., Tio, P., Schoevers, R. A., Borsboom, D., & Waldorp, L. J. (2022). Comparing network structures on three aspects: A permutation test. *Psychological Methods, 28*(6), 1273–1285. https://doi.org/10.1037/met0000476

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

Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *The Journal of Experimental Education, 80*(1), 26–44. https://doi.org/10.1080/00220973.2010.531299

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). https://doi.org/10.1037/10222-009

Williams, D. R., Rast, P., Pericchi, L. R., & Mulder, J. (2020). Comparing gaussian graphical models with the posterior predictive distribution and Bayesian model selection. *Psychological Methods, 25*(5), 653–672. https://doi.org/10.1037/met0000254

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 435–463. https://doi.org/10.1080/10705510701301677

Zhang, M., & Yang, L. (2022). Detecting measurement noninvariance with continuous indicators using three different statistical methods under the framework of latent variable modeling.

Structural Equation Modeling: A Multidisciplinary Journal, 29(4), 550–568.
https://doi.org/10.1080/10705511.2021.2021533