Metric Invariance in Exploratory Graph Analysis via Permutation Testing

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

[1] University of Virginia

[2] Vanderbilt University

Author Note

Correspondence concerning this article should be addressed to Laura Jamison.

E-mail: lj5yn@virginia.edu

Abstract

Establishing measurement invariance (MI) is vital to ensure applicability and comparability across groups (or time points) in psychological measurement. If MI is violated, differences between groups could be due to measurement rather than true differences between groups. Factor analytic methods are commonly used to test MI; however, many existing methods have reduced power to detect MI due to model misspecification (e.g., noninvariant referent indicators, reliance on data-driven methods). Literature reviews on MI studies have reported inaccurate or inadequately described models with modeling errors primarily predicted by software choice. Another reduction in power may be due to goodness of fit measures when group sample sizes vary. Network psychometrics methods to test MI are limited and primarily focus on partial correlation differences. In the present research, we propose a novel network psychometrics method to test MI within the Exploratory Graph Analysis framework. This method leverages so-called network loadings by calculating their differences between groups and uses permutation testing to stasitically compare these differences to the permutated null distribution. A simulation study was conducted using data structures common in psychological research (factor models) that included unequal group sample sizes. The proposed network psychometrics method demonstrated comparable ability to factor analytic methods in detecting MI, with some improvement in certain conditions such as lower noninvariance effect sizes in smaller or unequal sample sizes.

Metric Invariance in Exploratory Graph Analysis via Permutation Testing

Introduction

Measurement invariance assesses the equivalence of a measure across groups at a single time point (cross-sectional) or across time (longitudinal). Equivalent measurement indicates that a measure has the same meaning in each group and is therefore measuring the same construct in the same way across groups. Demonstrating measurement invariance is vital for the application of any psychological measurement. Tests for measurement invariance usually occur prior to administration across qualitatively distinct groups or time points to ensure reliable and valid measurement (Vandenberg & Lance, 2000). Measurement between groups where measurement invariance does not hold cannot be credibly interpreted across cultures, languages, sociodemographic categories, or administration modes (Borsboom, 2006). When measurement invariance is violated, overall score differences between groups may be the result of the measure itself rather than true differences between groups (F. F. Chen, 2007).

Traditionally, measurement invariance is tested using latent variable methods such as Item Response Theory (IRT) or factor analysis (FA) (Stark et al., 2006). In applied psychological research, FA is more common (Putnick & Bornstein, 2016) and therefore serves as the conceptual framework for the paper. Within FA, four consecutive tests are used to establish measurement invariance: configural invariance (equivalence of factor structure), metric invariance (equivalence of factor loadings), scalar invariance (equivalence of item intercepts), and strict invariance (equivalence of item residuals) (Widaman & Reise, 1997). Full measurement invariance is only achieved when a measure is determined to be invariant on all of these tests.

The present research investigated a novel network psychometrics method to test metric invariance within the the Exploratory Graph Analysis framework (H. F. Golino & Epskamp, 2017). Before introducing this method, a brief overview of FA measurement

invariance and some issues associated them (focusing on metric invariance) is provided. After, Exploratory Graph Analysis and the proposed method are introduced and formulated, respectively. Finally, a simulation study that compares the proposed method against traditional FA measurement is conducted.

Measurement Invariance in Traditional Psychometrics

Factorial Invariance

In the FA framework, measurement invariance is conducted by comparing a constrained model to a less constrained model from a weaker level of invariance (e.g., comparing the more constrained factor loading equivalence model to the less constrained factor structure equivalence model). The constrained model is formed by setting relevant parameters across groups (e.g., loadings) to be equal and comparing model fit to the unconstrained model (i.e., the same parameters are freely estimated). A model comparison is conducted and change in model fit statistics (e.g., ΔCFI, $\Delta\chi^2$) is used to determine invariance. If the constrained model fits about as well as the less constrained model (e.g., $\Delta\text{CFI} \leq 0.02$), then invariance at that level is considered to be established. This process is staged starting from the least constrained model (i.e., configural invariance) to the most constrained model (strict invariance). Each level is only tested if the previous level demonstrates invariance.

Configural invariance (equivalence of factor structure) is established by assessing the fit of a Multi-Group Confirmatory Factor Analysis model on all groups. The common factor model is defined as,

$$X_{jk} = \tau_{jk} + \sum_{m=1}^{r} \lambda_{jmk} W_m + U_j, \tag{1}$$

where X_{jk} is the jth random variable in the kth population. This value is defined by a linear function where τ_{jk} is the latent intercept for the jth variable in the kth population, λ_{jmk} is

the factor pattern parameters for the jth variable corresponding to the r common factors ($m = 1, \ldots, r$) in the kth population, W_m are the common factor scores for the r factors, and U_j is the unique factor score for the jth measured variable. When $\mathbf{W}' = (W_1, W_2, \ldots, W_r)$ and $\mathbf{U}' = (U_1, U_2, \ldots, U_p)$ (where p is the number of measured variables) it is assumed that $E_k(\mathbf{U}) = 0$ and uncorrelated with $\mathbf{W}$.

From (1), the unconditional mean ($\boldsymbol{\mu}_{\mathbf{X}_k}$) and covariance structure ($\boldsymbol{\Sigma}_{\mathbf{X}_k}$) for the measured variables $\mathbf{X}$ can be expressed as,

$$E_k(\mathbf{X}) = \boldsymbol{\mu}_{\mathbf{X}_k} = \boldsymbol{\tau}_k + \boldsymbol{\Lambda}_k \boldsymbol{\kappa}_k$$

and

$$\text{Cov}_k(\mathbf{X}) = \boldsymbol{\Sigma}_{\mathbf{X}_k} = \boldsymbol{\Lambda}_k \boldsymbol{\Phi} \boldsymbol{\Lambda}_k' + \boldsymbol{\theta}_k,$$

where $\mathbf{X}_k$ represents the measured variables, $\boldsymbol{\Lambda}_k$ represents the factor loading matrix, $\boldsymbol{\kappa}$ represents the common factor scores for r common factors, $\boldsymbol{\Phi}_k = \text{Cov}_k(\mathbf{W})$, and $\boldsymbol{\theta}_k = \text{Cov}_k(\mathbf{U})$ for the kth population. Following Millsap (2011), configural invariance can then be defined as,

$$\boldsymbol{\mu}_{\mathbf{X}_k} = \boldsymbol{\tau}_k + \boldsymbol{\Lambda}_{kc} \boldsymbol{\kappa}_k \tag{2}$$

and

$$\boldsymbol{\Sigma}_{\mathbf{X}_k} = \boldsymbol{\Lambda}_{kc} \boldsymbol{\Phi}_k \boldsymbol{\Lambda}'_{kc} + \boldsymbol{\Theta}_k, \tag{3}$$

for groups $k = 1, \ldots, K$ where $\boldsymbol{\Lambda}_{kc}$ denotes that the pattern matrices have the same structure with configural invariance.

The pattern matrices imply that each population has the same number of factors with the same distribution of variables. If this model fits satisfactorily on all groups, then the organization of items into these constructs is appropriate for all groups (Putnick & Bornstein, 2016). In other words, configural invariance establishes that the pattern of zero and nonzero loadings (fixed and free loadings) exists in all groups (Widaman & Reise, 1997). This pattern only demonstrates that similar, but not equivalent, latent factors exist in all groups. To perform comparisons across groups, model parameters must also be established as invariant.

Metric invariance implies, for groups $k = 1, \ldots, K$,

$$\boldsymbol{\mu}_{\mathbf{X}_k} = \boldsymbol{\tau}_k + \boldsymbol{\Lambda}\boldsymbol{\kappa}_k \tag{4}$$

and

$$\boldsymbol{\Sigma}_{\mathbf{X}_k} = \boldsymbol{\Lambda}_k \boldsymbol{\Phi}_k \boldsymbol{\Lambda}_k' + \boldsymbol{\Theta}_k, \tag{5}$$

which implies loadings are equivalent across groups. This model is compared against the configural model, and if demonstrated to have similar 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 (the model fits significantly worse), then there can be no comparison of factor variances and covariances (and subsequently scaled correlations) across groups (Widaman & Reise, 1997). Without the foundation of metric invariance, further testing for scalar and strict invariance should not be conducted. Testing for partial invariance of loadings, however, is often appropriate.

Partial Invariance

Partial invariance was introduced because of the difficulties establishing full measurement invariance in practice (Byrne et al., 1989). Partial invariance is when only a portion of the parameter set demonstrates noninvariance. For metric invariance, the goal was

to determine how many noninvariant item loadings existed per latent factor. Opinions vary concerning what type of partial invariance is acceptable (e.g., partial scalar invariance could still impact the mean of the latent factor), as well as what extent of partial invariance is permissible (i.e., the proportion of noninvariant items) (Putnick & Bornstein, 2016). Nonetheless, testing for partial invariance is useful to identify specific item parameters that are noninvariant. In the case of metric invariance, individual invariance constraints can be selectively introduced to $\mathbf{\Lambda}$ and tested.

Arguably, identifying partial invariance provides more useful information than an omnibus test. 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. Partial invariance testing could also identify noninvariance not identified by an omnibus test. This difference is a well documented with omnibus tests (Raykov et al., 2013), and the potential effect of misidentifying items as invariant is quite concerning. 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. In some cases, failing to identify a truly noninvariant item (Type II error) could have worse consequences for measurement than incorrectly classifying a truly invariant item (Type I error).

Problems With Traditional Testing

Some partial invariance tests use a referent indicator or an item whose relevant parameter (e.g., loading) is set to be equal across groups in the unconstrained model. Employing referent indicators assumes that the chosen indicator is itself invariant (often without testing for whether it is invariant). This strategy relies heavily on proper implementation by the researcher, leading to potential problems if carried out without care. Therefore, we focus on three partial invariance methods that do not require a referent

indicator: factor-ratio test (Rensvold & Cheung, 1998), a data-driven method that applies modification indices in a sequential manner proposed by Yoon and Millsap (2007), and a method proposed by Raykov et al. (2013) using a multiple testing procedure.

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. Since this method investigates all possible combinations of referent indicators, it is computationally expensive as the number of variables increases.

Yoon and Millsap (2007) proposed a data-driven method to sequentially evaluate modification indices. A modification index is the change in model fit (based on a likelihood ratio test) after a particular parameter constraint was freed. For two groups 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. Then, for both groups, all factor loadings are constrained to be equal. Based on fit indices, if this model fits as well as the configural model, then full invariance is established. If not, then each individual constrained parameter is freed and the change in χ^2 from a likelihood ratio test is estimated. This process is repeated until adequate model fit is established. Their simulation study found that this method controls false positives very well but primarily in "ideal" data conditions (large sample sizes, greater difference in loadings, low cross loadings). One limitation of this approach is that model misspecifications can lead an artificial inflation of Type I errors (Kim & Yoon, 2011; Whittaker, 2012), especially as model modifications are made throughout the testing process (Yoon & Millsap, 2007).

The third method uses the Benjamini-Hochberg procedure (BH-procedure; Benjamini

& Hochberg, 1995), a multiple comparison method that controls the Type I error, to compare two models using χ^2 testing (Raykov et al., 2013). One model (the baseline model) is a fully constrained model. In the other model, a parameter (e.g., loading) is freed for one item across groups. The two models are compared and this process is repeated for each item. At each level of invariance testing, 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 does circumvent the choice of referent indicator, using a fully constrained baseline model (i.e., including any model with constrained noninvariant items) could negatively impact accuracy (Benjamini & Hochberg, 1995). Given the computationally intensive nature of the factor-ratio test and model (mis)specification issues with the data-driven approach proposed by Yoon and Millsap (2007), we chose to focus on Raykov et al.'s (Raykov et al., 2013) method in this paper.

Network Psychometrics

The primary goal of the current work is to provide a network psychometrics method to test for metric invariance. Network psychometric methods have become a popular alternative to latent variable modeling. A network psychometric model represents a measure (e.g., self-report inventory) as a network where nodes (circles) represent variables (e.g., items) and edges (lines) represent the associations between them. There are many existing methods to compare psychometric network models across groups such as using a grouping variable as a moderator (Haslbeck & Bork, 2022), comparing the total sum of connections in the networks (Network Comparison Test) (Van Borkulo et al., 2022), recursive partitioning of covariance structures (Jones et al., 2020), Fused Graphical Lasso (Danaher et al., 2014), and a Bayesian method (Williams et al., 2020). Although these methods are robust and widely applicable, they compare individual edges or the entire network rather than sub-structures

that may exist within them (e.g., dimensions). The Exploratory Graph Analysis framework was introduced as a way to investigate dimensions in psychometric networks.

Exploratory Graph Analysis (EGA)

Exploratory Graph Analysis (EGA) first estimates a network and then applies a community detection algorithm to identify communities or dimensions in the network (H. Golino et al., 2020; H. F. Golino & Epskamp, 2017). A common approach to estimate a network in psychology is to apply the graphical least absolute shrinkage and selection operator (*glasso*) (Friedman et al., 2008). The *glasso* estimates a Gaussian Graphical Model (Lauritzen, 1996) where edges represent the partial correlations between two variables conditioned on all other variables. The glasso is an extension of the least absolute shrinkage and selection operator (Tibshirani, 1996) regularization approach to the covariance matrix. Through regularization, partial correlations in the network shrink toward zero with some becoming zero. In the *glasso* algorithm, there is a parameter called *lambda* that determines the strength of the shrinkage (i.e., tendency for partial correlations to shrink to zero). In the network psychometrics literature, the extended Bayesian information criterion (EBIC) (J. Chen & Chen, 2008) is used to select the lambda parameter based on model fit to the data (Epskamp & Fried, 2018).

After the network is estimated, a community detection algorithm is be applied. Community detection algorithms identify *communities* or sets of nodes that tend to be more densely connected to one another than the rest of the nodes in the network. In their seminal paper, H. F. Golino and Epskamp (2017) demonstrated that communities in psychometric networks are statistically consistent with latent factors when data are generated from a factor model. The Walktrap algorithm (Pons & Latapy, 2006) is a commonly applied community detection algorithm in the network psychometrics literature (Christensen et al., 2023). The Walktrap algorithm applies a random walk process over the nodes and edges of the network, using the edge weights (i.e., partial correlations) to inform that probability of

"walking" from one node to another. Ward's hierarchical clustering algorithm (Ward, 1963) is applied the transition matrix and *modularity*, or the extent the network is partitioned into communities where nodes are more strongly connected to nodes in their community than other nodes (Newman, 2006), is used to select the community partition (i.e., which nodes belong to which communities). Figure 1 depicts the end result of the EGA process on the `bfi` or 25-item Big Five Inventory dataset in the {psych} package (version 2.3.6) (Revelle, 2017) in `R` (version 4.1.0) (R Core Team, 2022).

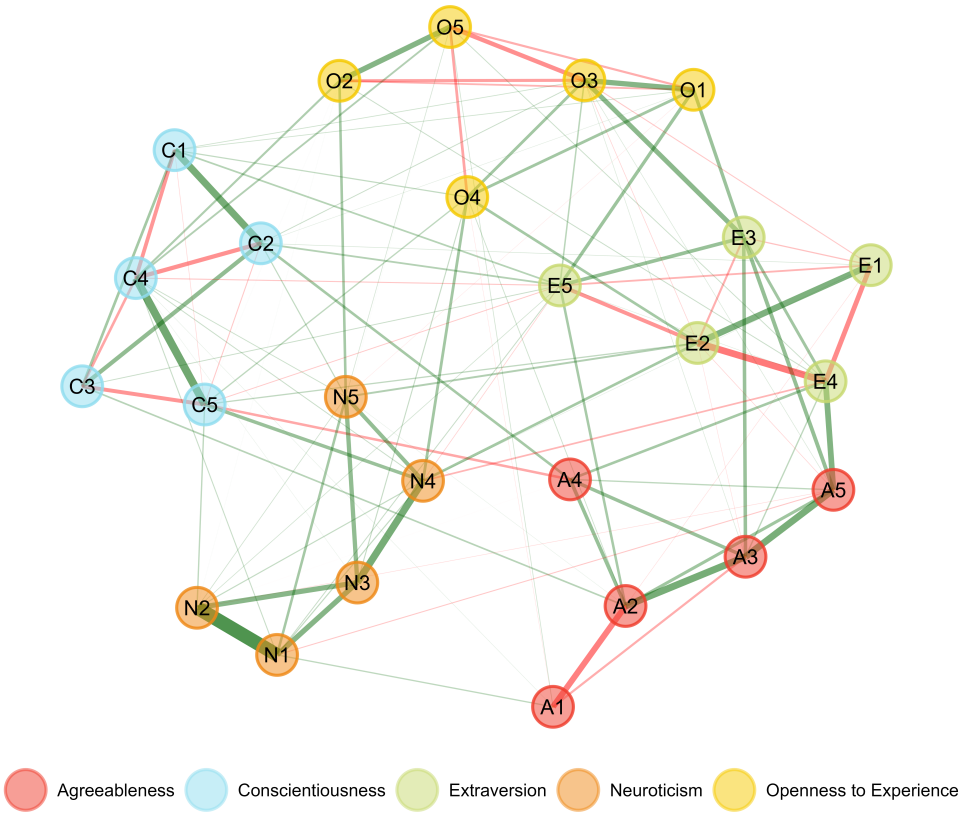

Figure 1

EGA network plot of the `bfi` *dataset in the {psych} package. Circles represent items in the BFI dataset and their colors correspond to their community. Lines represent edges with green indicating a positive partial correlation and red indicating a negative partial correlation. Line thickness corresponds to the magnitude of the partial correlation between two items.*

Beyond community detection, there are other measures to quantify the structure of networks. The majority of the network psychometric literature has focused on the local

structure of networks using *centrality* measures. Centrality measures quantify the relative position of a node in a network. To date, *node strength* or the (absolute) sum of a node's connections has been the most common measure applied (Bringmann et al., 2019). Recent work by Hallquist et al. (2021) showed that node strength is comprised of both dominant and cross loadings (from a FA perspective). They argue that unless dimensional structure is accounted for the metrics used to quantify networks, such as node strength, will be biased due to latent confounding. This finding prompted Christensen and Golino (2021b) to investigate whether splitting a node strength based on the communities identified by EGA could resolve the latent confounding issue. Through a series of simulations, they showed that this issue can be mitigated using this strategy. They coined the standardized version of this measure *network loadings* because they were statistically consistent with the factor loadings generated from the simulated data. These network loadings open the door for more traditional psychometric applications in network psychometrics including measurement invariance.

Measurement Invariance using Network Psychometrics

In this section, a method to test configural invariance is established and a novel method to test metric invariance is proposed. These methods do not extend to scalar and strict invariance because no latent variables are estimated in the network models and therefore there are no item means or residuals to compare. The goal of these methods is to establish configural and metric invariance in network psychometrics that are statistically consistent with traditional psychometric approaches.

Configural Invariance in the EGA Framework

Before introducing the proposed method to test metric invariance, a method to test configural invariance must be established. Configural invariance in the EGA framework exists when the same nodes have been identified in the same communities for all groups. This configuration can be 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 found by random sampling variability. In other words, additional testing should be conducted to test whether items are consistently organized into the same communities and whether the number of communities and their structure fluctuates.

A recent approach called Bootstrap EGA (Christensen & Golino, 2021a) produces a sampling distribution of EGA results that can then used to evaluate the stability of the identified structure. Bootstrap EGA uses *structural consistency* to assess the stability of dimensions and items. Structural consistency refers to the proportion of bootstraps in which the same structure as the initial EGA was recovered. If the groups are pooled together into one sample, higher structural consistency indicates it is more likely for this structure to be representative of the population structure for all groups. Lower structural consistency, or if it is found that the structure varies along with the sample size or the specific samples drawn, indicates that configural noninvariance may be present. Structural consistency can then be further broken down to assess the stability of dimensions (number of dimensions identified) and the stability of items (proportion of bootstraps in which an item was assigned to the same dimension). Items showing a stability of < 0.70 are considered to be less stable (Christensen & Golino, 2021a), indicating that these items may not be reproducing in the same community for all groups.

To test for configural invariance in the EGA framework, we recommend a straightforward approach by conducting bootstrap EGA on the entire sample. If the goal is to achieve invariance, then items with < 0.70 stability can be removed. Next, without these items, the structures identified by EGA across groups can be evaluated. This process should be repeated until a consistent common structure can be identified across all groups within a sample. It should be noted that this approach allows for no amount of partial configural invariance to be present in order to test for metric invariance.

Metric Invariance in the EGA Framework

Once configural invariance is established, then 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 considerably fewer assumptions about the population distribution than traditional hypothesis testing procedures (Chihara & Hesterberg, 2022). The concept of permutation testing can be applied to any test statistic, providing flexibility to easily adapt the model to any hypothesis or statistic (Chihara & Hesterberg, 2022; Ludbrook & Dudley, 1998). First, we define network loadings, then we discuss the permutation procedure.

Network loadings are mathematically defined as follows. Let $\mathbf{\Psi}$ represent a symmetric network $p \times p$ made up of edge weights (e.g., partial correlations) where p is the number of variables. Node strength is then defined as,

$$S_i = \sum_{j=1}^{n} |\psi_{ij}|,$$

where $|\psi_{ij}|$ is the absolute weight between node i and j. Then S_i is the sum of the absolute weights between node i and all other n nodes—that is, the strength of node i. Node strength can then be split between the communities identified by EGA,

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

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 communities. This measure can be standardized using,

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

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, $\boldsymbol{\aleph}$, are absolute weights with the signs being added after the loadings are computed. Unlike factor analysis, the number and content of communities is extracted from the network's structure before computing network loadings. Due to sparsity induced by network estimation methods, it is possible for a node to have a network loading of zero because it has no connections to one or more other communities.

To test the equivalence of network loadings across groups, we propose applying a permutation test. We start with two groups and then discuss how it can be extended to more than two groups. The original $n \times p$ data, D (where n is sample size and p 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. EGA is performed separately using 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). With identical community structures, the corresponding $p \times C$ network loading matrices $\boldsymbol{\aleph}_1$ and $\boldsymbol{\aleph}_2$ are computed where C is the number of communities. The difference between the two matrices is then computed by,

$$\boldsymbol{\mathcal{T}} = \boldsymbol{\aleph}_2 - \boldsymbol{\aleph}_2,$$

to form an $p \times C$ matrix $\boldsymbol{\mathcal{T}}$ which contains test statistics for each item. Only the elements of $\boldsymbol{\mathcal{T}}$ that correspond to each variable's assigned community are retained and the rest of the values are set to zero (in factor analytic terms, the dominant loadings are retained and cross-loadings are set to zero).

To form a null distribution for each test statistic to be compared to, grouping variable

G is randomly reordered and becomes G_R. Then the original data D is split by G_R to form two new datasets, D_{R1} and D_{R2}, thereby removing any relationship between item responses and group membership. This procedure is done repeatedly an i number of times creating i new datasets D_{R1i} and D_{R2i}. EGA is performed on each permuted dataset D_{R1i} and D_{R2i}. Then, network loadings are computed and the difference between the network loadings for each item is taken using,

$$t_i = \boldsymbol{\aleph}_{R1i} - \boldsymbol{\aleph}_{R2i}.$$

These differences are put in ascending order, $t_{(1)} \leq \cdots \leq t_{(i)}$, forming a null distribution representing the difference in network loadings if there was no 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 are calculated as,

$$g = \begin{cases} 1 & \text{if } t_i \geq \boldsymbol{\mathcal{T}} \\ 0 & \text{otherwise,} \end{cases}$$

$$l = \begin{cases} 1 & \text{if } t_i \leq \boldsymbol{\mathcal{T}} \\ 0 & \text{otherwise,} \end{cases}$$

$$G = \frac{(\sum_{i=1}^{N} g_i) + 1}{(N+1)},$$

$$L = \frac{(\sum_{i=1}^{N} l_i) + 1}{(N+1)},$$

$$p = \begin{cases} 2G & \text{if } G \leq L \\ 2L & \text{otherwise.} \end{cases}$$

If any p-value is determined to be ≤ 0.05, it can be asserted that metric invariance was violated, although if not all p-values are ≤ 0.05, then partial metric invariance has been found. As previously mentioned, there is no agreement in the literature, to our knowledge, as to what constitutes an acceptable level of partial invariance.

Conveniently, this approach is 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.

Simulation Study

The following section outlines the methods used for each portion of the simulation study. We compare the proposed method against the FA model with Raykov et al.'s (Raykov et al., 2013) method. After introducing the FA approach, we discuss the multiple comparison procedure (aforementioned BH-procedure) and how it is applied in the current study. The data generation methodology, data conditions tested, and metrics used to assess model

accuracy are then outlined in detail. For purposes of comparison, we focus on continuous data. Although factor and network models can generalize to ordinal variables using polychoric correlations, our intent was to evaluate the proposed method under ideal conditions before submitting it to more extensive conditions.

Factor Analytic Testing Approach

To test metric and partial metric invariance using FA models, we estimated two models: a configural (unconstrained model, see (2) and (3)) and a model with loadings constrained to be equal across k populations (constrained metric model, see (4) and (5)). In order to directly compare the FA approach to the method proposed for EGA, 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) in `R`, the Fixed and Wald methods are run simultaneously with Free. Because this procedure is commonly used in practice, we evaluated the results of all three methods.

In all methods, an original model was chosen to be either the constrained or unconstrained model. Then, loadings were iteratively either fixed or freed 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 any variables by fixing the loading of one variable per factor to 1. In this way, we could make direct comparisons across the EGA and FA approaches. Since we are only interested in whether an item was identified as noninvariant, we do not report fit statistics (e.g., CFI) from these models. Instead, we summarize the effectiveness of each methods ability to classify items correctly as invariant or noninvariant.

The Free method uses the constrained model as the original model. Iteratively, each variable j is freed in the matrix Λ to create 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 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 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 J tests. The Wald method is similar to Free. It uses the constrained model as the original model, but rather than iteratively freeing each jth variable and conducting likelihood ratio tests it uses a multivariate Wald test. In each method, multiple hypotheses are being tested. It should be noted that these methods do not adjust for Type I error. Therefore, a multiple comparison test could be applied.

Multiple Comparison Problem

Within both frameworks, multiple hypotheses are tested which can artificially inflate the Type I error and require a multiple comparison procedure (MCP) to be applied (Raykov et al., 2013; Steinberg, 2001). To select which MCP to apply, we first need to define which type of error we are most interested in controlling. A Type I error in this context is to identify a truly invariant item as noninvariant; a Type II error is to identify a truly noninvariant item as invariant. The practical risk of classifying a truly invariant item as noninvariant is relatively low because it could still leave a set of adequate and applicable invariant items. In contrast, falsely classifying a truly noninvariant item as invariant could have serious consequences for measurement. Therefore, a conservative approach would favor greater accuracy the identification of noninvariant over invariant items. Based on this premise, greater emphasis was placed on correctly identifying noninvariant items and controlling the Type II error.

Most MCPs focus on controlling the Family Wise Error Rate (FWER) or the probability of making a Type I error at all, which is important to control if there could be serious implications to falsely rejecting a null hypothesis. In the instance of partial invariance, however, the opposite is true. The adverse impact of falsely identifying an item as noninvariant is greater than falsely identifying an item as invariant, particularly if the

construct will be used to compare across groups (Shi et al., 2019). In this vein, we propose the investigation of False Discovery Rate (FDR). FDR represents the expected number of falsely rejected null hypotheses if any null hypotheses are rejected. Formally FDR, ϕ, is defined as,

$$\phi = \begin{cases} E(V|R), & \text{if } R > 0 \\ 0, & \text{otherwise,} \end{cases}$$

where V is the number of falsely rejected null hypotheses, and R is the total number of rejected null hypotheses out of the set of all hypotheses tested. The BH-procedure, which was designed to control the FDR, maintains a FWER at $\alpha = .05$ as well as demonstrates marked improvements in power above and beyond traditional MCP methods (e.g., Tukey, Bonferroni, Scheffe). Raykov et al. (2013) first proposed the use of the BH-procedure because it allows for a more accurate depiction of which parameters are truly noninvariant when noninvariance is found rather than focusing on lowering the risk of mistakenly identifying noninvariant parameters at all.

Data Generation

Data were generated from a factor model following Golino and colleagues' (2020) approach. First, a population correlation matrix for each group, $\mathbf{R}_{R_G}$, with communalities on the diagonal is defined as,

$$\mathbf{R}_{R_G} = \mathbf{\Lambda}_G \mathbf{\Phi} \mathbf{\Lambda}_G',$$

where $\mathbf{R}_{R_G}$ is the reproduced population correlation matrix for each group G, $\mathbf{\Lambda}_G$ is a $p \times r$ factor loading matrix for p variables and r factors for each group G, and $\mathbf{\Phi}$ is the structure matrix of the latent variables (i.e., a $r \times r$ matrix of correlations among factors). This structure means the population does not contain any correlated residuals and thereby no

minor factors.

Then, by inserting unities on the diagonal of $\mathbf{R}_{R_G}$ it becomes a full rank matrix and is now population correlation matrix $\mathbf{R}_{P_G}$. Each group in G is assigned a $\mathbf{R}_{P_G}$ matrix. A Cholesky decomposition is performed on the population correlation matrix for each $\mathbf{R}_{P_G}$ to obtain new random values that maintain the original correlations, resulting in $\boldsymbol{\Upsilon}_G$ where,

$$\mathbf{R}_{\mathbf{P_G}} = \boldsymbol{\Upsilon}'_G \boldsymbol{\Upsilon}_G.$$

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

$$\mathbf{X}_G = \mathbf{Z}_G \boldsymbol{\Upsilon}_G,$$

where $\mathbf{Z}_G$ is an $n \times p$ matrix where each value is a random draw from a standard normal distribution and all variables are uncorrelated. The result is a dataset of continuous variables for each group.

Design

The overall design of the simulation study closely followed Kim and Yoon (2011) with a few modifications. A two factor model was simulated with 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 design allowed us to manipulate interfactor correlation and investigate any impact on the power of the proposed method. Only one variable in one factor was simulated to have different loadings across group. Because our main goal was to assess each method's ability to identify noninvariant items correctly, having only one noninvariant

item allowed an "all-or-nothing" hit rate. Further, comparisons between methods to detect invariant items within factors, both with and without noninvariant items, could be made.

For simplicity, only two groups were simulated but we note that each method's expansion into testing three or more groups is tractable. Factor loadings were set to be the same across factors: Factor 1 = (0.80, 0.70, 0.60); Factor 2 = (0.80, 0.70, 0.60). Keeping large, static factor loadings allowed us to make sure configural invariance was not negatively impacted, particularly for data conditions with a large difference in loadings and/or a large interfactor correlation. Similar to H. F. Golino and Epskamp (2017), the correlation between factors was set to be small (0.30), moderate (0.50), or large (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). Because we used static factor loadings, the magnitude of loading differences 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 disparate sample sizes per group (500 in G_1 and 1000 in G_2). These sample sizes allowed us to compare the proposed method's ability to detect noninvariant items in conditions that traditional methods typically struggle (i.e., small or disparate samples). In total, there were 18 separate conditions. For each condition, 500 datasets were simulated.

Measurement invariance was tested on each simulated dataset using EGA in the {EGAnet} package (version 1.1.1) (H. Golino & Christensen, 2022), FA in the {lavaan} package (version 0.6.15) (Rosseel, 2012), and {semTools} for FA measurement invariance. All analyses were conducted in `R` and full code can be found at https://osf.io/2ewyn/?view_only=7147ed15f51c45cda17628e666a91bf2.

Data Analysis

Confusion matrix metrics were used to assess the accuracy of each model's noninvariance detection. 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, there is only one possible true positive (TP) (or false negative; FN) based on whether the model (in)correctly identifies noninvariance in Variable 5. True negatives (TN) occur when all other variables are identified as invariant; false positives (FP) occur when all other variables are identified as noninvariant.

Four separate metrics assessing the proportions of TP, TN, FP, and FN are employed to emphasize different aspects of model accuracy: *Hit Rate*, *Specificity*, *Sensitivity*, and *F1*. The {caret} package (version 6.0.94) (Kuhn, 2022) in R was used to calculate *Specificity*, *Sensitivity*, and *F1*. All metrics were calculated separately using both MCP corrected and uncorrected p-values.

Hit Rate provides a straight forward, overall assessment of method accuracy for correctly identifying invariance or non-invariance. If a variable was correctly identified as a TP or TN, then it was assigned a 1 for *Hit Rate*; otherwise, it was assigned a 0. *Hit Rate* equals the mean number of correct identifications for each variable and condition. *Specificity* ($\frac{TN}{TN+FP}$) represents the proportion of correctly identified invariant items for each method. *Sensitivity* ($\frac{TP}{TP+FN}$) represents the mean of correctly identifying the noninvariant item for each method (i.e., Variable 5). *F1* is calculated using *Sensitivity* and *Precision*. *Precision* assesses the proportion of TP's out of all rejected null hypotheses ($\frac{TP}{TP+FP}$) or accuracy of identifying the truly noninvariant item at the cost of overidentifying other items as noninvariant. *F1* is the harmonic mean of these two metrics: $\frac{2(Precision \times Sensitivity)}{(Precision+Sensitivity)}$. The *F1* metric is penalized when *Precision* and *Sensitivity* are unequal. For example, if $Precision = .50$ and $Sensitivity = .50$, then $F1 = .50$. If they both increased to 0.80, then *F1* would increase to 0.80 as well. If, however, only one metric increased to 0.80, then *F1* would be 0.62.

To summarize, *Hit Rate* provides a metric for accuracy that can assessed for each

variable across conditions; *Sensitivity* and *F1* assess the ability of each method to accurately identify noninvariant items; *Specificity* assesses the ability of each method to accurately identify invariant items. For all metrics, values range from 0 to 1 with values going to 1 indicating better performance. Although high *Specificity* and *Sensitivity* are desirable, *Sensitivity* should be given greater weight in our study. *Sensitivity* represents the hit rate of detecting the noninvariant variable: it must be either a *TP* or *FP* and therefore *Sensitivity* boils down to the average accuracy for identifying the noninvariant variable in each condition. *Specificity* represents the extent to which all invariant variables are correctly identified as invariant. In most applied cases, there is a preference to avoid the conclusion that a measure is invariant when it is not; therefore, *Sensitivity* is the key metric in our simulation study.

Results

Effect of MCP on *p*-Values

Configural invariance was recovered in 99.73% of the simulated datasets using EGA and 100% using FA. To provide a direct, full comparison, the iterations that did not reach configural invariance for EGA were removed. Within method, we assessed accuracy in terms of metric invariance. Figures 2 and 3 show the mean and 95% confidence interval for the p-values (both MCP corrected and uncorrected) of each variable split by method, sample size, correlation between factors, and loading difference. A dashed line intercepts the y-axis at .05 representing the α level.

Across methods, MCP corrected and uncorrected p-values were lowest in Variable 5, regardless of condition. This result demonstrates that the manipulation was effective. Looking at Figure 2 in the Free column, the 6 variables in Factor 1 had lower p-values than the 6 variables in Factor 2. This pattern was not present for the other methods. Comparing Figure 3 (MCP corrected p-values) to Figure 2 (uncorrected p-values), we see that the average p-value is higher for MCP corrected p-values than uncorrected, regardless of whether an item is invariant. Looking at the MCP corrected p-values (Figure 3), all 3 FA methods

have a higher average *p*-value for Variable 5 when the difference in loading for Variable 5 is set to 0.20 and sample size is either 500 or disparate. Under these same conditions, this same trend in EGA is only noticeable when the correlation between factors increases to 0.70. In other words, having a lower effect size with a smaller or disparate sample size affects the uncorrected *p*-value for all 3 FA methods more so than EGA except when there is a higher correlation between factors.

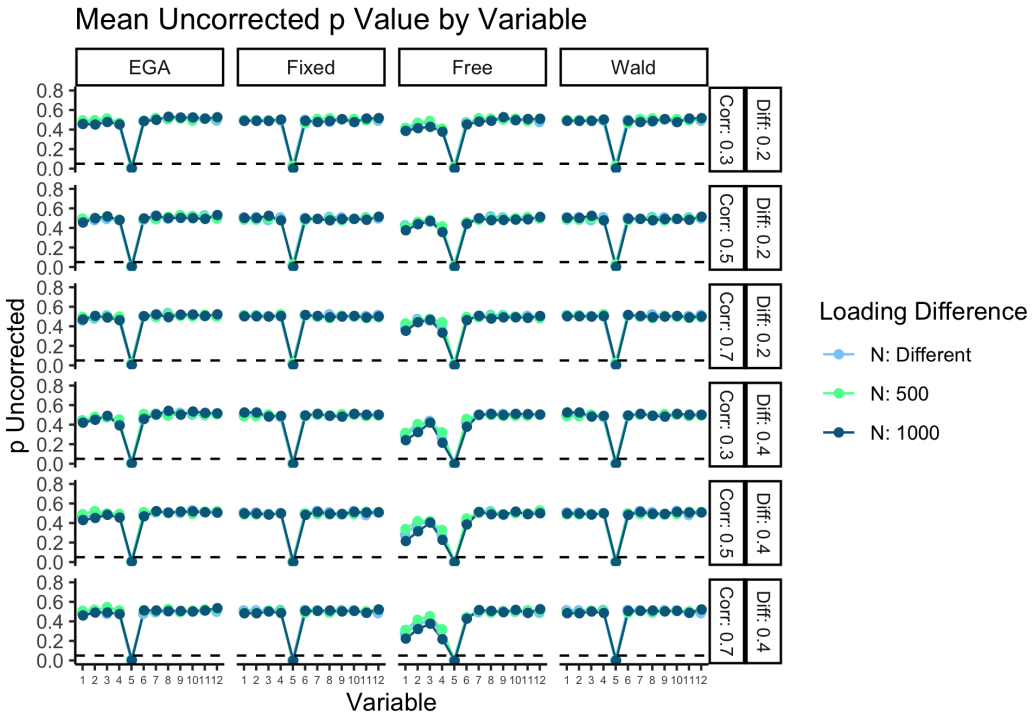

Figure 2

Mean uncorrected p-value for each item split by method and condition.

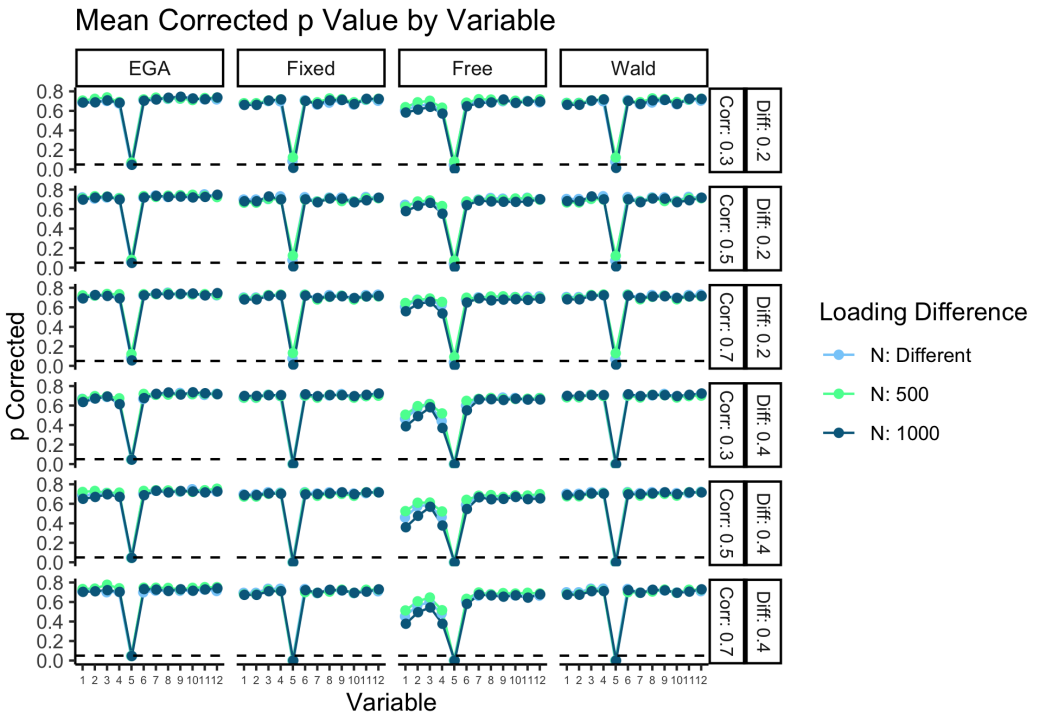

Figure 3

Mean MCP corrected p-value for each item split by method and condition.

Hit Rate

Figure 4 shows the average *Hit Rate* for both the MCP corrected and uncorrected *p*-values across items for each method by condition. In almost all cases, the MCP corrected *p*-values produce a higher mean *Hit Rate* than uncorrected *p*-values. When the difference in loadings is 0.40, EGA, Fixed, and Wald all have almost perfect *Hit Rate* across all variables. In this condition, the same trend arises for Free as in Figures 2 and 3: mean *Hit Rate* is lower in general for items in Factor 1 (Variables 1-6); however, its level of mean *Hit Rate* for Factor 2 (Variables 7-12), is more similar to that of the other three methods. In other words, having a higher effect size for EGA, Fixed, and Wald produces a higher *Hit Rate*, but for the Free method, in the factor where noninvariance is present, having a higher effect size does not improve the *Hit Rate*. For the Free method, all items in Factor 1 (where noninvariance is present) have a lower *Hit Rate* when compared to items in Factor 2 (where no noninvariance is present).

When the difference in loading is set to 0.20 for Variable 5, all four methods had lower mean *Hit Rate* when the *p*-value is MCP corrected as compared to the uncorrected *p*-value. This trend is most notable for Fixed, Free, and Wald when sample size is smaller or disparate, but does not appear when sample size is 1000. EGA only shows this trend when sample size is smaller and gradually becomes more different as the correlation between factors increases from 0.30 to 0.70. 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.

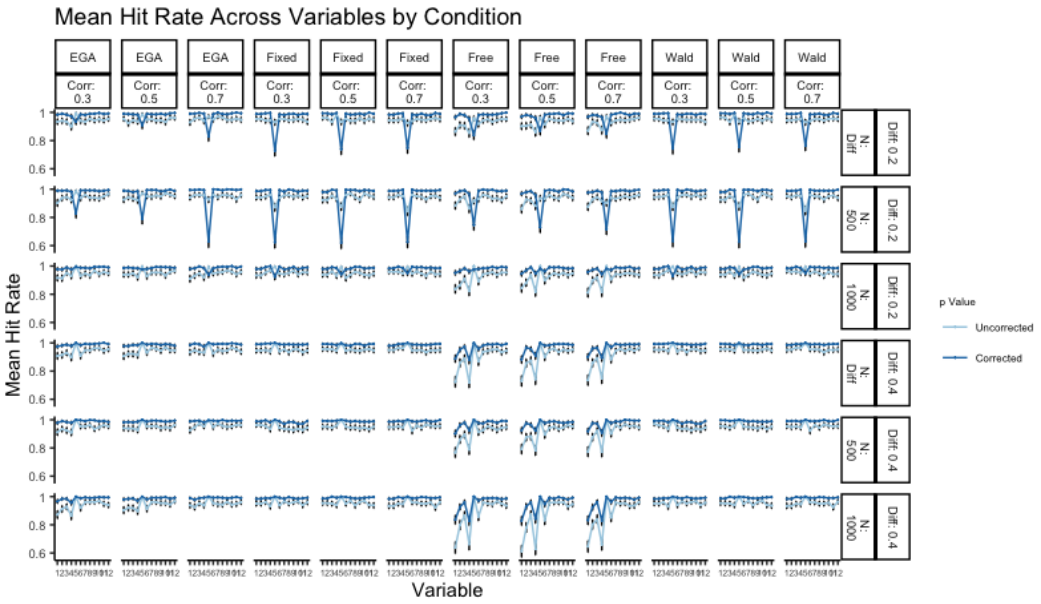

Figure 4

Mean Hit Rate across all items split by MCP corrected and uncorrected p-values, method, and condition.

Overall Metrics

Table 1 shows the overall metrics across the methods for both the MCP corrected and uncorrected *p*-values. When a correction was applied, an interesting pattern appears. Both *F1* 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 was applied, *Sensitivity* decreased for all four methods, most dramatically for Fixed and Wald, falling below 0.90. Conversely, *Sensitivity* increased slightly by applying the BH-procedure while *F1* values were dramatically increased by applying the BH-procedure going up, on average, by 0.14. When using MCP corrected *p*-values, EGA has the highest values for *F1* (0.91) and is tied for the highest *Specificity* with Fixed and Wald at 0.99. Free has the lowest values (using the MCP corrected *p*-values) of all of four methods for both *F1* (0.83) and Sensitivity (0.97).

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

Figure 5 shows the *Sensitivity* values split by MCP corrected and uncorrected *p*-values by method and data structure. When the difference in loadings is 0.40, all methods in all conditions have perfect *Sensitivity* regardless of whether or not the *p*-value was MCP corrected. When the difference in loadings is 0.20, uncorrected *p*-values lead to a higher level of *Sensitivity*. In this condition, almost perfect *Sensitivity* was achieved using MCP corrected *p*-values when sample size was 1000 for all methods. When the difference in loadings is set to 0.20, MCP corrected *p*-values were used, and sample size was smaller or disparate, EGA and Free performed better than Fixed and Wald. EGA, however, was more heavily influenced by the increase in correlation between factors; when the correlation between factors reached 0.70, EGA's performance fell below Free's to the same level as Fixed and Wald. Though

when the correlation between factors was 0.30 or 0.50, EGA outperformed Free. To sum up, a larger effect size always produced a perfect *Sensitivity* regardless of data condition or method. When the effect size is lower, using uncorrected *p*-values improves *Sensitivity*.

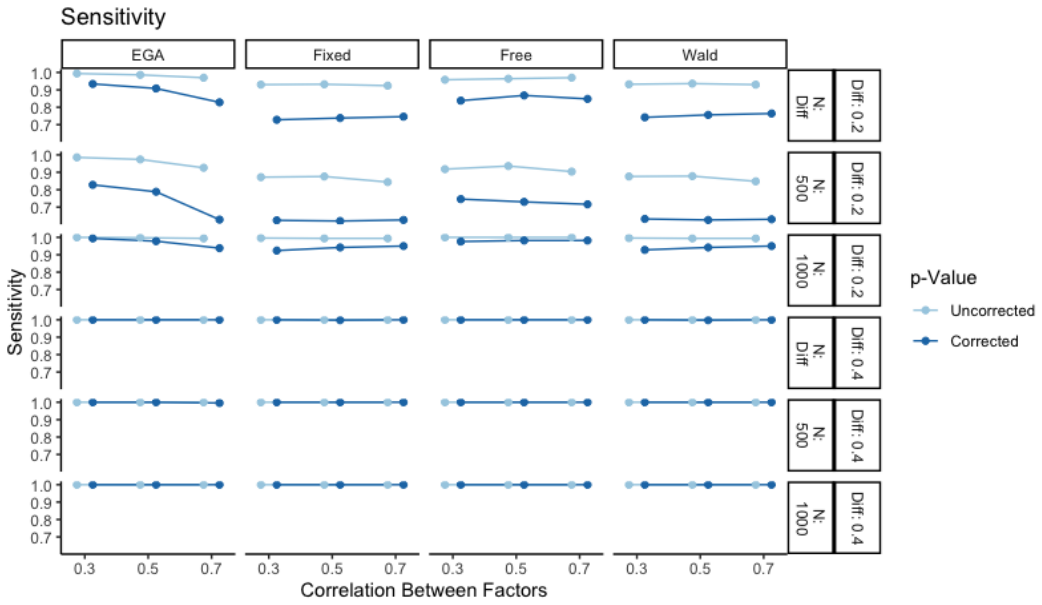

Figure 5

Sensitivity split by MCP corrected and uncorrected p-values, method, and condition.

F1

Figure 6 shows the *F1* values for MCP corrected and uncorrected *p*-values by method and data structure. In all conditions and across all four methods, MCP corrected *p*-values produce higher *F1* values than uncorrected *p*-values. When the difference between loadings is set to 0.40, 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 0.20 and *F1* is calculated using the MCP corrected *p*-values, a similar pattern arises that was seen in *Sensitivity*. EGA outperforms the other three methods when sample size is smaller or disparate. EGA, however, was more heavily influenced by the increase in correlation between factors; when the correlation between factors reached 0.70, EGA's performance fell below

Free's to the same level as Fixed and Wald. When the correlation between factors was 0.30 or 0.50, EGA outperformed Free. Overall, EGA, Free, and Wald perform similarly in their *F1* values and MCP corrected *p*-values produced higher *F1* values than uncorrected.

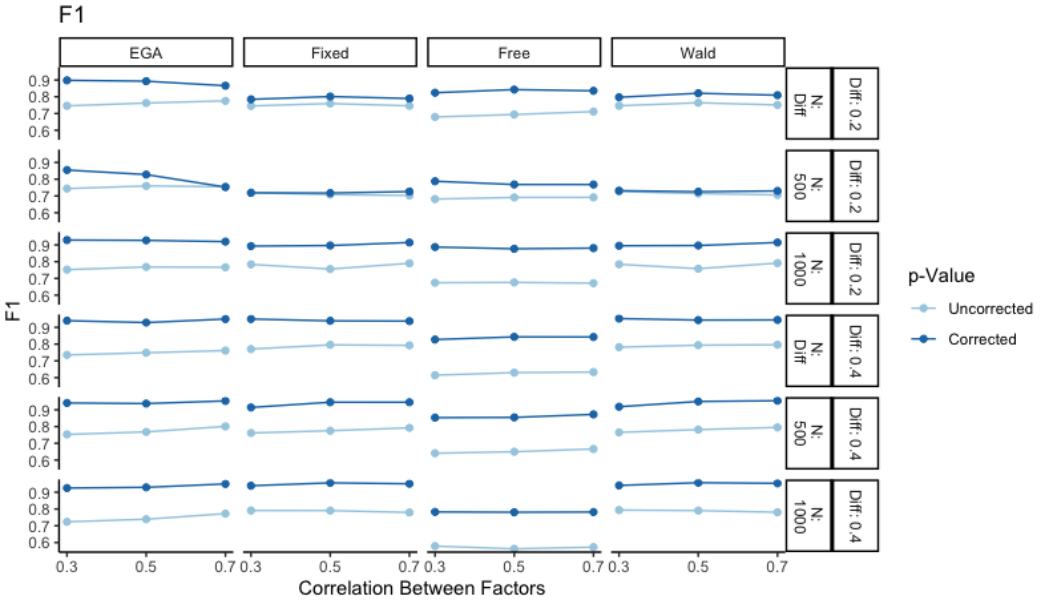

Figure 6

F1 split by MCP corrected and uncorrected p-values, method, and condition.

Specificity

Figure 7 shows the *Specificity* values for MCP corrected and uncorrected *p*-values by method and data structure. Across all these conditions, *Specificity* calculated using MCP corrected *p*-values is higher than uncorrected *p*-values. 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 0.20 to 0.40, 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).

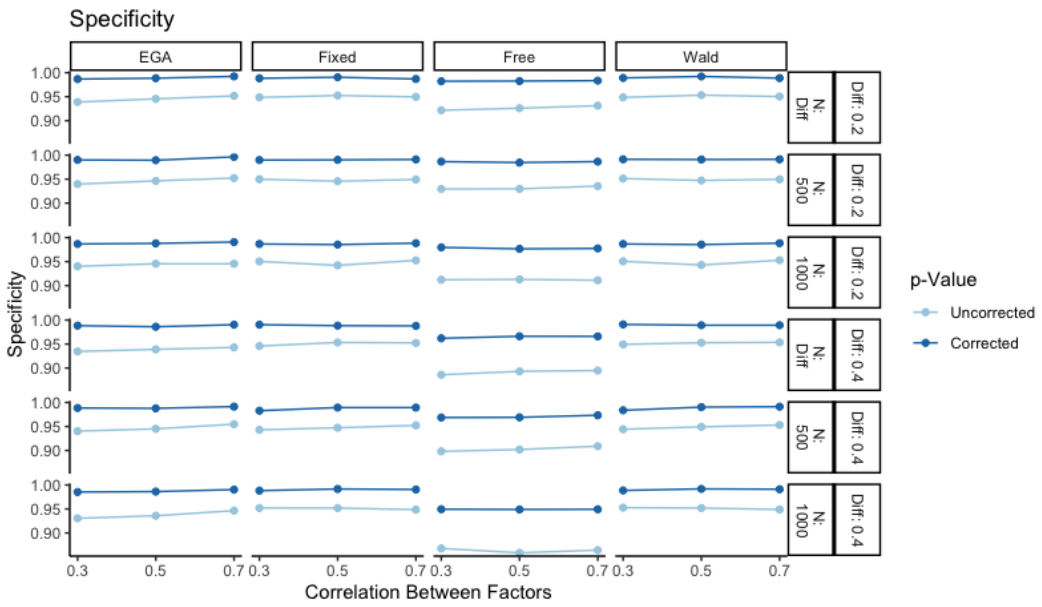

Figure 7

Specificity split by MCP corrected and uncorrected p-values, method, and condition.

Empirical Example: Testing Metric Invariance in the BAPQ

To demonstrate a substantive application of the proposed approach, a large dataset of the Broad Autism Phenotype Questionnaire (BAPQ; Hurley et al., 2007) was used. Data were provided by 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.

Notably, our simulation evaluated continuous data whereas most applied cases in psychology, including this example, use ordinal data (usually 2, 5, or 7 categories). Similarly, the *glasso* network estimation method is suited for continuous multivariate normal data.

Consistent with much of the (network) psychometrics literature, polychoric correlations were used, which assumes a bivariate normal distribution underlies the association between two variables. Some researchers have argued in favor of Spearman's correlations (Isvoranu & Epskamp, 2021); however, our goal for the applied example is to be consistent with the applied psychometrics literature, which uses polychoric correlations for ordinal variables in both factor and network analysis.

This questionnaire was given to the parents (mother and father) of an autistic child to assess the parent's 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. First, we load the {EGAnet} package and then we load the BAPQ dataset into the Global Environment as an objected called `bapq.all`.

```
# Load {EGAnet} package
library(EGAnet)

# Load data
load("bapq.all.RData")

# Column names in the `bapq.all` data frame
colnames(bapq.all)
```

```
 [1] "individual"         "measure"            "calculation_method"
 [4] "q01"                "q02"                "q03"
 [7] "q04"                "q05"                "q06"
[10] "q07"                "q08"                "q09"
[13] "q10"                "q11"                "q12"
[16] "q13"                "q14"                "q15"
```

```
[19] "q16"                "q17"                "q18"
[22] "q19"                "q20"                "q21"
[25] "q22"                "q23"                "q24"
[28] "q25"                "q26"                "q27"
[31] "q28"                "q29"                "q30"
[34] "q31"                "q32"                "q33"
[37] "q34"                "q35"                "q36"
[40] "status"             "Parent"
```

Next, we will run the `EGA()` function just for the mothers. To do this, we will need to specify four arguments in the `EGA()` function: `data`, `model`, `algorithm`, and `plot.EGA`. Since we are only interested in the mothers for this graph, we will subset the dataset to only include those rows which have the parent listed as mother. The data argument takes only those variables which we wanted included in the graph. Therefore, we select only columns 4 through 39 from the dataset as these are the items from the BAPQ. Next, we will set the model to `"glasso"`, algorithm to `"walktrap"`, and `plot.EGA` to `FALSE` so that we can create a plot with a customized title.

```
# Run EGA on each group and compare their structure
# EGA Mother
ega.mother <- EGA(
  data = bapq.all[bapq.all$Parent == "Mother", 4:39],
  model = "glasso", algorithm = "walktrap", plot.EGA = FALSE
)


# Plot EGA
plot(ega.mother, title = "Mother")
```

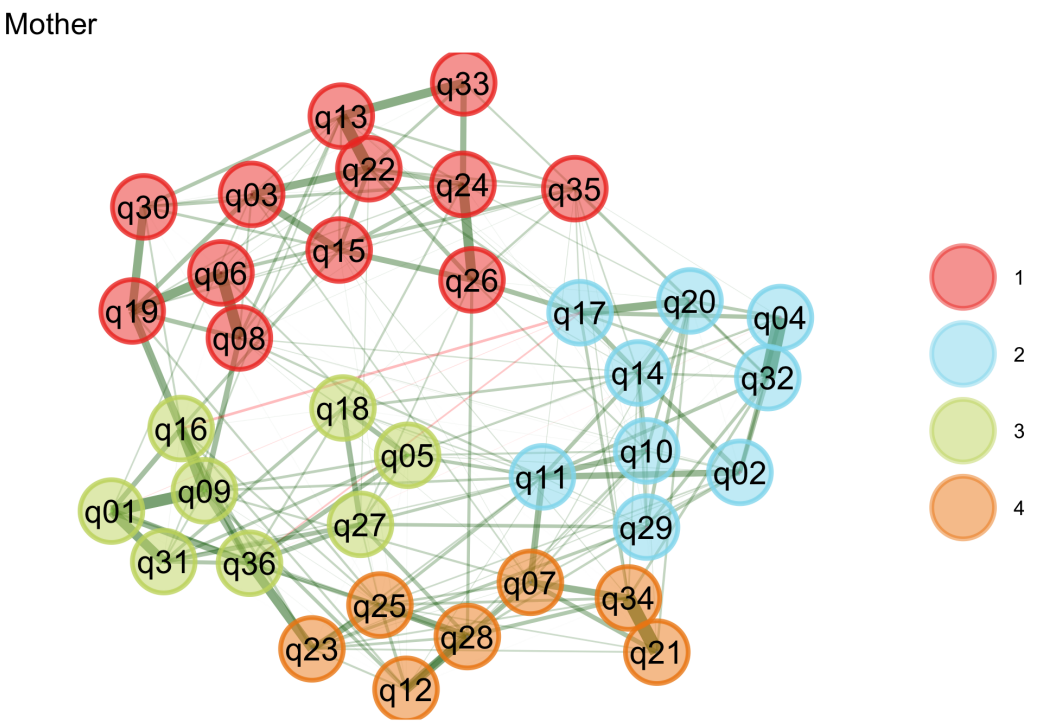

Figure 8
EGA of BAPQ data for mothers only.

This process can be repeated for fathers:

```
# Run EGA on each group and compare their structure

# EGA Father

ega.father <- EGA(

  data = bapq.all[bapq.all$Parent == "Father", 4:39],

  model = "glasso", algorithm = "walktrap", plot.EGA = FALSE

)


# Plot EGA

plot(ega.father, title = "Father")
```

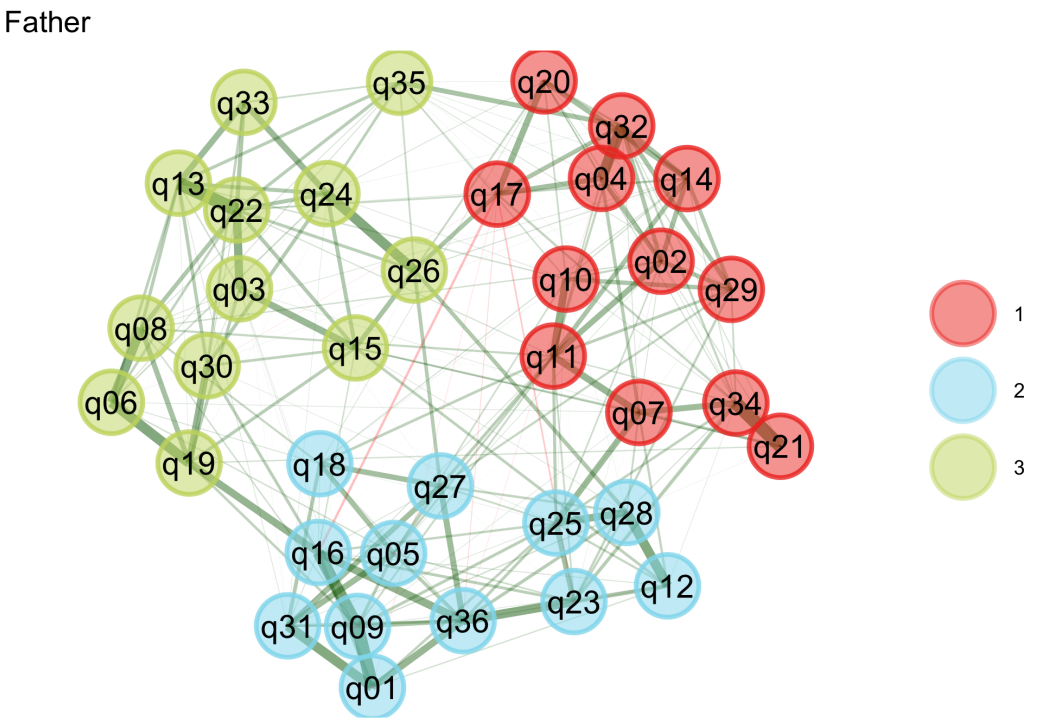

Figure 9
EGA of BAPQ data for fathers only.

Visually we can see that the two graphs contain nonequivalent community structures. In order to reach configural invariance, we assess the stability of the items on the full sample. Investigating item stability allows us to see which items are consistently organized into the same communities. We look at the stability of the items in the entire sample because instability at this level could be due to items being partitioned into different communities when split into subsamples. First, `bootEGA()` is run to iteratively resample and estimate EGA. We only need to set two arguments: `data` and `iter`. The data argument takes the full sample so we do not need to subset the rows. Similarly to the `EGA()` function it can only take those columns we wish to include in the graph so we only select columns 4 through 39. Then, `dimensionStability()` is used to compute the stability of the items. This function only needs to be given the bootEGA object to run.

```
# Bootstrap EGA

set.seed(1) # for reproducibility

boot.bapq <- bootEGA(data = bapq.all[,4:39], iter = 500,

                     uni.method = "LE", algorithm = "walktrap")


# Dimension Stability

bapq.stability <- dimensionStability(boot.bapq)
```

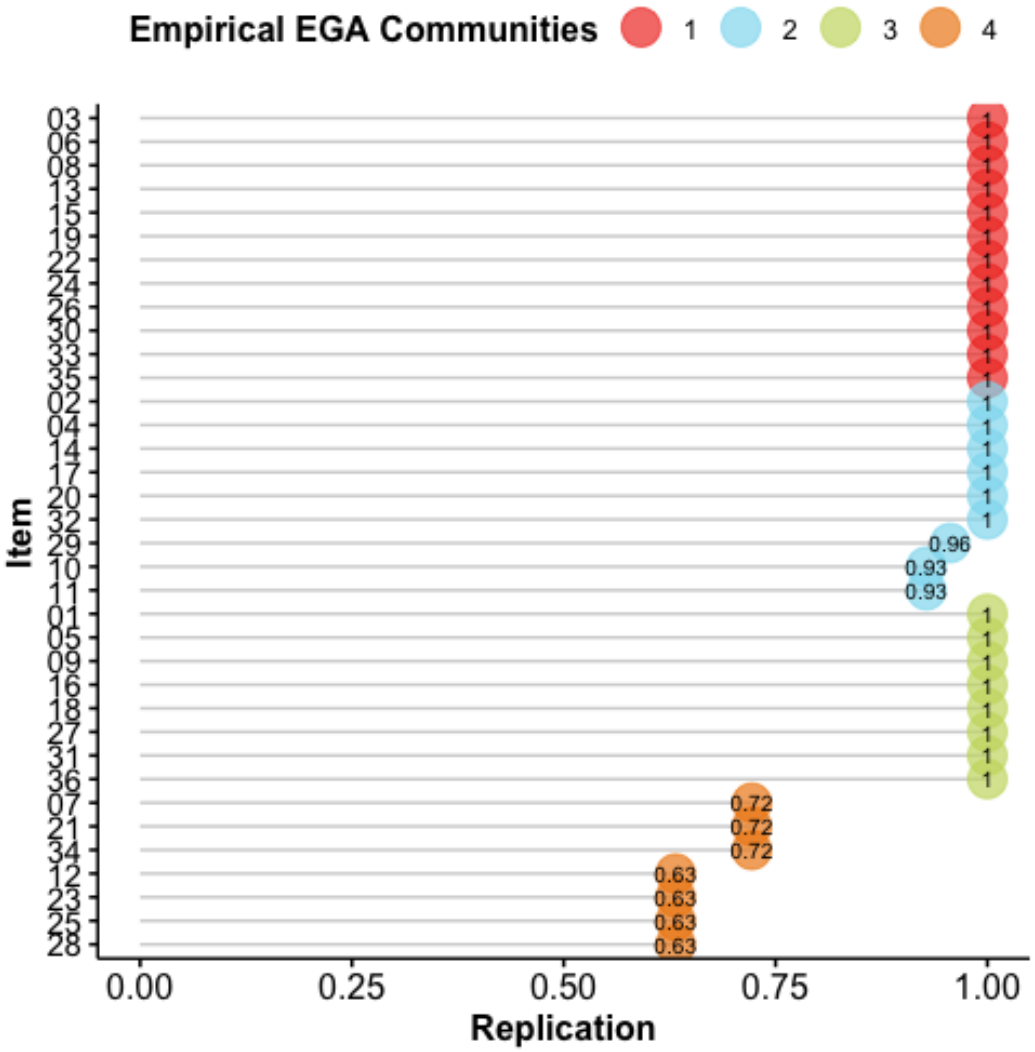

Figure 10
BAPQ item stability for all items using the full sample.

Looking at Figure 10, there are four items with an item stability less than 0.70. Let's see what happens to the configural invariance if we remove them. To remove them, we will create an object called `remove` which contains all the names of the items with a stability of less than 0.70. Then, we will create a dataset called `bapq.stable` which does not contain those four variables. After, we will rerun `bootEGA()` and `dimensionStability()` on the new dataset.

```
# Creating a vector containing the names of unstable items
remove <-
  bapq.stability$item.stability$item.stability$empirical.dimensions < 0.70


# Print variable names
names(remove[which(remove==TRUE)])
```

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

```
# Creating a new data frame containing only stable items
bapq.stable <- bapq.all[,4:39][,names(remove[which(remove==FALSE)])]
bapq.stable <- cbind(bapq.all[,c(1:3)],bapq.stable,bapq.all[,c(40,41)])
# Column names of the bapq.stable dataframe
colnames(bapq.stable)
```

```
 [1] "individual"         "measure"            "calculation_method"
 [4] "q01"                "q02"                "q03"
 [7] "q04"                "q05"                "q06"
[10] "q07"                "q08"                "q09"
[13] "q10"                "q11"                "q13"
[16] "q14"                "q15"                "q16"
```

```
[19] "q17"                "q18"                "q19"
[22] "q20"                "q21"                "q22"
[25] "q24"                "q26"                "q27"
[28] "q29"                "q30"                "q31"
[31] "q32"                "q33"                "q34"
[34] "q35"                "q36"                "status"
[37] "Parent"
```

```r
# bootEGA
set.seed(1)
boot.bapq2 <- bootEGA(bapq.stable[,4:35], iter = 500)


# dimensionStability
bapq.stability2 <- dimensionStability(boot.bapq2)
```

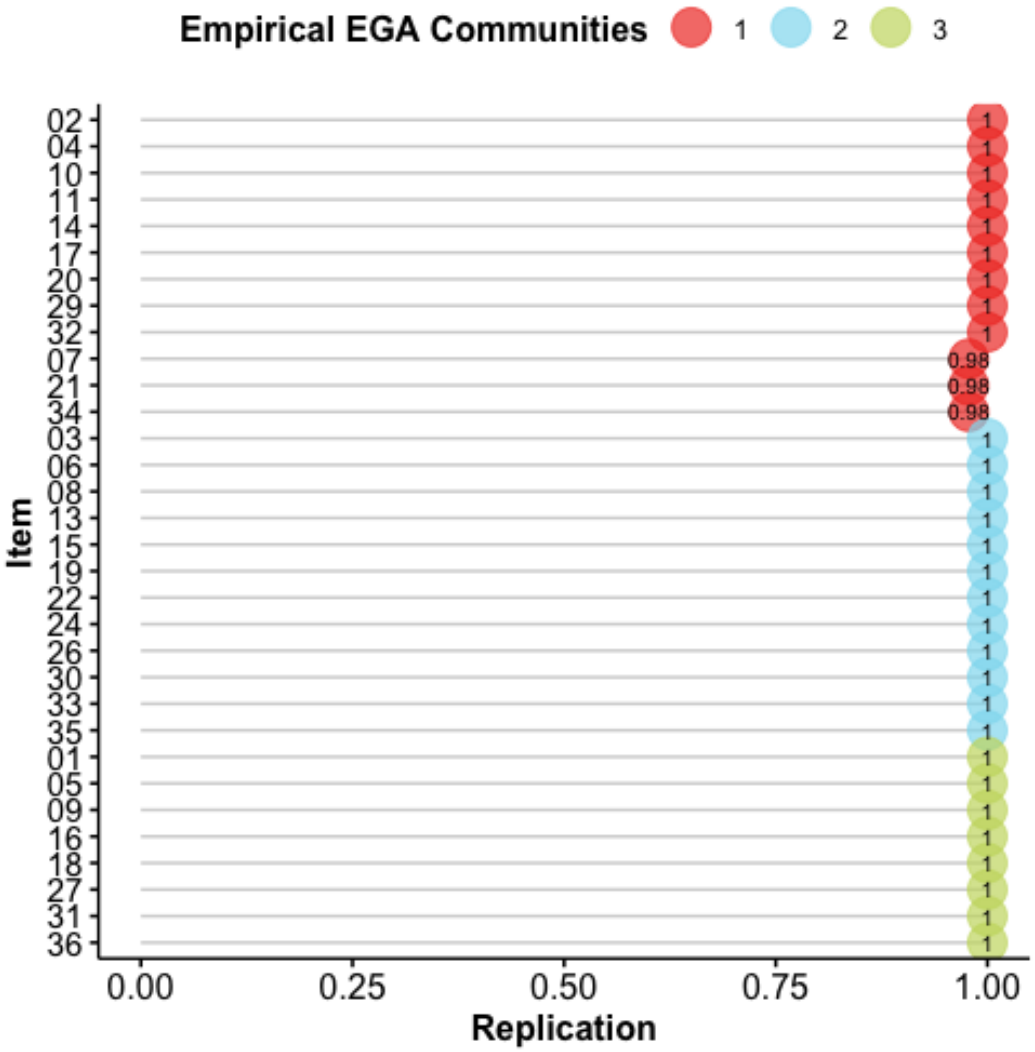

Figure 11
BAPQ item stability without unstable items using the full sample.

Now, we have a highly stable three community structure. We can now rerun EGA on each group individually and check to see if this holds within the samples. We will rerun the `EGA()` function on the `bapq.stable` dataset in the same way as above separately for mothers and fathers.

```
# Run EGA on each group and compare their structure

# EGA Mother

ega.mother2 <- EGA(
```

```
  bapq.stable[bapq.all$Parent == "Mother", 4:35],
  model = "glasso", algorithm = "walktrap", plot.ega = FALSE
)
plot(ega.mother2, title = "Mother")
```

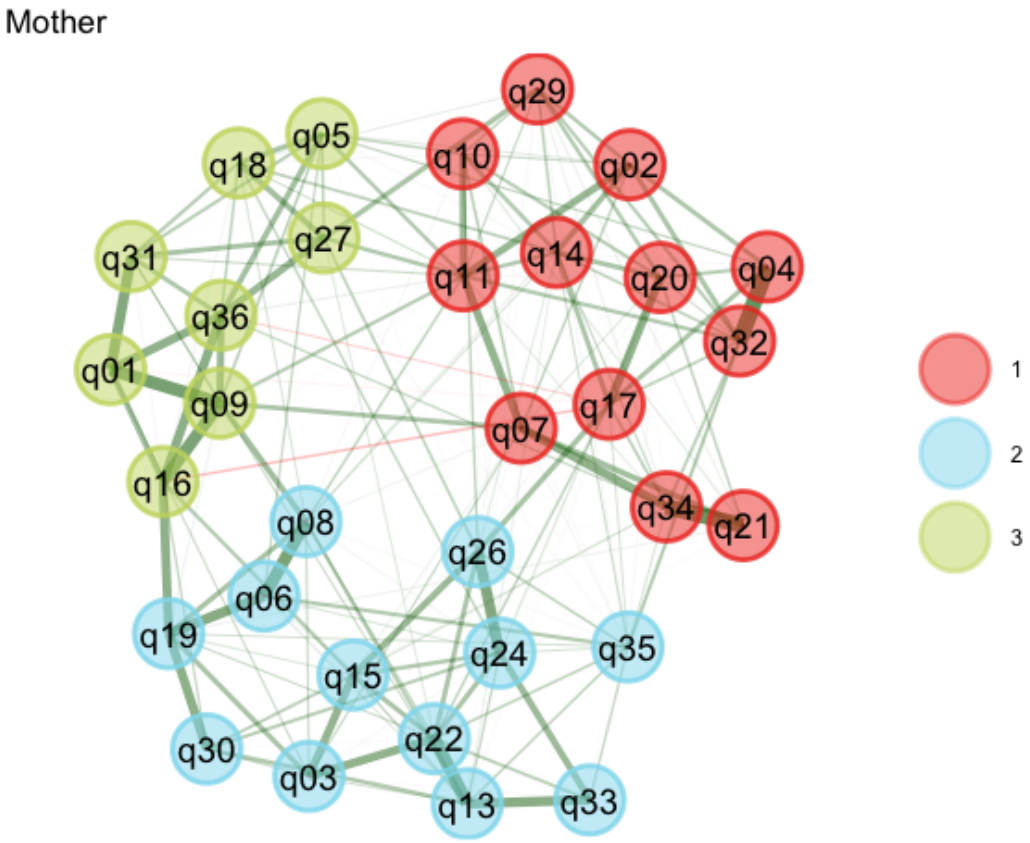

Figure 12
EGA of BAPQ data for mothers only without unstable items.

```
# Run EGA on each group and compare their structure
# EGA Father
ega.father2 <- EGA(
  bapq.stable[bapq.all$Parent == "Father", 4:35],
  model = "glasso", algorithm = "walktrap", plot.ega = FALSE
)
```

```
plot(ega.father2, title = "Father")
```

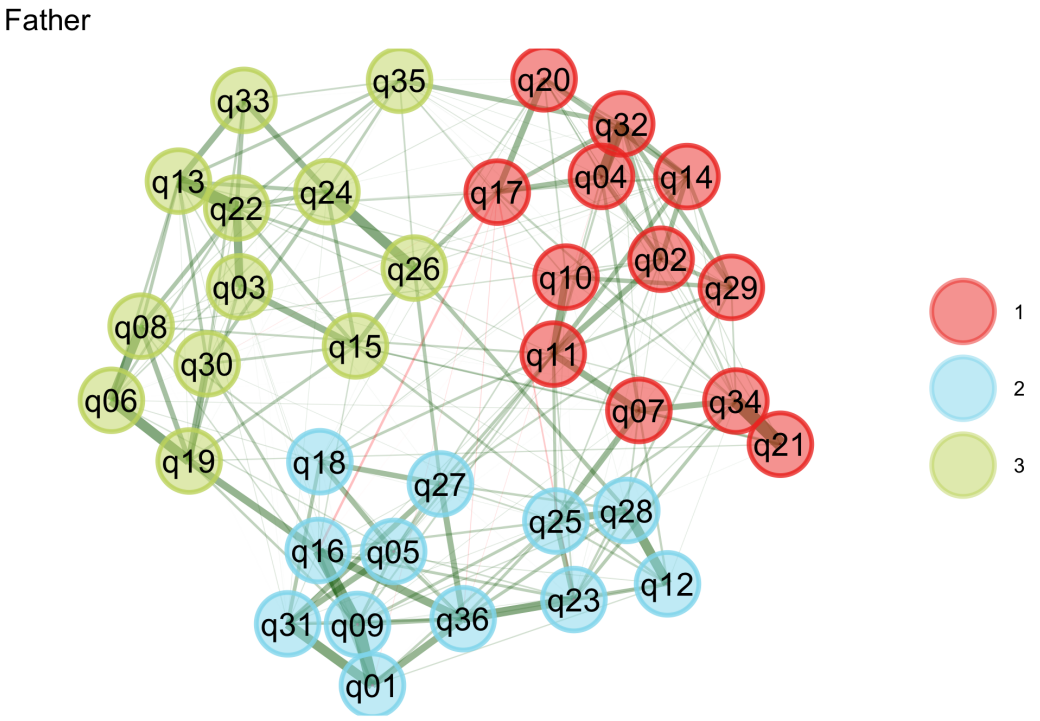

Figure 13
EGA of BAPQ data for fathers only without unstable items.

Visually, we can see that now each graph contains three communities, all of which appear to have the same items in them. We can confirm that by comparing the `wc` object within each saved EGA analysis. The `wc` object produced by `EGA()` is a vector of the item to community assignments in the network.

```
ega.mother2$wc
```

```
q01 q02 q03 q04 q05 q06 q07 q08 q09 q10 q11 q13 q14 q15 q16 q17 q18 q19 q20 q21
  3   1   2   1   3   2   1   2   3   1   1   2   1   2   3   1   3   2   1   1
q22 q24 q26 q27 q29 q30 q31 q32 q33 q34 q35 q36
  2   2   2   3   1   2   3   1   2   1   2   3
```

We can compute the normalized mutual information between the `wc` objects with a value of 1 meaning identical:

```
igraph::compare(ega.mother2$wc, ega.father2$wc, method = "nmi")
```

```
[1] 1
```

Since all `wc` are the same, then we can say we have established configural invariance. We can now test for metric invariance. This can be done using the `invariance()` function from the {EGAnet} package and then investigate which items are significant (noninvariant). We will investigate both uncorrected and MCP corrected p-values. To obtain MCP corrected p-values, we can use the `p.adjust()` function to apply the BH-procedure.

```
met.invariance <- invariance(data = bapq.stable[,4:35],
                             groups = bapq.stable$Parent, gamma = 0)
plot(met.invariance)
```

```
# Metric Invariance
set.seed(1)
results <- invariance(bapq.stable[,4:35],bapq.stable$Parent)
```

```
# Applying BH-procedure
adjusted.p <- p.adjust(
  results$results$p, method = "BH",
  n = length(results$results$p)
)

# Uncorrected p
results$results[results$results$p < .05,]
```

	Node	Membership	Difference	p	sig	Direction
q31		1	0.033	0.034	*	Father > Mother
q11		2	0.038	0.026	*	Father > Mother
q20		2	-0.059	0.002	**	Father < Mother
q21		2	-0.043	0.004	**	Father < Mother
q29		2	0.044	0.014	*	Father > Mother
q03		3	-0.032	0.048	*	Father < Mother
q22		3	0.039	0.030	*	Father > Mother

Figure 14 shows a visual representation of these results. The graphs for both Mothers and Fathers are shown, with more transparently shaded nodes indicating metric invariance, while solidly shaded nodes indicate metric non-invariance.

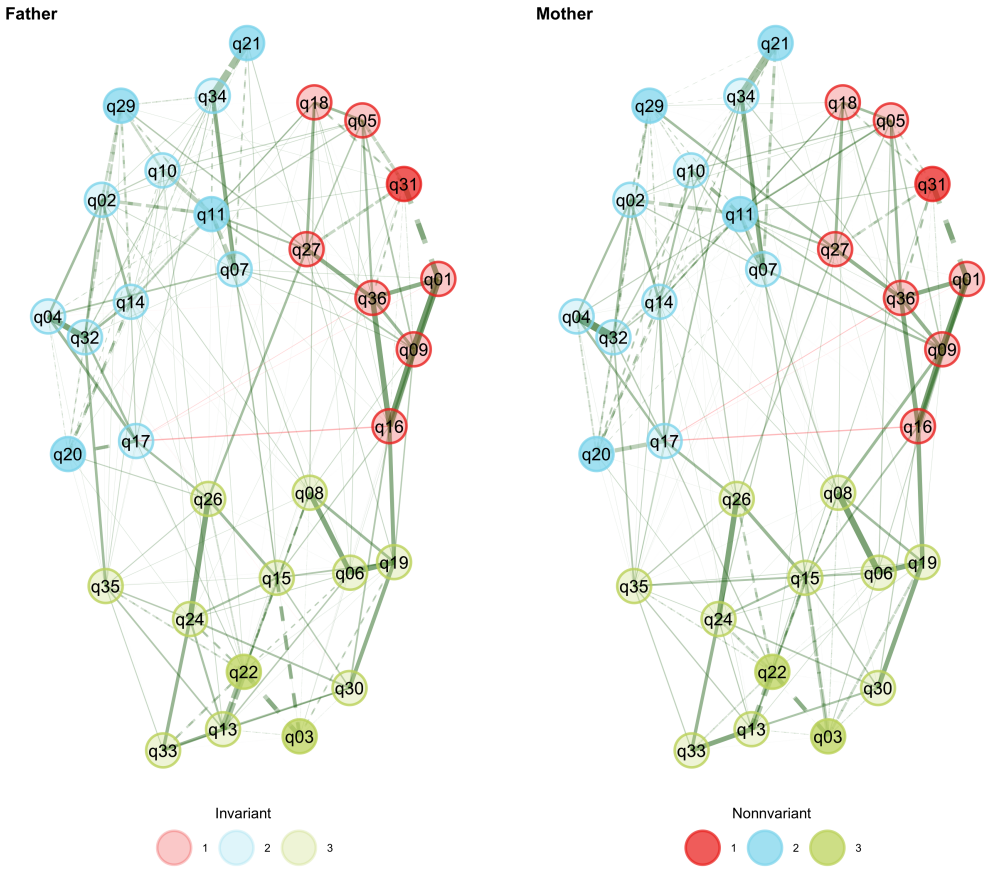

Figure 14
EGA metric invariance plot with solid nodes indicating non-invariant items.

```r
# Corrected p

results$results[adjusted.p < .05,]
```

```
    Node Membership Difference   p-Adjusted sig Direction
q20          2     -0.059  0.064  ** Father < Mother
q21          2     -0.043  0.064  ** Father < Mother
```

The item text for the 7 items showing metric noninvariance using an uncorrected *p*-value are as follows:

- "I prefer to be alone rather than with others."
- "I feel disconnected or "out of sync" in casual interaction with acquaintances."
- "I speak too loudly or softly."
- "I can tell when someone is not interested in what I am saying."
- "I leave long pauses in conversation."
- "I am comfortable with unexpected changes in plans."
- "I have a hard time dealing with changes in my routine."

Using the BH-procedure to account for multiple comparisons, only items *"I speak too loudly or softly."* and *"I can tell when someone is not interested in what I am saying."* are identified as metric noninvariant at the 0.05 level.

Discussion

Establishing measurement invariance is crucial for the use of any measurement across groups in any applied setting. FA approaches are the most commonly used methods to test measurement invariance. Previous research within network psychometrics has established a handful of methods for comparing networks, but nothing comparable to FA that accounts for dimensionality. With network loadings, metric invariance in network psychometrics is tractable. The current paper proposed a method to test for metric invariance using a permutation test on network loadings in the EGA framework. To compare the proposed method to existing FA methods, a simulation study was conducted manipulating sample size, loadings difference, and correlation between factors to compare the proposed method to traditional methods in FA. Three methods in the FA framework were used to test partial metric invariance: Fixed, Free, and Wald. In all four methods, we tested for configural, metric, and partial metric invariance. In general, the novel network psychometrics approach performed as well as, and in some cases, better than the FA methods.

Testing for partial metric invariance in all methods necessitated a MCP due to multiple hypotheses being tested. Most MCPs control FWER with the BH-procedure correcting for a portion of significant findings. Given the negligible consequences of falsely identifying noninvariant items, our suggestion is that there should be greater emphasis on correctly identifying noninvariant items rather than invariant items. The BH-procedure is recommended because it balances what might otherwise be overly conservative correction using other MCP methods (e.g., Bonferroni).

With data simulated across two groups with two factors and one item on one factor, results indicated that applying the BH-procedure provided a gain in the correct identification of invariant items but not noninvariant items. This result is particularly true when the difference between loadings was small suggesting that the p-value correction may cause truly noninvariant items to be classified as invariant. Our results demonstrate this effect on the

Hit Rate: average *Hit Rate*, in general, was higher for for all items except noninvariant items.

For the specific methods evaluated in this study, the proposed method performed similarly to FA methods Fixed and Wald and in certain data conditions performed slightly better. The Free method showed some inconsistencies and was outperformed by the proposed method, Fixed, and Wald. 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 affected the accuracy for the noninvariant item with the uncorrected p-value being more accurate in some cases. The proposed method was slightly less influenced by this effect than the other methods. With a smaller or different sample sizes, the proposed method's accurate identification of noninvariant item was slightly better than the other methods. Importantly, as the correlation between factors increased, the accuracy decreased when sample size was either smaller or disparate. All four methods are performing highly and comparably at identifying invariant variables. The Free method overall showed lower accuracy overall but performed comparably at correctly identifying noninvariant items.

All together, these results indicate that the proposed method is comparable to traditional FA methods. One promising result from our study is that the proposed method was less impacted by disparate sample sizes, which are a known challenge for FA methods (F. F. Chen, 2007; Kaplan & George, 1995). Another benefit of the network psychometrics approach is that it does not require any intensive model specifications (in contrast to FA methods) and can be feasibly implemented in a few lines of code in commonly used software (`R`). The ease of this approach mitigates many of the concerns that are raised about the proper application of measurement invariance methods (Schroeders & Gnambs, 2018).

In terms of applying an MCP, 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 FA. This finding is problematic because the goal of the method is to properly identify noninvariant items. Our findings question whether an MCP is useful to identify noninvariant items. The *Hit Rate* results 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. As is also seen by the results of *Specificity*, there is a negligible effect of falsely identifying an item as noninvariant. As a general guideline, we recommend that noninvariant variables identified both MCP corrected and uncorrected p-values should be evaluated. The researcher can at that point can conduct a risk assessment based on their specific research question and context. Another alternative is to change the α level when applying the MCP. In the Appendix we have included the all results with an additional condition where the MCP corrected p-values are assessed for significance at the $\alpha = .10$ level. The results indicate that this method slightly improves the accuracy of identifying noninvariant items for the proposed method but makes no impact for FA.

The use of any latent variable measure across qualitatively distinct groups should necessitate the testing of measurement invariance. The proposed method performs well at identifying noninvariant items across data conditions and shows promise at improved identification over FA when sample sizes are smaller or disparate. Although network psychometrics have different substantive interpretations than FA models, the proposed method is statistically consistent with FA measurement invariance methods when the data generating model is a factor model. An added benefit of the network psychometric approach is that it holds less stringent assumptions about the data (e.g., local independence) and therefore can be applied in broader contexts (e.g., topic modeling) (H. Golino et al., 2022; Kjellström & Golino, 2019).

One limitation of the present study is that the simulated conditions were narrow. The generating model always had two factors with six variables and only one variable as noninvariant. Further, we generated data was continuous and without skew. In applied data, it is more common to have ordinal data with skew. Future work should verify our findings in

more extensive conditions that better mirror applied data (e.g., categorical and skew data, multiple factors, different number of variables per factor, multiple noninvariant items on one or more factors). Another limitation is that this study only measurement invariance when the data generation model was a factor model. Under applied circumstances, psychologists have considered nearly all scales as a latent variable model and therefore our data generation follows the assumptions that most applied researchers already hold about their data. That said, the effectiveness of measurement invariance methods, including the one proposed here, should be evaluated with alternative data mechanisms such as a small-world network model. Before recommending the proposed approach over more traditional FA methods, further investigation into these limitations is necessary.

Appendix

Hit Rate

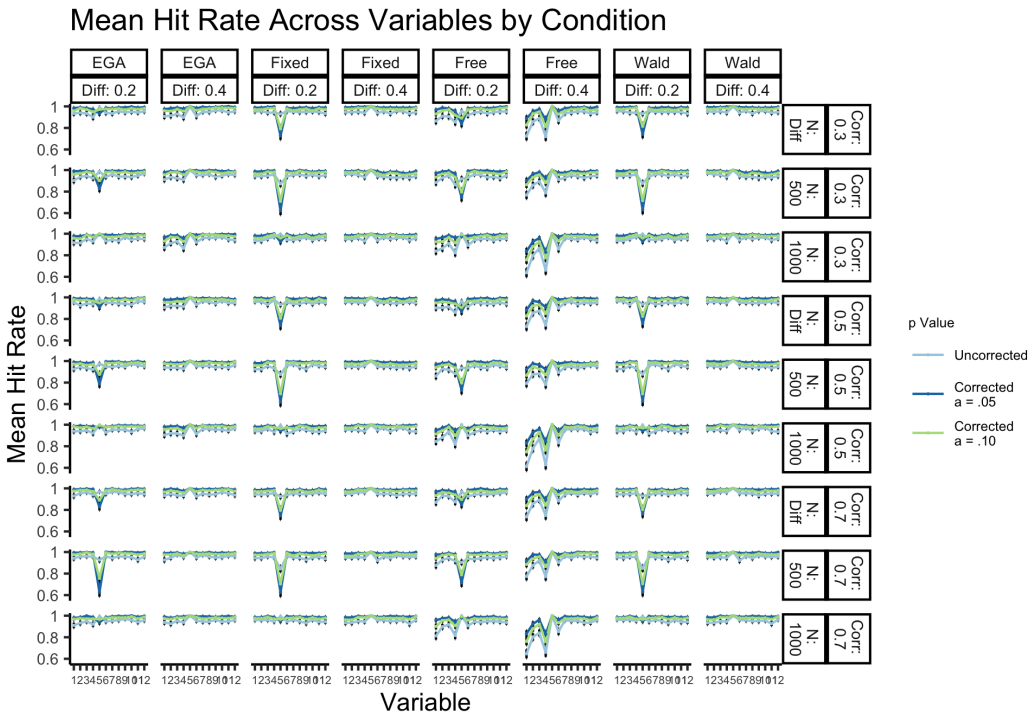

Figure 15

Mean Hit Rate across all items split by MCP corrected and uncorrected p-values, method, and condition.

Overall Metrics

Table 2

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

Sensitivity

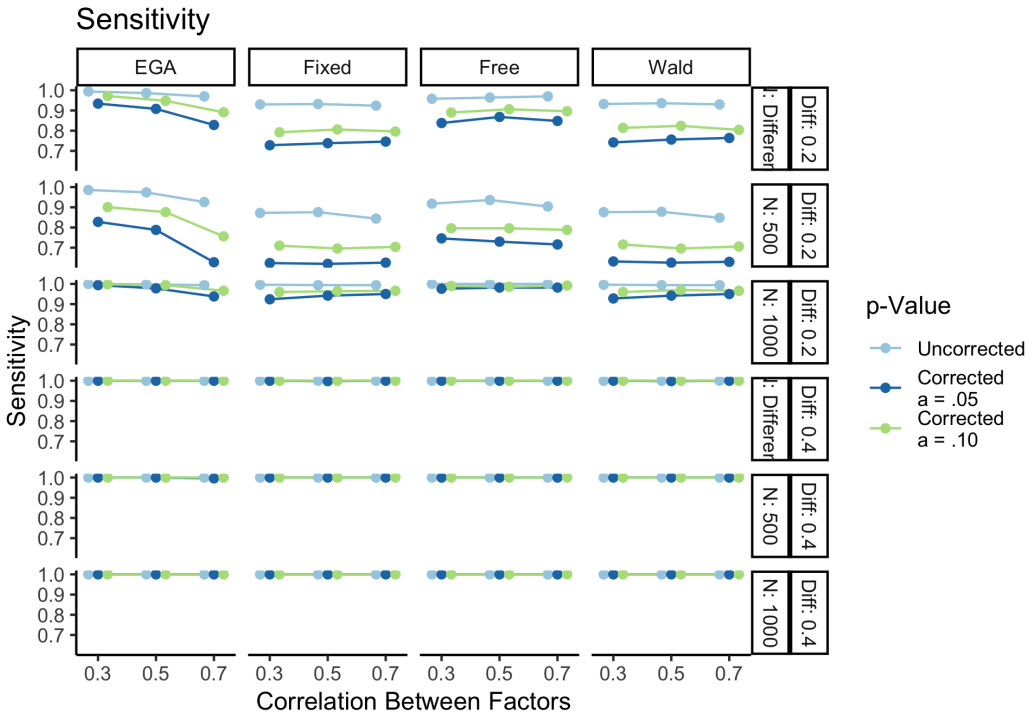

Figure 16

Sensitivity split by MCP corrected and uncorrected p-values, method, and condition.

F1

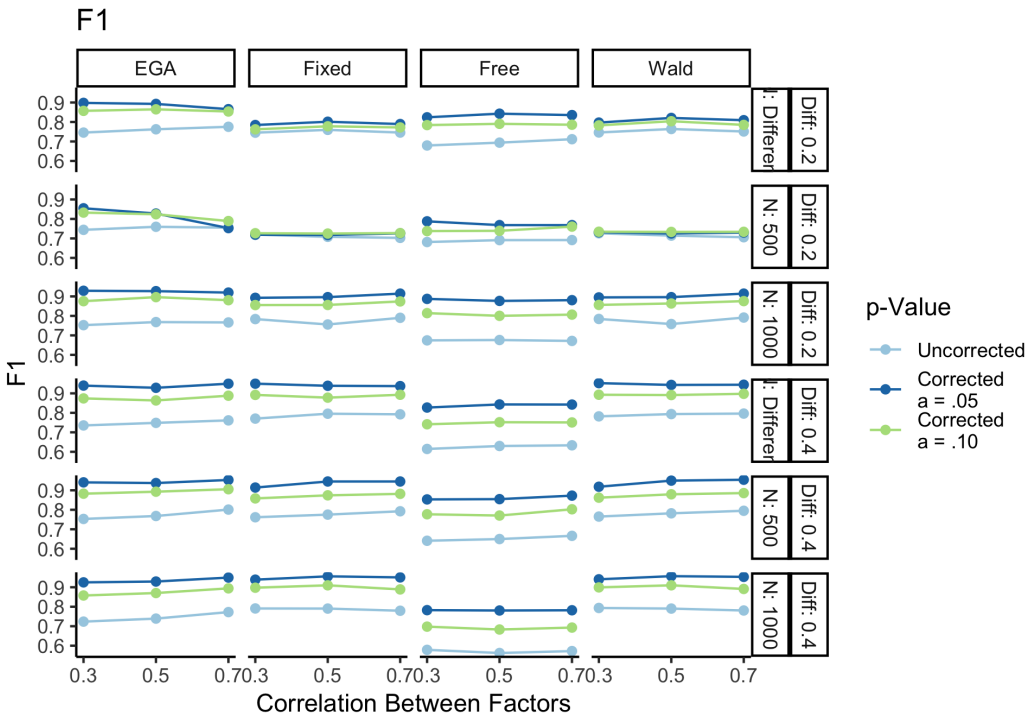

Figure 17

F1 split by MCP corrected and uncorrected p-values, method, and condition.

Specificity

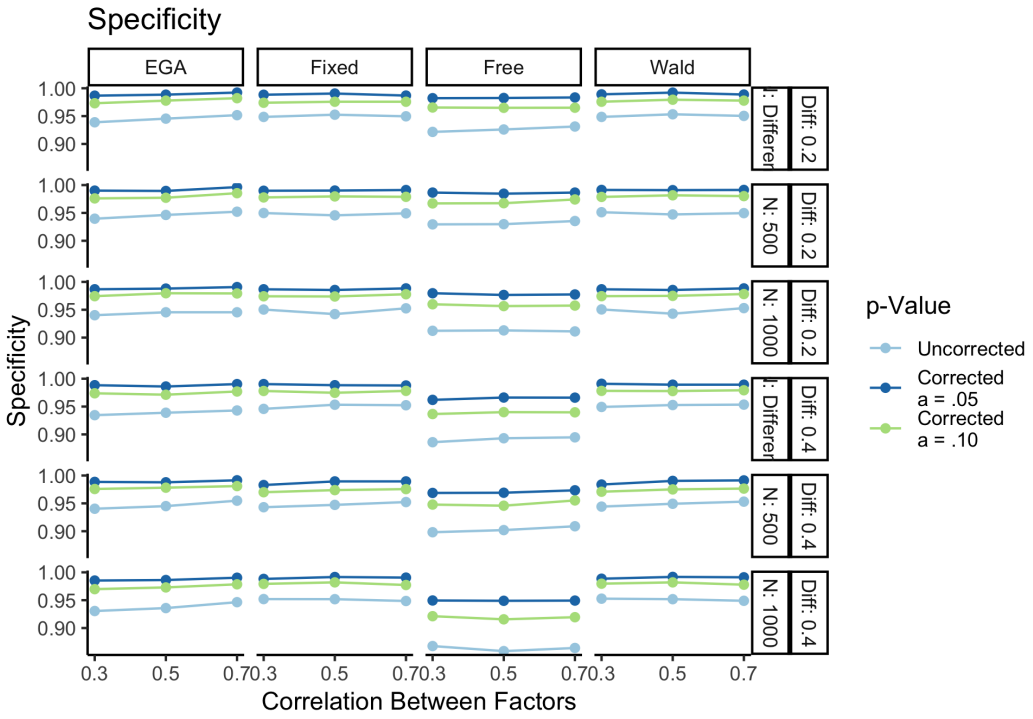

Figure 18

Specificity split by MCP corrected and uncorrected p-values, method, and condition.

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.

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

Bringmann, L. F., Elmer, T., Epskamp, S., Krause, R. W., Schoch, D., Wichers, M., Wigman, J., & Snippe, E. (2019). What do centrality measures measure in psychology networks? *Journal of Abnormal Psychology*, *128*, 892–903.

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

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

Chen, J., & Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), 759–771.

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

Christensen, A. P., Garrido, L. E., Guerra-Peña, K., & Golino, H. (2023). Comparing community detection algorithms in psychometric networks: A monte carlo simulation. *Behavior Research Methods*, 1–21.

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.

Christensen, A. P., & Golino, H. (2021b). On the equivalency of factor and network loadings. *Behavior Research Methods*, *53*(4), 1563–1580.

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.

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*, *23*(4), 617.

Feliciano, P., Daniels, A. M., Snyder, L. G., Beaumont, A., Camba, A., Esler, A., Gulsrud, A. G., Mason, A., Gutierrez, A., Nicholson, A., et al. (2018). SPARK: A US cohort of 50,000 families to accelerate autism research. *Neuron*, *97*(3), 488–493.

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.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PloS One*, *12*(6).

Golino, H., & Christensen, A. P. (2022). *EGAnet: Exploratory graph analysis – a framework for estimating the number of dimensions in multivariate data using network psychometrics*.

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*, 1–32.

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*.

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.

Haslbeck, J., & Bork, R. van. (2022). Estimating the number of factors in exploratory factor analysis via out-of-sample prediction errors. *Psychological Methods*.

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.

Isvoranu, A.-M., & Epskamp, S. (2021). Which estimation method to choose in network psychometrics? Deriving guidelines for applied researchers. *Psychological Methods*.

Jones, P. J., Mair, P., Simon, T., & Zeileis, A. (2020). Network trees: A method for recursively partitioning covariance structures. *Psychometrika*, *85*(4), 926–945.

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

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.

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.

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.

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.

Kuhn, M. (2022). *Caret: Classification and regression training*. https://CRAN.R-project.org/package=caret

Lauritzen, S. L. (1996). *Graphical models* (Vol. 17). Clarendon Press.

Ludbrook, J., & Dudley, H. (1998). Why permutation tests are superior to t and f tests in

biomedical research. *The American Statistician*, *52*(2), 127–132.

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.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*, 8577–8582.

Pons, P., & Latapy, M. (2006). Computing communities in large networks using random walks. *J. Graph Algorithms Appl.*, *10*(2), 191–218.

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.

R Core Team. (2022). *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.

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.

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

Revelle, W. (2017). *psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. https://CRAN.R-project.org/package=psych

Rosseel, Y. (2012). Lavaan: An r package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36.

Schroeders, U., & Gnambs, T. (2018). Degrees of freedom in multigroup confirmatory factor

analyses. *European Journal of Psychological Assessment*.

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

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.

Steinberg, L. (2001). The consequences of pairing questions: Context effects in personality measurement. *Journal of Personality and Social Psychology*, *81*(2), 332.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288.

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*.

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.

Ward, J. H. (1963). Hierarchical clustering to optimise an objective function. *Journal of the American Statistical Association*, *58*, 238–244.

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.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. *The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research*.

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.

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.

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*, 1–19.