

CHRISTENSEN, ALEXANDER P., Ph.D. Towards a Network Psychometrics Approach to Assessment: Simulations for Redundancy, Dimensionality, and Loadings. (2020)

Directed by Dr. Paul J. Silvia. 106 pp.

Research using network models in psychology has proliferated over the last decade. The popularity of network models has largely been driven by their alternative explanation for the emergence of psychological attributes—observed variables co-occur because they are causally coupled and dynamically reinforce each other, forming cohesive systems. Despite their rise in popularity, the growth of network models as a psychometric tool has remained relatively stagnant, mainly being used as a novel measurement perspective. In this dissertation, the goal is to expand the role of network models in modern psychometrics and to move towards using these models as a tool for the validation of assessment instruments. This paper presents three simulation studies and an empirical example that are designed to evaluate different aspects of the psychometric network approach to assessment: reducing redundancy, detecting dimensions, and estimating loadings. The first simulation evaluated two novel approaches for determining whether items are redundant, which is a key component for the accuracy and interpretation of network measures. The second simulation evaluated several different community detection algorithms, which are designed to detect dimensions in networks. The third simulation evaluated an adapted formulation of the network measure, node strength, and how it compares to factor loadings estimated by exploratory and confirmatory factor analysis. The results of the simulations demonstrate that network models can be used as an effective psychometric tool and one that is on par with more

traditional methods. Finally, in the empirical example, the methods from the simulations are applied to a real-world dataset measuring personality. This example demonstrated that these methods are not only effective, but they can validate whether an assessment instrument is consistent with theoretical and empirical expectations. With these methods in hand, network models are poised to take the next step towards becoming a robust psychometric tool.

This represents the final committee formatted dissertation manuscript. Studies in this manuscript will be published separately. This manuscript will remain unpublished. If citing, please use:

Christensen, A. P. (2020). Towards a network psychometrics approach to assessment: Simulations for redundancy, dimensionality, and loadings (Unpublished doctoral dissertation). University of North Carolina at Greensboro, Greensboro, NC, USA.

TOWARDS A NETWORK PSYCHOMETRICS APPROACH TO ASSESSMENT:
SIMULATIONS FOR DIMENSIONALITY, LOADINGS, AND REDUNDANCY

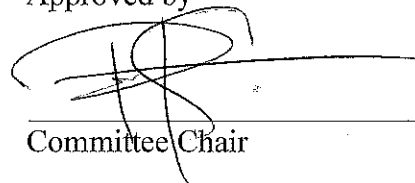
by

Alexander P. Christensen

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2020

Approved by



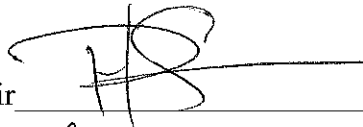
Committee Chair

To Gilbert, my furry collaborator.

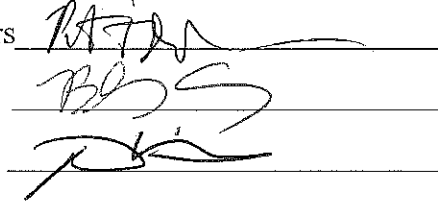
APPROVAL PAGE

This thesis written by ALEXANDER P. CHRISTENSEN has been approved by the following committee of the Faculty of The Graduate School at the University of North Carolina at Greensboro.

Committee Chair



Committee Members



03/13/2020
Date of Acceptance by Committee

03/13/2020
Date of Final Oral Examination

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my parents for their continued support and encouragement. Second, I would like to thank the many collaborators, specifically, Roger Beaty, Yoed Kenett, and Hudson Golino, that have influenced the development of my research program. Third, I would like to especially thank Hudson Golino for graciously allowing me to use his workstation to complete the series of simulations performed in this dissertation. Last but not least, I would like to thank Paul Silvia whose mentorship was beyond exception and expectation.

TABLE OF CONTENTS

CHAPTER	Page
I. INTRODUCTION	1
Aims of the Present Research	3
II. REDUNDANCY	5
Detecting Redundancy in Networks	6
Present Research	8
Method	8
Data Generation	8
Psychometric Network Model	10
Redundant Node Approach.....	12
Design	14
Simulating Redundancy	15
Statistical Analyses	16
Results.....	18
False Discovery Rate	18
False Negative Rate	20
Critical Success Index.....	22
Discussion	23
III. DIMENSIONALITY	27
Recent Simulation Studies	28
Present Research	31
Method	32
Data Generation	32
Psychometric Network Models.....	32
Modularity.....	35
Community Detection Algorithms.....	35
Unidimensionality Adjustment	39
Parallel Analysis	40
Design	41
Statistical Analyses	41
Results.....	43
Accuracy and Bias	43
Item Placement.....	48

Best Algorithms	51
Discussion	54
IV. LOADINGS	57
Review of Hallquist, Wright, and Molenaar (2019)	59
Present Research	61
Method	62
Data Generation	62
Psychometric Network Model	62
Network Loadings	63
EFA Loadings	64
CFA Loadings	64
Design	65
Statistical Analyses	65
Results	66
Discussion	68
V. EMPIRICAL EXAMPLE	72
Node Redundancy Strategies	73
Node Redundancy Guidelines	74
SAPA Inventory	77
Results and Discussion	79
Redundancy	79
Dimensionality	82
Loadings	83
Summary	88
VI. CONCLUSIONS	90
REFERENCES	95

CHAPTER I

INTRODUCTION

Network models have become the definitive approach for modeling complexity across the sciences (Barabási, 2012). From mapping the worldwide web (Newman, 2010) to the intricate interactions of the brain (Rubinov & Sporns, 2010), networks have advanced our understanding of complex systems in nearly every domain. Networks are relatively simple, with nodes (circles) representing an element of the system and edges (lines) representing relationships between these elements. One type of network in psychology has been termed *psychometric network models*, which are depicted with nodes representing variables (e.g., psychopathological symptoms) and edges representing the partial correlation between two nodes conditioned on all other nodes (Epskamp & Fried, 2018).

Psychometric network models provide an alternative explanation for the formation of psychological attributes (i.e., properties that exist prior to and independent of measurement; Loevinger, 1957). Traditionally, observable variables that reflect an attribute are thought to co-occur because of an underlying common cause—that is, a latent (unobserved) attribute causes the covariation between observed variables (often referred to as the common cause model; Schmittmann et al., 2013). Network models instead propose that observable variables co-occur because they directly and reciprocally reinforce one another, forming a causally connected system (Borsboom, 2008). Psychological attributes (e.g., personality traits) therefore reference this system of

causally connected components (e.g., observable variables; Cramer, 2012; Schmittmann et al., 2013). This perspective is referred to as the mutualism model (van der Maas et al., 2006).

These two perspectives provide contrasting views on what is being measured and how researchers should measure it. On the one hand, the common cause perspective proposes that observable variables measure an underlying attribute. On the other hand, the mutualism perspective proposes that observable variables do not measure the attribute but are instead part of it (Borsboom, 2008; Schmittmann et al., 2013). The former places the emphasis on measuring the attribute itself, while the latter places the emphasis on measuring the parts of the attribute. The ramifications of these emphases can be significant: Should clinicians treat an underlying psychopathological disorder or the symptoms that constitute the disorder?

These differing emphases also have important implications for the development and validation of assessment instruments (e.g., questionnaires) in psychology. A notable example comes from personality questionnaires, where item content tends to overlap in order to measure a specific attribute (e.g., extraversion). This same attribute from the network perspective is suggested to be comprised of unique causal components, which makes the overlap of item content problematic due to latent confounding (Cramer et al., 2012; Hallquist, Wright, & Molenaar, 2019). Although this may seem to suggest that there is a need to start anew, this is not necessarily the case. It does, however, suggest that the validation of existing and newly developed assessment instruments should be reconsidered.

Recently, my colleagues and I developed a conceptual framework for validating existing and newly developed assessment instruments from the network perspective (Christensen, Golino, & Silvia, under review). In our framework, there is a focus on the identification of the unique components in an instrument and how the psychometric evaluation of the instrument such as dimensionality and item selection should be executed using network models. In the former, the conceptual foundations for a statistical measure to identify redundant items in an instrument was introduced. In the latter, network models were suggested to provide equivalent statistical information as latent variable models (e.g., factors and factor loadings) but were argued to have different substantive interpretations.

Aims of the Present Research

In this paper, my goal is to systematically and empirically investigate our conceptual framework by performing a series of simulation studies, specifically I examine the capacity of network methods and measures to identify redundant items, detect dimensions, and estimate loadings. To achieve this aim, I've organized this dissertation paper into five sections. In Chapter II, I briefly review measurement from the network perspective and discuss the importance of identifying unique components in networks. The simulation study in this chapter focuses on the evaluation of two novel approaches that can be used for detecting redundant nodes in networks. In Chapter III, I review the substantive meaning of dimensions from the network perspective and discuss current methods of estimating dimensions in networks. The simulation study in this chapter evaluates several community detection algorithms that are used to identify

dimensions in networks. In Chapter IV, I review a recent set of simulation studies that demonstrate that the network measure *node strength* (i.e., the sum of connections to a node) is roughly redundant with confirmatory factor analysis (CFA) loadings. The simulation in this chapter evaluates a novel formulation of so-called *network loadings*, which are derived to be roughly equivalent to factor loadings. In Chapter V, I apply these network measures and methods to an empirical example to demonstrate the application of our conceptual framework in a real-world personality data. Results of each section are presented and discussed in turn. In Chapter VI, I conclude with the general implications of these studies.

CHAPTER II

REDUNDANCY

Psychometric network models propose that attributes arise not because of a common cause but instead from the mutual interactions between observed variables. This implies that some attributes, such as personality traits, do not exist—or at least they do not exist in a classical sense of measurement (i.e., causing variation in observable variables; Cramer, 2012). Instead, the relationship between a personality trait and an assessment instrument (e.g., questionnaire) is a mereological one: items in a questionnaire do not measure the trait but are part of it (Borsboom, 2008; Cramer et al., 2012).

This suggests that a personality trait is a summary statistic for how components of a trait's network are influenced by one another: the components *liking to talk to people*, *liking to go to parties*, and *liking to meet new people* of extraversion are causally coupled such that liking to talk to people may lead a person to go to more parties and meet new people (Cramer, 2012). In this sense, extraversion is the *state* of the network or the stable organization of dynamic components that are mutually reinforcing one another (Cramer et al., 2012; Schmittmann et al., 2013). The network thus represents a system of causally connected components that we refer to as extraversion.

What then are the components of networks? Sticking with the personality example, components are defined as “every feeling, thought, or act” that is associated with a “unique causal system” (Cramer et al., 2012, p. 415). More generally, components

refer to causally distinct parts of the system that are not exchangeable with any other part of the system. A key part of this definition is that these components are *unique* in that they are causally autonomous (i.e., distinct causal processes). A recent set of simulation studies corroborated this point by demonstrating that network measures are affected by latent confounding (e.g., similar item phrasings, underlying common causes; Hallquist et al., 2019). Therefore, there is a need to identify unique components in networks to (a) align with the theoretical perspective of a causal system and (b) ensure the accurate interpretation of network measures.

Detecting Redundancy in Networks

Identifying unique components of the system is thus the first step of assessment from the network perspective. This step holds for whether the assessment instrument already exists or is being developed. To do this, identifying components that are redundant and handling that redundancy (e.g., removing all but one component or merging components) is necessary. My colleagues and I have proposed two approaches to identify redundancy: one from the network perspective and the other from the traditional psychometrics perspective (Christensen et al., under review).

The network approach uses the network measure called *weighted topological overlap* (Zhang & Horvath, 2005). The weighted topological overlap measure quantifies the extent to which two nodes share the same connections and similar weights in those connections. Such a measure has been useful for identifying genes or proteins that share similar biological pathways or functions (Nowick, Gernat, Almaas, & Stubbs, 2009). In this sense, greater topological overlap suggests that two genes may belong to the same

functional class relative to other genes. In the context of a psychological network, greater topological overlap would suggest that two observed variables have similar processes or an underlying common cause.

The traditional approach can be derived from more traditional psychometrics where residual correlations of a factor model can provide inference into which variables have redundant information. A simpler method would be to simply obtain a partial correlation matrix where the relationship between each pair of variables is conditioned over all other variables. This matrix is often referred to as the *precision matrix*. In psychometric networks, it is precisely this matrix that is used to estimate the network with some elements in the matrix being zero.

This makes determining what a “high” partial correlation means more complicated than computing statistical significance because significance is already one of the criteria used in the estimation of the network (i.e., determining which edges should be retained). An alternative that is still based on statistical significance is to obtain the empirical distribution of the non-zero values of the topological overlap or partial correlations. Using the absolute values between each unique pair of nodes, a best fitting distribution can be obtained, and the parameters of the distribution can be used to then determine statistical significance. Importantly, there are large number of parameters that are estimated (e.g., every value in the lower triangle of the partial correlation matrix), so a multiple comparison method should be applied.

Present Research

For the purpose of this study, my goal was to investigate whether these two approaches could effectively detect redundant items in a factor model. Within these approaches, I also wanted to examine several different multiple comparison methods to determine which method was most effective for this purpose. To do so, I derived an algorithm that implemented the approaches I described above. A Monte Carlo simulation was used to determine how well each approach (and their multiple comparison methods) could identify redundant items in a factor model. To evaluate the performance of these methods, I used sensitivity and specificity measures. The focus of these performance measures was on the accurate detection of redundant nodes (true positives and false negatives) and the avoidance of detecting non-redundant nodes (false positives).

Method

Data Generation

The data generation for all population models across all simulations generally followed the same approach (Golino et al., in press), unless otherwise noted. First, the reproduced population correlation matrix was computed:

$$\mathbf{R}_R = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}',$$

where \mathbf{R}_R is the reproduced population correlation matrix, $\mathbf{\Lambda}$ is the k (variables) $\times r$ (factors) factor loading matrix, and $\mathbf{\Phi}$ is the $r \times r$ correlation matrix. The population correlation matrix, \mathbf{R}_P , was then obtained by putting the unities on the diagonal of \mathbf{R}_R . Next, Cholesky decomposition was performed on the correlation matrix such that:

$$R_P = U'U.$$

If the population correlation matrix was not positive definite (i.e., at least one eigenvalue ≤ 0) or any single item's communality was greater than 0.90, then Λ was re-generated and the same procedure was followed until these criteria are met. Finally, the sample data matrix of continuous variables was computed:

$$X = ZU,$$

where Z is a matrix of random multivariate normal data with rows equal to the sample size and columns equal to the number of variables.

To generate polytomous data, each continuous variable was categorized with a random skew ranging from -2 to 2 on a 0.5 interval from a random uniform distribution (Table 1).

Table 1. Skew Values for Polytomous Data

Boundaries	Skew								
	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2
1	-1.77	-1.62	-1.45	-1.16	-1.80	-0.34	0.05	0.41	0.68
2	-1.34	-1.16	-0.94	-0.63	-0.60	0.16	0.51	0.78	1.00
3	-1.00	-0.78	-0.51	-0.16	0.60	0.63	0.94	1.16	1.34
4	-0.68	-0.41	-0.05	0.34	1.80	1.16	1.16	1.62	1.77

To provide an example: if a continuous variable had a skew of -1.5, then the value ranges from the second skew column would be used to categorize its values, specifically values less than the first boundary in the column (i.e., -1.62) would be categorized as 1. Values greater than or equal to the first boundary in the column and less than the second boundary in the column (i.e., -1.16) would be categorized as 2. Categorization continues down the skew column until the last boundary where values greater than or equal to the last boundary (i.e., -0.41) would be categorized as 5.

It's important to note that factor models were used to generate data for network models. Recent research has pointed out that despite different hypothesized data generating mechanisms (i.e., factors causing covariation between items vs. direct causal relations between items) these models can be shown to be statistically equivalent (Epskamp et al., 2018a; Fried, 2020; Marsman et al., 2018; van der Maas, 2006). These equivalences extend into the first (means) and second (variance-covariance matrix) moments, which means that any covariance matrix can be represented as a latent variable and network model (van Bork et al., 2019). Therefore, simulating data from a factor model does not inhibit the effectiveness of network models.

Psychometric Network Model

The Gaussian Graphical Model (GGM; Lauritzen, 1996) was used as the psychometric network model. The GGM is a network model where nodes represent variables and edges represent the partial correlation between two nodes given all other nodes in the network. The graphical least absolute shrinkage and selection operator (GLASSO; Friedman, Hastie, & Tibshirani, 2008) has been the most commonly applied

GGM network estimation method in the network psychometrics literature (Epskamp, Waldrop, Möttus, & Borsboom, 2018b). The least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996) of the GLASSO is a statistical regularization technique that reduces parameter estimates, with some estimates becoming exactly zero (for the mathematical notation, see Epskamp & Fried, 2018). The aim of this technique is to achieve a *sparse* model—non-relevant edges are removed from the model, leaving only a subset of relevant (not necessarily significant) edges.

This sparsity is controlled by a parameter called *lambda* (λ). Lower values of λ remove fewer edges, increasing the possibility of including spurious associations, while larger values of λ remove more edges, increasing the possibility of removing relevant edges. When $\lambda = 0$, then the estimates are equal to the ordinary least squares solution (i.e., the partial correlation matrix). This parameter is thus an important part of model selection, striking a balance between sensitivity (i.e., selecting relevant edges that are *truly* relevant) and specificity (i.e., removing edges that are *truly* not relevant).

The popular approach in the network psychometrics literature is to compute models across several values of λ (usually 100) and to select the model that minimizes the extended Bayesian information criterion (EBIC; Chen & Chen, 2008; Epskamp & Fried, 2018). The EBIC model selection uses a hyperparameter (γ) to control how much it prefers simpler models (i.e., models with fewer edges; Foygel & Drton, 2010). Larger γ values lead to simpler models, while smaller γ values lead to denser models. When $\gamma = 0$, the EBIC is equal to the Bayesian information criterion. In the psychometric network literature, this approach has been termed *EBICglasso* and is applied via the *qgraph*

package (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012) in R (R Core Team, 2020). For continuous data, Pearson's correlations were computed; for polytomous data, polychoric correlations were computed.

Redundant Node Approaches

To evaluate whether nodes are redundant, I've developed two novel approaches that were described in the Introduction section of this chapter. The first approach uses what's called *weighted topological overlap* (wTO), which uses the network's structure to determine how much the shared (and not shared) connections of two nodes "overlap" with respect to weight, signs, and quantity (Zhang & Horvath, 2005). The second approach simply uses the absolute values of the partial correlation matrix. The former approach is specifically designed for network models, while the latter is a more general form. Both approaches produce a symmetric matrix where the elements are weights (either topological similarity or partial correlation) between two nodes.

The general strategy for both approaches uses the lower triangle of the symmetric weight matrix to avoid redundant values (i.e., weights are counted only once). The absolute values of the lower triangle are obtained and values equal to zero are removed. The largest values that remain are likely redundant; however, a statistical criterion is necessary. It's important to note that these values imply that two nodes, rather than a single node, are redundant with each other.

To derive statistical significance, a normal and gamma distribution are fit to the distribution of the weights using the *fitdistrplus* package (Delignette-Muller & Dutang, 2015) in R. These two distributions were chosen because they can be efficiently

estimated with maximum likelihood and reflect the distributions that were most often found in several datasets I tested. The *fitdist* function of the *fitdistrplus* package outputs Akaike information criterion (AIC), which is used to determine which distribution has the lowest (best fitting) AIC value. The parameters of the best fitting distribution—mean and standard deviation for normal, and rate and shape for gamma—are then derived using the *MASS* package (Venables & Ripley, 2002) in R. *p*-values for each weight are obtained using this empirical distribution. Because there are typically a substantial number of comparisons being made, I tested several multiple comparison correction methods: standard alpha ($\alpha = .05$), Bonferroni correction, false-discovery rate (FDR; Benjamini & Hochberg, 1995), and adaptive alpha (α_{adapt} ; Pérez & Pericchi, 2014).

The standard alpha simply selects all weights that have a *p*-value less than .05. The Bonferroni correction (also known as the familywise error rate) is the standard alpha value divided by the number of comparisons (e.g., total number of weights). The FDR controls the false positive rate of significance tests by using the expected number of false positive results (e.g., 5% with an $\alpha = .05$) to adjust for the total number of significant results. The number of false positives is controlled by a *q*-value, which is can be set with a slightly more liberal value (e.g., $q = .10$). The *q*-value suggests that rather than 10% of all tests resulting in false positives, 10% of all *significant* results will be false positives. Finally, the adaptive alpha adjusts the standard alpha level by accounting for a reference sample size. It's well-known that as sample size increases, the likelihood of a small effect becoming significant also increases.

To account for this, Pérez and Pericchi (2014) provide the following formula:

$$\alpha_{adapt} = \frac{\alpha * \sqrt{n_0 \times (\log(n_0) + \chi^2_{\alpha}(1))}}{\sqrt{n^* \times (\log(n^*) + \chi^2_{\alpha}(1))}}, \quad (11)$$

where n_0 is the reference sample size, n^* is the actual sample size, and α is the standard alpha. The reference sample size can be computed using a power analysis. For my purposes, this power analysis was computed using the *pwr* package (Champely, 2018) in R for a correlation with a medium effect size, alpha of .05, and power of .80. This yields a reference sample size of 84.07. The actual sample size will be the number of weights used in the distribution. Both approaches were applied using the `node . redundant` function in the *EGAnet* package (Golino & Christensen, 2020) in R.

Baseline comparison. To provide a baseline comparison method, I used a threshold of partial correlations where if a connection between two nodes was greater than .20, then the nodes were considered redundant. This threshold serves as a benchmark independent of statistical significance, which may sometimes produce false positives because there can always be points on a distribution in which values are considered significant.

Design

The population models were simulated from a multidimensional multivariate normal distribution. Across population models, factor loadings for each item were randomly drawn from values between .40 and .70 to mimic more realistic data conditions. Similarly, cross-loadings were generated following a random normal distribution with a mean of zero and a standard deviation of .10. This procedure follows previous simulation work described in Garcia-Garzón, Abad, and Garrido (2019). These cross-loadings

represent data conditions that are more likely to be found in real-world data (Bollmann, Henne, Küchenhoff, & Bühner, 2015).

Two and four factors were simulated to provide multidimensional structures that are commonly found in the psychological literature (Henson & Roberts, 2006). There were six, twelve, and eighteen variables per factor, which were chosen to evenly split the number of variables for the percentage of redundant items. These percentages were 0%, 16.7%, 33.3%, and 50%. The condition of zero redundant items is particularly important for estimating the consistency for which methods identify false positive redundancy (i.e., redundancy when there is none). Correlations between factors were manipulated to be orthogonal (.00), small (.30), moderate (.50), and large (.70). Finally, very small (250), small (500), medium (1000), and large (5000) samples sizes were generated.

The simulation design of the current study allowed for a mixed factorial design: $2 \times 3 \times 4 \times 4 \times 4 \times 2$ (number of factors \times variables per factor \times percentage of redundant items \times correlations between factors \times sample size \times number of responses) for a total of 768 simulated condition combinations.

Simulating Redundancy

To simulate redundancy, the following approach was used. First, the number of redundant items per factor was manipulated a priori (i.e., percentage of redundant items). Second, from each factor, a subset of items (equivalent to the percentage of redundant items; e.g., $33.3\% \times 18 \text{ items} = 6 \text{ items}$) was randomly sampled *without* replacement. This subset is referred to as the *replace* set. Third, excluding the subset of items already

selected in the replace set, another subset of items within the same factor were randomly sampled *with* replacement. This subset is referred to as the *copy* set.

From the copy set, 20% of the values in each item were copied to “replace” the corresponding values in the replace set. Because direct copies of values would introduce perfect collinearity, random noise was added to the values that reduced this effect. This random noise, on average, added or subtracted one standard deviation from the copied value. This strategy generally led to larger increases in correlation between items that started out with smaller correlations—that is, smaller correlations had a greater increase in magnitude than larger correlations. It’s worth noting that because the copy set was sampled *with* replacement, it was possible for items to be redundant with more than one item. The same values, however, were not used to avoid increasing the number of redundant items beyond the intended manipulation.

Statistical Analyses

To evaluate the performance of the two redundancy approaches, four types of alpha, and a threshold method, I used sensitivity and specificity measures (Table 2).

Table 2. Sensitivity and Specificity

		Estimated	
		Redundant	Not Redundant
Population	Redundant	True Positive (TP)	False Negative (FN)
	Not Redundant	False Positive (FP)	True Negative (TN)

More specifically, I used false discovery rate ($\frac{FP}{(TP+FP)}$), false negative rate ($\frac{FN}{(FN+TP)}$), and critical success index ($\frac{TP}{(TP+FP+FN)}$). Given that there were a large number of true negatives (i.e., nodes that are not redundant that are identified as not redundant), measures were chosen that did not include them.

False discovery rate was used to determine the number of incorrectly estimated redundant items versus the total number of the estimated redundant items. This measure represents an approach's (or type of alpha's) tendency to over-identify redundant items relative to the number of actual redundant items. False negative rate was used to determine the number of type II errors or the number of items that were estimated as not redundant when they were redundant versus the total number of the estimated redundant items. This measure represents an approach's (or type of alpha's) tendency to under-identify redundant items relative to the number of actual redundant nodes. Critical success index was used as an overall accuracy measure, giving an equal weight to true positives as false positives and negatives. This measure represents an approach's (or type

of alpha's) tendency to correctly identify redundant nodes, with few false positives and false negatives.

Results

For the presentation of the results, I focused on breaking down the FDR, false negative rate (FNR), and critical success index (CSI) by number of responses (continuous \times polytomous), percentage of item redundancy (per factor), and sample size. The percentage of item redundancy was the most critical factor in each approach's and alpha type's performance across sensitivity and specificity measures. The next most important factor was sample size, which is the most obvious factor when analyzing the data (i.e., a researcher may not know how many items per factor are redundant). The other conditions (number of factors, variables per factor, and correlation between factors) were not substantial contributors to variability in performance and therefore are not discussed.

False Discovery Rate

For the general trends, the number of responses did appear to have an effect on FDR, with both approaches and most alpha types having better performance when the data were continuous (except for weighted topological overlap when $n = 250$; Figure 1). This is somewhat expected as the continuous number of responses has greater variability in the responses, which leads to better differentiation of whether items are redundant. In contrast, when the continuous data were categorized, the variability between items is reduced and collapsed into bins, which allows for greater redundancy to appear when there may not be. Similar to the number of responses, the FDR across all approaches and alpha types decreased as the sample size and number of redundant items increased.

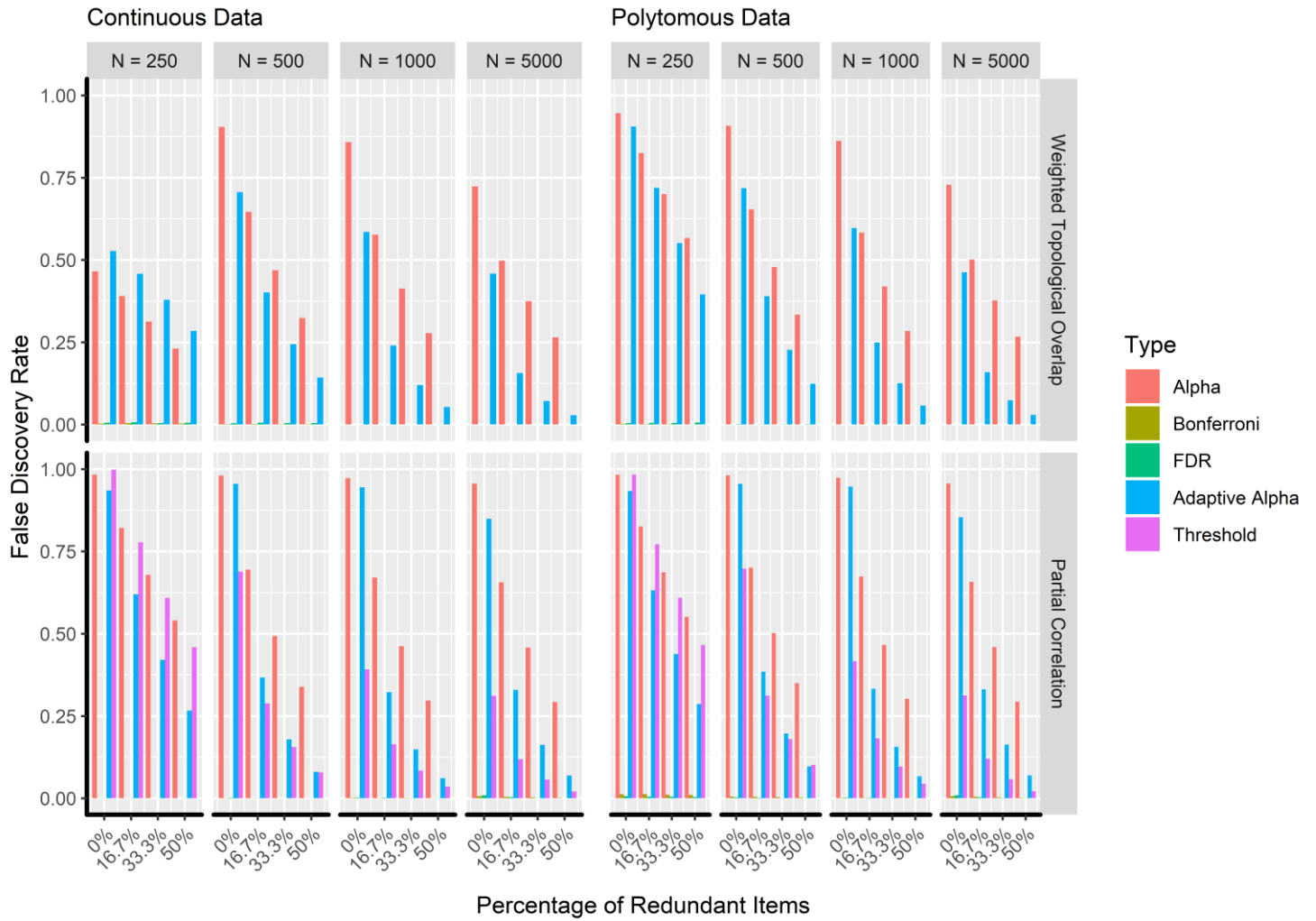


Figure 1. False discovery rate broken down by number of responses, percentage of redundant nodes, and sample size.

As far as the best performing methods, the Bonferroni and false discovery rate (multiple comparisons method) had very few false positives relative to their total positives. This result is relatively misleading, however, because these alpha types generally did not identify redundant nodes across conditions. This can be seen in the breakdown of the false negative rate (FNR) results (Figure 2) where both of these alpha types had FNRs near 1, suggesting that they were consistently not detecting *any*

redundant items (regardless of approach). Because these methods performed so poorly (see Figure 3), they won't be discussed.

This turns the attention to the other three alpha types: standard, adaptive, and threshold. In general, the adaptive alpha had the lowest FDR across approaches and conditions. Adaptive alpha for the weighted topological overlap approach appeared to fare better when there were fewer redundant items (i.e., 0% and 16.7%) and was comparable to the partial correlation approach when there were more redundant items (i.e., 33.3% and 50%). Notably, the threshold method performed comparably to all other approaches and alpha type combinations when the sample size was small ($n = 500$), and outperformed them when sample size was moderate ($n = 1,000$) or large ($n = 5,000$).

False Negative Rate

Similar to the FDR results, the FNR decreased as the sample size increased (Figure 2). In contrast to the FDR results, the percentage of redundant items did not appear to affect the FNR values (except for weighted topological overlap when $n = 250$). For the approaches, the partial correlation approach generally had fewer false negatives than the weighted topological overlap approach. It's worth noting that across all approaches and alpha types that there were FNR values of 0 for the 0% redundant items condition. This is because there were no redundant items in the population and therefore could be no false negatives.

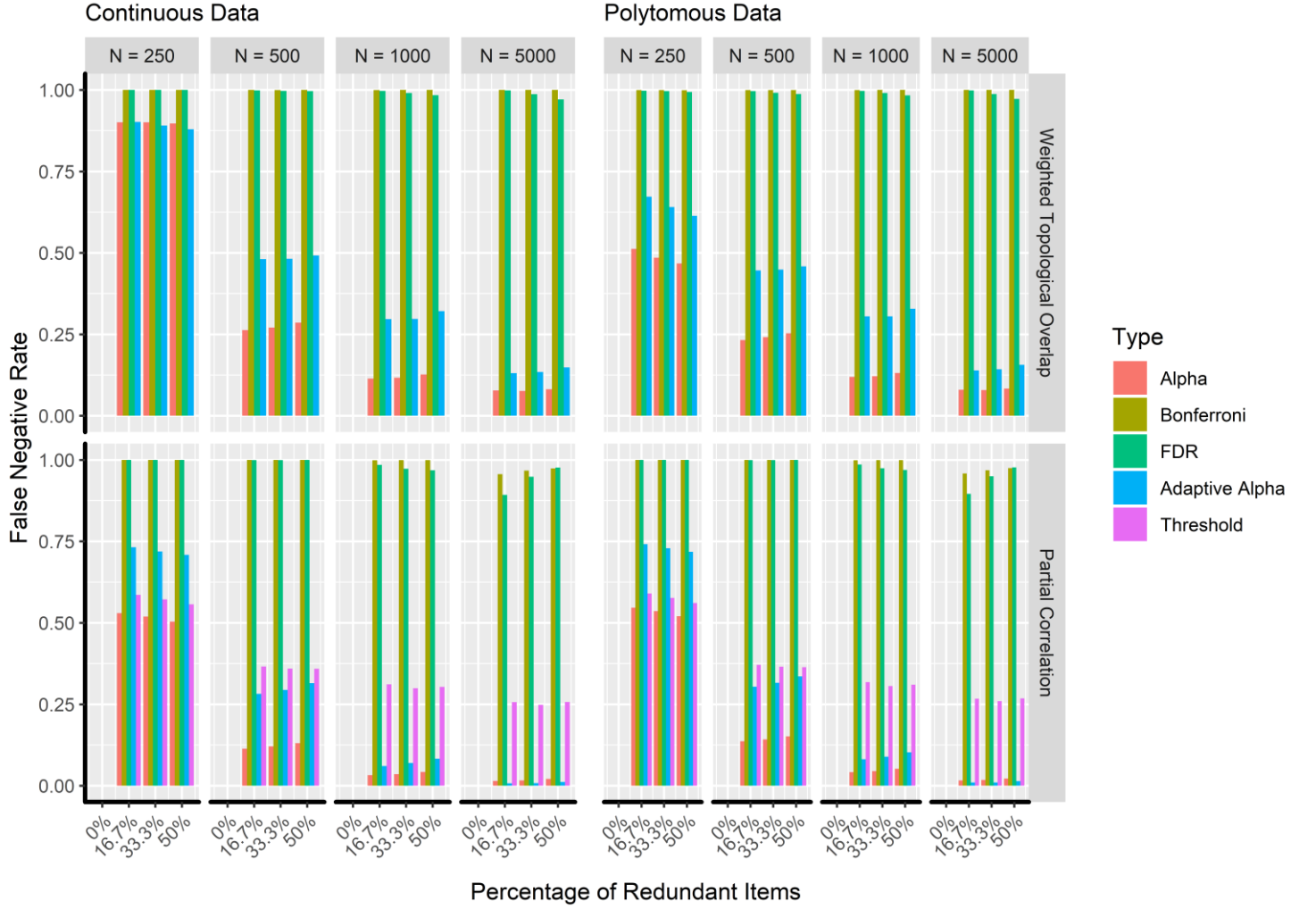


Figure 2. False negative rate broken down by number of responses, percentage of redundant nodes, and sample size.

When looking across the alpha types, the standard alpha generally had the fewest false negatives, suggesting it was more likely to discover most redundant nodes. This contrasts with the FDR where it had the most false positives. Taken in combination, this result suggests that the standard alpha may not have discriminated which items were redundant very well. The adaptive alpha, however, tended to have lower FNR values than the threshold, which usually held across the number of responses (except when $n = 250$).

Interestingly, these results are the reverse of the FDR results, specifically the threshold method has a higher FDR and lower FNR when the sample size is very small, while adaptive alpha has a higher FDR and lower FNR when the sample size is small, moderate, and large.

Critical Success Index

As the overall metric for accuracy, the CSI reflects the combination of the FDR and FNR where their minimization leads to the most optimal outcome. For the general trends, CSI increased as sample size and percentage of redundant items increased. The number of responses did not seem to affect the CSI values for either approach. Across these conditions, the partial correlation approach had higher CSI values than the weighted topological overlap approach for each respective alpha type (Figure 3).

For the alpha types, the adaptive alpha (particularly for the partial correlation approach) had the largest CSI across conditions and the difference from other alpha types increased as the sample size and percentage of redundant items increased. The standard alpha and threshold method had comparable CSI when there were 50% of items that were redundant; otherwise, the threshold method had the second largest values of CSI.

Similar to the FNR, there was no measure of CSI for the 0% redundant items because it was not possible to have true positives. Therefore, the best marker of performance for 0% redundant items is the FDR metric (Figure 1). Here, the threshold method had the best performance when the sample size was very small; otherwise, the adaptive alpha should be preferred (weighted topological overlap approach for continuous data and partial correlation approach for polytomous data).

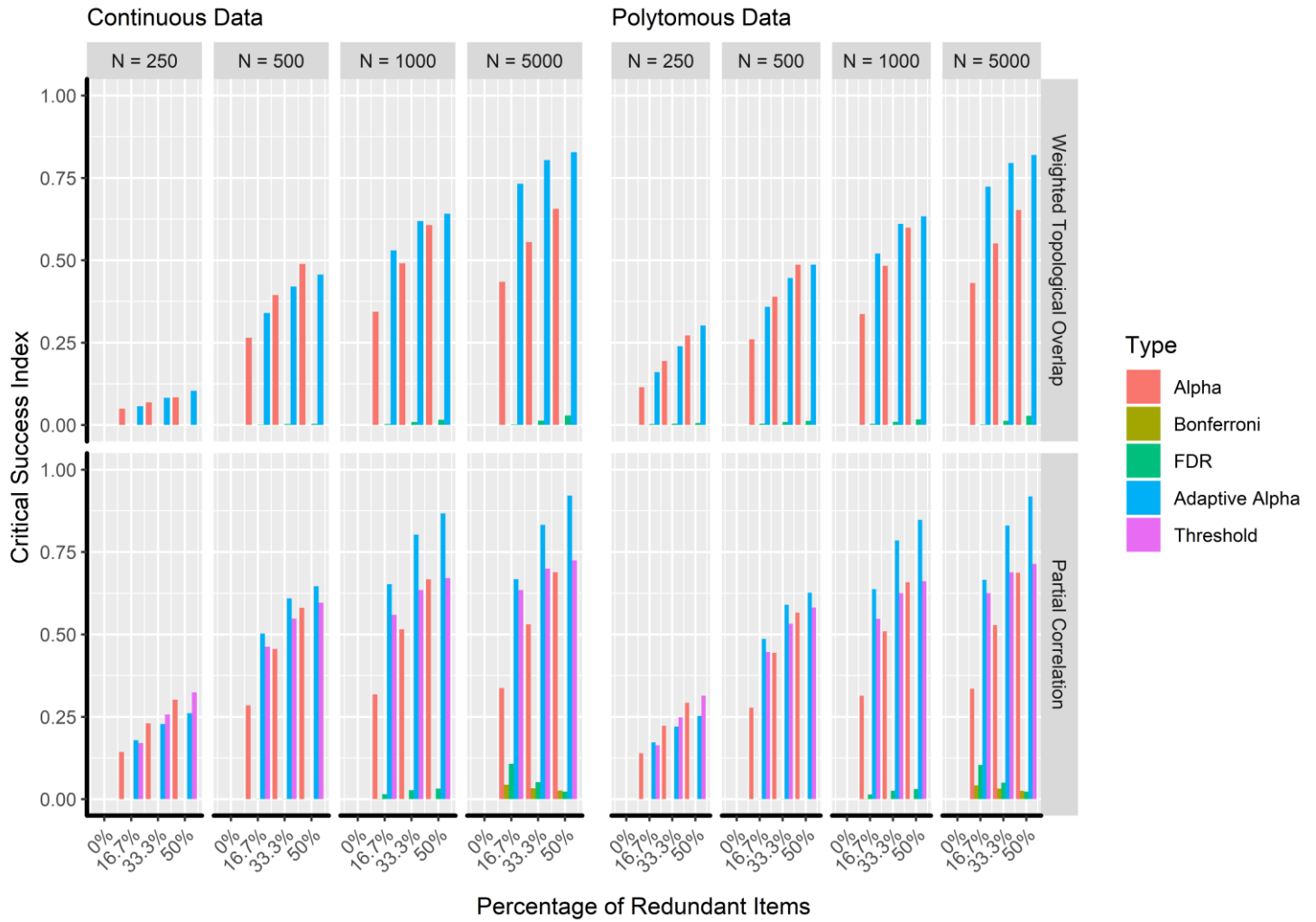


Figure 3. Critical success index broken down by number of responses, percentage of redundant nodes, and sample size.

Discussion

This study evaluated the effectiveness of two approaches to estimating statistical redundancies among items in an assessment instrument. Across the conditions tested, both approaches appeared to be effective but were limited in their effectiveness by the type of alpha used, specifically standard alpha and adaptive alpha performed the best of the alpha types with Bonferroni and FDR multiple comparison corrections being too

stringent to detect any redundancies when they were present. In the end, only the adaptive alpha method surpassed the baseline comparison of the threshold method for achieving better rates of false positives (lower), false negatives (lower), and overall accuracy (higher).

When taking in the results as a whole, there seemed to be a trade-off between detecting all of the redundant items (i.e., avoiding false negatives) and detecting only the redundant items (i.e., avoiding false positives). A primary example was the standard alpha, which had consistently high false positives and low false negatives relative to the adaptive alpha and threshold methods. When considering this trade-off, greater emphasis should be placed on avoiding false negatives rather than false positives. This emphasis is because this procedure serves as a statistical basis for human judgment of whether items are from a theoretical common cause. Therefore, it's better to err on the side of detecting too much redundancy rather than too little because researchers will have the definitive decision for whether two or more items are theoretically redundant. Nonetheless, an optimal approach would strike a balance between the two.

The adaptive alpha and threshold methods struck the best balance between false positives and negatives under certain conditions. When sample size was very small ($n = 250$), the threshold method had the best CSI. It's notable, however, that no combination of approach and alpha type fared well in this condition. This suggests that this redundancy analysis should be avoided when sample sizes are very small. When sample size is small ($n = 500$), moderate ($n = 1,000$), or large ($n = 5,000$), the adaptive alpha method struck the best balance between false positives and negatives. This was

particularly true for the partial correlation approach. In general, the partial correlation approach appeared to outperform the weighted topological approach in most conditions, but these differences were relatively small. The choice between either approach will likely come down to the conditions expected in the data such as the expected number of redundant items per factor.

When evaluating these results, there was a significant limitation to acknowledge: the factor loadings varied randomly while the redundant items were also selected at random. This means that an item with a high loading may have been made more redundant with an item that had a low loading (and vice versa) rather than making already similar items more similar (i.e., high loading items more redundant with high loading items). As a consequence, some of the redundant items may have been harder to detect because although they were becoming more similar, they were perhaps not as similar as some high loading items on the same factor that were not manipulated to be redundant. This limitation may mean that the results of this study are more conservative estimates of the effectiveness of these approaches and alpha types.

To my knowledge, this is the first simulation to attempt to statistically detect redundancy in assessment instrument conditions. This simulation therefore serves as a starting point more than a definitive conclusion. There is clearly room for improvement, such as the strategy for generating redundant items. Future research, for example, may consider generating redundancy by introducing minor factors that have large loadings within major factors. This choice would better reflect more common scale development practices and likely lead to more robust results. As for the approaches implemented here,

the adaptive alpha appears to be the decisive go-to method for the best results. The adaptive alpha was consistently better than the baseline of using a threshold (both approaches) and demonstrated the lowest false discovery rate when there was no redundancy between items (weighted topological overlap). In sum, the statistical detection of redundancy in assessment instruments seems feasible and will be a useful tool for deriving concise assessment instruments when paired with a researcher's theoretical knowledge.

CHAPTER III

DIMENSIONALITY

The next step in our psychometric network assessment framework is to identify dimensions. Dimension identification in assessment is a critical part of validating an instrument. Traditional psychometric approaches apply factor analytic techniques such as exploratory factor analysis (EFA) to assess the dimensionality of an instrument (Flora & Flake, 2017). Factor analytic methods typically correspond to common cause models where items are regressed on the factors (Borsboom et al., 2003). From the common cause perspective, dimensions represent evidence of an underlying cause of a set of variables.

From the network perspective, dimensions emerge from densely causally connected sets of nodes and represent a coherent sub-network (i.e., smaller network) within the overall network (Christensen, Golino, & Silvia, under review). For network models, community detection algorithms are the commonly applied to identify dimensions (Fortunato, 2010). These algorithms typically identify the number of *communities* (or dimensions) in the network by maximizing a function called *modularity*, which quantifies the extent to which a set of nodes has a higher number of connections within its group than what is expected at random (Newman, 2006; Newman & Girvan, 2004).

Although the hypothesized data generating mechanisms behind these perspectives differ, they are based on the same data structure (van Bork et al., 2019). Indeed, a

researcher can fit a factor model to a data structure generated from a network model with good model fit (van der Maas et al., 2006). Similarly, a network model with a community detection algorithm can be fit to a data structure generated from a factor model and identify factors (Fried, 2020; Golino & Epskamp, 2017). This underlying equivalence follows from the fact that any covariance matrix can be represented as a latent variable or network model (van Bork et al., 2019). Therefore, factors of a latent variable model and communities of a network model are statistically equivalent (Golino & Epskamp, 2017) and the difference is purely the hypothesized data generating mechanism (Fried, 2020).

Recent Simulation Studies

The most extensive work on dimensionality in the psychometric network literature has been with a technique called Exploratory Graph Analysis (EGA; Golino & Epskamp, 2017; Golino et al., in press). The EGA algorithm works by first estimating a Gaussian Graphical Model (Lauritzen, 1996) using the graphical least absolute shrinkage and selection operator (GLASSO; Friedman, Hastie, & Tibshirani, 2008). Edges in the GGM represent (regularized) partial correlations between nodes after conditioning on all other nodes in the network. After network estimation, EGA applies the Walktrap community detection algorithm (Pons & Latapy, 2006), which uses random walks to determine the number and content of communities in the network (discussed in more detail in the Method section of this chapter). Several simulation studies have shown that EGA has comparable or better accuracy when identifying the number of population dimensions than the most accurate factor analytic techniques (e.g., parallel analysis; Golino & Demetriou, 2017; Golino & Epskamp, 2017; Golino et al., in press).

Despite the effectiveness of EGA, there has been only one investigation, to my knowledge, into the effect of different network estimation methods and no investigations into the effect of different community detection algorithms. To date, the GLASSO has been the standard network estimation method applied across psychological network studies (Epskamp & Fried, 2018). Notably, there are other network estimation methods, each of which will estimate a different network structure, which ultimately affects the dimensionality estimate. One simulation study compared the dimension identification accuracy of the GLASSO and triangulated maximally filtered graph (TMFG; Massara, Di Matteo, & Aste, 2017) network estimation methods using the Walktrap community detection algorithm (Golino et al., in press). This study found that the GLASSO network estimation method had better accuracy and less bias than the TMFG but both performed comparable to the best factor analytic techniques.

More recently, non-regularized network estimation methods have been put forward in the literature (Williams, Rhemtulla, Wysocki, & Rast, 2019). These methods have been shown to have better performance when estimating the population network structure of dense (highly connected) networks, which are common in psychology (Williams & Rast, 2019). Despite their better performance when estimating the population network structure, there has yet to be an investigation in whether they perform better for estimating dimensions in networks.

Similarly, the Walktrap community algorithm has not been evaluated in the context of other community detection algorithms. Several other algorithms such as the Spinglass algorithm (Reichardt & Bornholdt, 2006) have been used in the psychometric

network literature (e.g., De Beurs et al., 2019). Despite their application, there has yet to be an investigation that compares these algorithms in a psychological network context. In general, most community detection algorithms were developed and validated on networks containing a large number of nodes (e.g., > 1,000; Lancichinetti & Fortunato, 2009; Yang, Algesheimer, & Tessone, 2016). Moreover, these algorithms are often designed to work well for one type of problem or data structure (Gates, Henry, Steinley, & Fair, 2016). Because most psychological networks consist of fewer than 100 nodes, there is a need to verify which of these algorithms work best and under conditions commonly found in the psychological literature.

A recent simulation study systematically examined several freely available community detection algorithms in the context of brain networks (Gates et al., 2016). Brains networks are perhaps the closest comparison to psychological networks in that they are typically represented by correlational (rather than count) data and generally have fewer than 1,000 nodes. In this study, they generated network models using a structural equation modeling method and manipulated several conditions, including number of nodes and communities, size of edge weights (i.e., correlations), and correlations between communities. Of the six algorithms they examined, the Walktrap and Louvain (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) algorithms performed the best across conditions. Importantly, their study investigated conditions where there were a small number of nodes (i.e., 25 and 75).

Present Research

The goal of this simulation study was twofold: compare the effects of (1) network estimation methods and (2) community detection algorithms on the accuracy of dimension identification in psychological network models. For the network estimation methods, I used the standard network estimation method in the psychometric network literature, the GLASSO, and compared its accuracy to two non-regularized partial correlation methods that are based on neighborhood selection (Williams et al., 2019). These two approaches solely differ in their criterion for model selection: Bayesian information criterion (BIC) and Akaike information criterion (AIC). For the community detection algorithms, I examined several freely available algorithms that were used in Gates et al.'s (2016) simulation study and included a few others that were available in the R package *igraph* (Csárdi & Nepusz, 2006).

My simulation study differs from previous studies that have compared these network estimation methods and community detection algorithms in a few ways. First, the data in this study were generated from a factor model rather than being generated from an empirical dataset or network model (Gates et al., 2016; Williams et al., 2019). As already discussed, factors and communities of factor and network models (respectively) are statistically equivalent and only differ in their substantive interpretations. Therefore, generating data from a factor model is advantageous because it allows me to vary conditions which are familiar to many researchers in psychology (e.g., factor loadings, number of variables per factor, correlations between factors).

Second, this study specifically analyzes the accuracy of dimension identification rather than whether the true network structure is identified (i.e., correct number of edges;

Williams et al., 2019). It's plausible that the true network structure may contain many edges that are not relevant for detecting dimensions, which may reduce the efficacy of contemporary community detection algorithms that are based on sparser network structures. Therefore, network estimation methods may differ in their utility (e.g., correct estimation of the true network structure vs. correct estimation of dimensions). Finally, the present simulation generates data that aligns with conditions more commonly found in psychological networks: specifically, multivariate data with a relatively low number of dimensions (e.g., 1, 2, and 4) and variables per dimension (e.g., 4, 8, and 12). With these conditions, the number of nodes in the network range from 4 to 48, which is considerably smaller than networks observed in brain data (Gates et al., 2016).

Method

Data Generation

The data generation approach followed the same approach applied in Chapter II's Method.

Psychometric Network Models

Similarly, the same GLASSO network estimation method used in Chapter II's Method was used in this study. The GLASSO was applied using the network estimation criteria found in the Exploratory Graph Analysis approach (Golino et al., in press). First, the minimum λ value is set to .01, which is slightly higher than the default of .001. This is selected to reduce the possibility of false positive edges in the network. Second, the γ value is set to .50, which is the default; however, it is iteratively decreased by .25, until reaching zero, based on whether any one node in the network is disconnected. If γ

reaches zero, then the network is used regardless of whether any nodes are disconnected. Finally, a node that forms its own community is not included in as a part of the number of dimensions identified (Golino et al., in press). This removes variables that are not identified to be a part of any dimension in the network.

Non-regularized partial correlation networks. In addition, two non-regularized partial correlation estimation methods were used. Both methods were based on a regression strategy called *neighborhood selection*, which uses node-wise multiple regression on each node in the network (Williams et al., 2019). Multiple regression coefficients have direct correspondence to the inverse covariance coefficients in that the negative regression coefficient ($-\beta_{ij}$) divided by the predictor variable’s variance (σ_j^2) is equal to the inverse covariance between the regressed variable and the predictor variable given all other variables (θ_{ij}).

The multiple regression coefficients for each regressed variable are placed across the row of each target variable with the regressed variable’s variance in its respective element’s position (θ_{ii}^2 ; i.e., variance of each variable is on the diagonal). A common method for computing partial correlations is to take the square root of the product of the corresponding regression coefficients in the matrix and replacing their signs (i.e., $\rho_{ij} = \text{sign}(\beta_{ij})\sqrt{\beta_{ij}}$, $i \neq j$). Notably, this leads to an asymmetric covariance matrix where coefficients do not correspond to their respective transpose element (i.e., $\theta_{ij}^2 \neq \theta_{ji}^2$).

There are two approaches for determining whether an edge should be non-zero: the “and-rule” where both β_{ij} and β_{ji} must be non-zero and the “or-rule” where only one

coefficient must be non-zero. Both approaches use a forward search strategy for determining non-zero coefficients, which removes predictor variables from each multiple regression that minimize some criterion until the minimum value of the criterion is achieved for the set of predictor variables. The coefficients that are not removed in the process of minimizing the criterion are retained in the network as non-zero edges, while the removed variables are set to zero.

This criterion is based on traditional model selection criteria AIC and BIC. The main difference between these criteria is that the BIC tends to penalize more complex models more severely than the AIC. In short, the AIC is better in conditions when a false negative is considered to be worse than a false positive, while BIC is better in conditions when a false positive is considered to be worse than a false negative.

For this study, I examined both the AIC and BIC approaches to edge selection because they were shown to have considerable differences in estimating the population network structure in previous simulations (Williams et al., 2019). The “and-rule” and “or-rule” had negligible effects on the estimation of population network structures, so I only investigated the “and-rule” in this study (Williams et al., 2019). Both non-regularized partial correlation network models were estimated using the *GGMnonreg* package (Williams, 2019) in R.

Modularity

A key definition for understanding many community detection algorithms is the concept of modularity (Newman, 2006). Modularity can be expressed as (Fan, Li, Zhang, Wu, & Di, 2007):

$$Q = \frac{1}{2w} \sum_{ij} \left(w_{ij} - \frac{w_i w_j}{2w} \right) \delta(c_i, c_j),$$

where w_{ij} is the edge strength for a given node pair, w_i and w_j are the node strength for node i and node j (respectively), w is the sum of all the edge weights in the network, c_i and c_j represents the community that node i and node j belong to, and δ is 1 if the nodes belong to the same community (i.e., $c_i = c_j$) and 0 if otherwise. Essentially, modularity reflects the extent to which communities have more connections within the community and fewer connections with other communities.

Community Detection Algorithms

This study focused on eight different community detection algorithms that are freely available via the R package *igraph*. These include the Walktrap (Pons & Latapy, 2006), Infomap (Rosvall & Bergstrom, 2008), Fast-greedy (Clauset, Newman, & Moore, 2004), Louvain (Blondel et al., 2008), Leading Eigenvalue (Newman, 2006), Label Propagation (Raghavan, Albert, & Kumara, 2007), Spinglass (Reichardt & Bornholdt, 2006), and Edge Betweenness (Girvan & Newman, 2002) community detection algorithms.

All community detection algorithms were implemented with their default arguments in order to evaluate their baseline performance without researcher direction (similar to Gates et al., 2016). Moreover, all network matrices were input with absolute values to avoid bias of some methods performing better than others because of their

ability to handle negative associations. Below, I briefly describe each algorithm (more detailed information can be found within their respective citations).

Walktrap. The Walktrap algorithm (Pons & Latapy, 2006) has been the most commonly applied algorithm in the psychometric network literature (Golino & Demetriou, 2017; Golino & Epskamp, 2017; Golino et al., in press). The Walktrap algorithm begins by computing a transition matrix where each element represents the probability (based on node strength) of one node traversing to another given a length of time. Using Ward’s agglomerative clustering approach (Ward, 1963), each node starts as its own cluster and merges with adjacent clusters (based on squared distances between each cluster) in a way that minimizes the sum of squared distances between other clusters. Modularity is then used to determine the optimal partition of clusters (i.e., communities).

Infomap. Similar to the Walktrap algorithm, the Infomap algorithm (Rosvall & Bergstrom, 2008) uses random walks. Different from the Walktrap algorithm, Infomap is derived from information theory with idea of “compressing” the conditional information of a random walk on the network into Huffman codes (a binary naming system; Rosvall & Bergstrom, 2008). The major difference between these two algorithms is that Infomap captures the conditional flow of information across the network in a way that maximizes the information (e.g., bits) of the random walk process. The partition function that optimizes this minimization is given by the entropy of movement between communities and the entropy of movement within communities. The space of possible partitions is

explored using a deterministic greedy search algorithm, which is refined using a simulated annealing approach.

Fast-greedy. The Fast-greedy algorithm (Clauset, Newman, & Moore, 2004) uses modularity to identify optimal partitions in the network. Like the Walktrap algorithm, the Fast-greedy algorithm begins with each node considered as its own community and follows a hierarchical clustering algorithm. The algorithm then proceeds by iteratively combining neighboring communities in a greedy way: Each node is moved into a community that maximizes the modularity function. These aggregate communities are then merged until the modularity function can no longer be increased.

Louvain. The Louvain algorithm (also referred to as Multi-level; Blondel et al., 2008) is very similar to the Fast-greedy algorithm in that it uses modularity to optimize its partitions. It differs in that its motivation is to identify hierarchical structures in large networks, specifically it iteratively exchanges nodes between communities and evaluates the change in modularity. The algorithm then further creates smaller networks by creating latent nodes representing a collection of nodes and identifies edge weights with other observed and latent nodes (Gates et al. 2016). In its use in this study, the algorithm was not used to identify hierarchical community structures in the network. Therefore, it's expected that this algorithm will closely align with the Fast-greedy algorithm. It's also important to note that the algorithm implemented in *igraph* is deterministic; however, other variants are not (Gates et al., 2016; Rubinov & Sporns, 2010).

Leading Eigenvalue. The Leading Eigenvalue algorithm (Newman, 2006) is based on spectral properties of the network using eigenvector of the first eigenvalue to

determine optimal community structures. Like Fast-greedy and Louvain algorithms, the Leading Eigenvalue algorithm uses modularity to optimize these structures. The algorithm begins by computing the first eigenvector of the modularity matrix and the network is split into two communities that improves the modularity. This process iteratively unfolds until there is no longer improvement in modularity.

Label Propagation. The Label Propagation algorithm (Raghavan et al., 2007) begins by assigning each node a unique label. Each node then adopts the same label that the majority of its neighbors have, with ties being broken randomly. This continues iteratively until each node has the same label as the majority of its neighbors. The general notion of the algorithm is that a consensus will develop among the nodes in the network. Notably, this algorithm is not deterministic in that it produces different results with each run. In this study, only one run was implemented for each sample in order to evaluate its accuracy in its current form. Other strategies such as repeated sampling could be used to arrive at a relatively stable organization of communities (e.g., median; De Beurs et al., 2019; Lancichinetti & Fortunato, 2012).

Spinglass. The Spinglass algorithm comes from statistical physics and is based on notion that “the problem of community detection can be mapped onto finding the ground state of an infinite ranged Potts spin glass” (Reichardt & Bornholdt, 2006, p. 1540). In essence, edges should connect nodes that are in the same spin state (i.e., community), while nodes in different states should be disconnected, which results in a “lower energy state” or ground state of the system. Similar to the Label Propagation algorithm, this algorithm is not deterministic and only one run was implemented in this study.

Edge Betweenness. The Edge Betweenness algorithm (Girvan & Newman, 2002) was one of the first algorithms used to identify communities in networks. This algorithm finds edges that are frequently “between” other nodes in the network known as *edge betweenness* (based on the betweenness centrality; Freeman, 1977). Edge betweenness is calculated for the entire network and the edge with the highest betweenness value is removed. All edges that are affected by this removal have their edge betweenness value recalculated. This process repeats iteratively until no edges remain.

Unidimensionality Adjustment

A well-known limitation of community detection algorithms is that they tend to favor multidimensional structures (Golino et al., in press). This is a consequence of what most algorithms were designed to do: identify modular components in large networks (i.e., > 1000 nodes). Because this issue lies in many of the community detection algorithms, all psychometric network models were adapted to the unidimensional approach found in Golino et al. (in press).

Their approach works in the following way: generate a random multivariate normal dataset with a certain number of variables (e.g., four) with high factor loadings (e.g., .70) on a single factor and add these variables to the original dataset before computation of the (partial) correlation matrix. Then, compute the network and apply the community detection algorithm. If the algorithm identifies one or two dimensions, then the original data is unidimensional. If more than two dimensions are identified, then the generated variables are removed, and the network and community detection algorithms are reapplied. The conceptual reasoning behind this is that the generated variables

represent a cohesive single factor that is independent of the original data. Therefore, it is known that if there are two factors, then one will be the generated data and the other will be the original data. Based on recommendations by Golino and colleagues (in press), the number of variables generated in the simulated data was set equal to the variables per factor in the data generation conditions.

Parallel Analysis

As a comparison, two parallel analysis (PA) methods—principal axis factoring (PAF) and principal component analysis (PCA)—were used. These two methods were chosen because they have been extensively evaluated in the literature (e.g., Garrido, Abad, & Ponsoda, 2013) and have shown comparable performance with EGA in a previous simulation study (Golino et al., in press). In short, PA generates a larger number of random datasets, with an equivalent number of cases as the original dataset, by resampling (with replacement) from the original dataset (Horn, 1965). The number of factors (PAF) or components (PCA) whose eigenvalues in the original dataset are greater than the mean of the resampled datasets is suggested as the dimensional solution. The number of dimensions were estimated using the minimum residual estimator.

Design

Similar to the Design of Chapter II, the population models were simulated from a multidimensional multivariate normal distribution with factor loadings for each item generated with $\pm .10$ deviance drawn from a uniform distribution. Cross-loadings were

also generated following a random normal distribution with a mean of zero and a standard deviation of .10. The same correlations between factors (.00, .30, .50, and .70) and sample sizes (250, 500, 1000, 5000) that were used in the Chapter II simulation were used in this study.

Different for this study, one, two, and four factors were simulated to provide unidimensional and multidimensional structures that are commonly found in the psychological literature (Henson & Roberts, 2006). There were four, eight, and twelve variables per factor, which represented conditions common in scale development and validation. Finally, factor loadings were manipulated to be small (.40), moderate (.55), and large (.70).

The simulation design of the current study allowed for a mixed factorial design: $4 \times 4 \times 3 \times 3 \times 3 \times 2$ (factor correlations \times sample size \times number of factors \times number of variables \times factor loadings \times number of responses) for a total of 864 simulated condition combinations.

Statistical Analyses

To evaluate the performance of the network and parallel analysis approaches, overall accuracy and bias were measured using the percentage of correct number of factors (PC), mean bias error (MBE; the average deviation away from the correct number of factors) and mean absolute error (MAE; the average absolute deviation away from the correct number of factors). These are defined below:

$$PC = \frac{\sum C}{N}, \text{ for } C = \begin{cases} 1 & \text{if } \hat{\theta} = \theta \\ 0 & \text{if } \hat{\theta} \neq \theta \end{cases},$$

$$MBE = \frac{\sum(\hat{\theta} - \theta)}{N},$$

$$MAE = \frac{\sum|\hat{\theta} - \theta|}{N},$$

where $\hat{\theta}$ is the estimated number of factors, θ is the population number of factors, and N is the number of sample data matrices simulated.

A second approach was used to quantify the accuracy of the item placement of the community detection algorithms, specifically, whether the items were being identified in the correct dimension. The number of dimensions, for example, could be estimated correctly; however, some dimensions may have items that belong to a different dimension than the population dimension.

One common approach from the network science literature is to use normalized mutual information (NMI; Danon, Díaz-Guilera, Duch, & Arenas, 2005). NMI defines a confusion matrix, N , where the rows correspond to the population dimensions and the columns correspond to the estimated dimensions. The element, C_{ij} , refers to the number of items that are found in population dimension i that are in the estimated dimension j . Using the information-theoretic measure of mutual information, this defines NMI as:

$$NMI = \frac{-2 \sum_{i=1}^{C_A} \sum_{j=1}^{C_B} N_{ij} \log (N_{ij}N / N_{i.}N_{.j})}{\sum_{i=1}^{C_A} N_{i.} \log (N_{i.}/N) + \sum_{j=1}^{C_B} N_{.j} \log (N_{.j}/N)},$$

where C_A is the number of population dimensions and C_B is the number of estimated dimensions. Notably, when there is only one dimension (either population or estimated), then NMI is equivalent to PC (i.e., all items are in one dimension = 1 or at least one item

is in a second dimension = 0). The NMI metric can be roughly interpreted as the proportion of items properly placed into the correct dimension, but with a slightly larger penalty for items not placed in the correct dimension.

Results

Accuracy and Bias

The overall performance of the network and PA algorithms are presented in Figure 4. As shown in Figure 4, the number of responses did not have much effect on the accuracy of the GLASSO method but did have a considerable effect on the accuracy of the AIC, BIC, and PA methods. In fact, both parallel analysis algorithms dropped over 10% overall accuracy from continuous responses to polytomous responses ($\Delta_{PCA} = 10.1\%$ and $\Delta_{PAF} = 20.2\%$). When collapsed across number of responses, the Louvain, Fast-greedy, and Walktrap algorithm of the GLASSO method had the best accuracy (88.6%, 87.8%, and 87.1%) followed by the PCA algorithm of PA method (86.7%; Table 3). For the network methods, there was a general trend for the GLASSO method (79.9%) to perform better than the two non-regularized partial correlation methods (AIC = 63.3% and BIC = 58.6%), which held regardless of number of responses (i.e., continuous vs. polytomous data; Figure 4). When looking between the number of responses, most methods appeared to have higher accuracy for continuous than polytomous data.

Percent Correct, Mean Absolute Error, and
Mean Bias Error by Method and Algorithm
Continuous and Polytomous Data

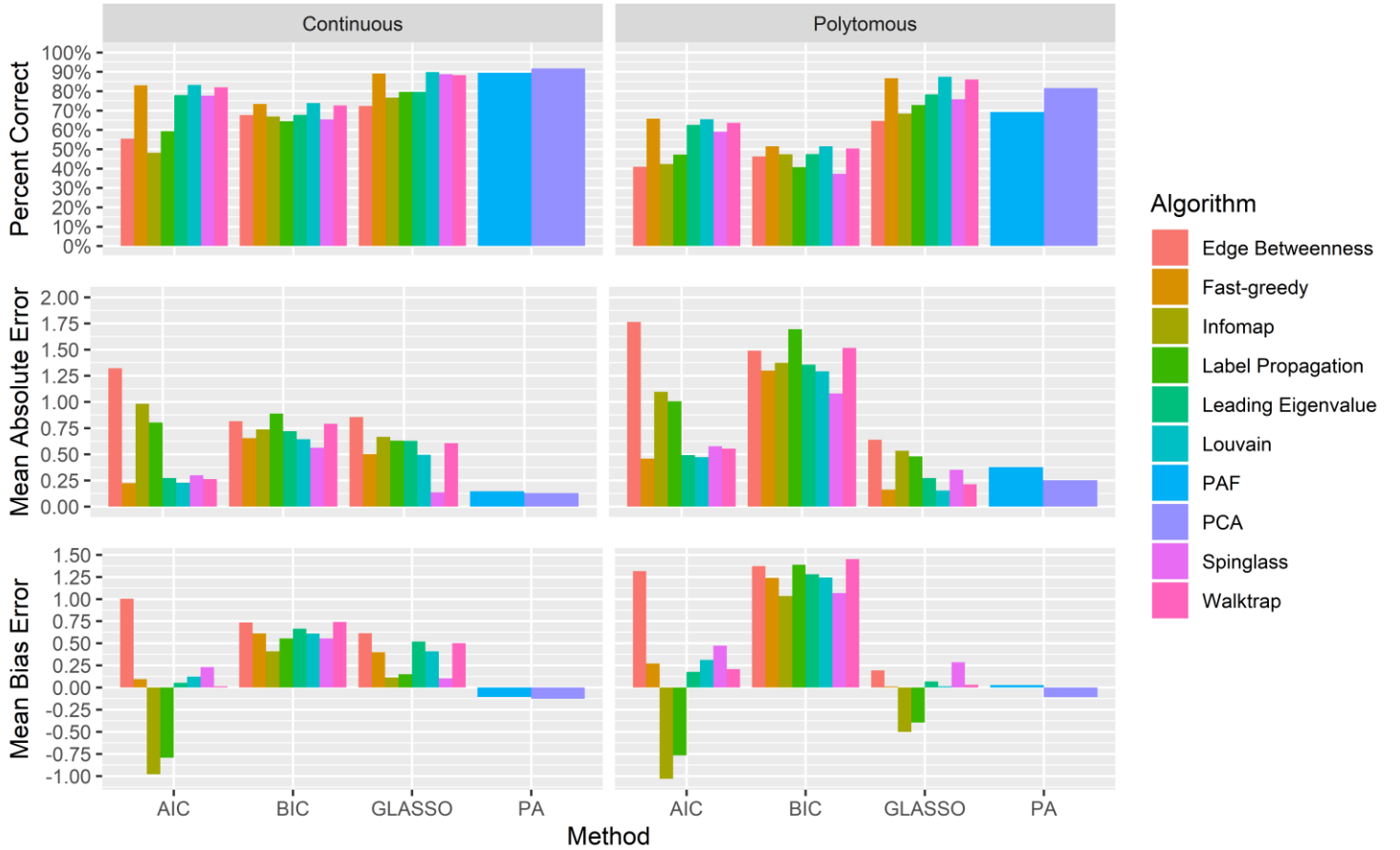


Figure 4. Accuracy and bias measures broken down by method, algorithm, and number of responses.

As for the community detection algorithms, the Louvain, Fast-greedy, and Walktrap had the highest overall percent correct (Table 3) and were the least affected by number of responses when used with the GLASSO method (Figure 4). The near equivalent performance of the Louvain and Fast-greedy algorithms was expected as the Louvain algorithm is very similar to the Fast-greedy algorithm with a modification for

hierarchical structures (i.e., communities are not merged but rather nodes are switched between communities). Notably, the Spinglass algorithm (70.7%) had a high overall accuracy but was unable to estimate a large proportion of the conditions when used with the BIC (continuous = 80.9%; polytomous = 83.9%) and GLASSO (continuous = 57.5%; polytomous = 34.1%) methods. This inability of the Spinglass algorithm to estimate dimensions in the networks was likely due to the sparsity of the networks estimated by the BIC and GLASSO methods, which tend to estimate sparser networks than the AIC method. Because of the Spinglass algorithm's lack of estimation for the majority of the simulated conditions, I refrain from interpreting the results of the Spinglass algorithm for the BIC and GLASSO methods.

Table 3. Percent Correct for Each Independent Condition

Algorithm	Method	Sample Size				# of Factors			# of Variables			Factor Correlations				Factor Loadings			Number of Responses		Overall
		250	500	1000	5000	1	2	4	4	8	12	0.00	0.30	0.50	0.70	0.40	0.55	0.70	Continuous	Polytomous	
Edge Betweenness (57.7%)	AIC	36.4	46.9	54.1	55.7	83.7	47.1	15.9	57.7	49.7	37.7	53.7	51.4	47.2	40.3	38.4	52.5	53.7	55.5	40.9	48.2
	BIC	31.7	53.7	68.8	73.1	77.8	53.2	41.5	65.4	58	48.4	65.7	61.8	55.3	45.2	37.4	64.2	68.4	67.6	46.2	57
	GLASSO	62.2	66.3	69.3	74.4	98	64.4	40.7	64.6	69.5	70.9	80.5	74.3	65.4	53.1	52.9	70.9	78.7	72.3	64.6	68.3
Fast-greedy (74.9%)	AIC	52.2	73.7	84.4	87.9	89.7	61.5	72.6	79.9	79.2	64.2	80.9	78.7	73.6	64	60.4	81.4	81.4	83	65.7	74.3
	BIC	30.9	55.1	73.7	89.8	76.3	54	58.2	75.4	63.1	50	69.6	66.8	61.6	51.9	42.6	69.7	74.2	73.3	51.5	62.5
	GLASSO	79.7	86.8	89.8	93.7	98.7	80	84.4	85.3	89.2	89	94.7	92.2	87.3	77	68.3	93	99.3	89.1	86.7	87.8
Infomap (58.1%)	AIC	38.2	43.6	49.7	49.6	99.3	18.2	21.4	41.7	48.8	44.9	52	48.6	43	37.2	34.2	41.9	59.9	48.1	42.3	45.2
	BIC	34.1	53.5	68.7	71.8	81.7	47.5	44.2	60.4	60.6	51.1	66.6	62.8	55.3	44.1	36.7	62.8	71.1	66.9	47.4	57.3
	GLASSO	67.1	71.7	74.3	75.4	99.3	51.5	65.1	60.9	75.3	80.9	87.1	80.7	68.3	53.1	50.3	72.8	90.4	76.6	68.4	72.3
Label Propagation (60.6%)	AIC	42.7	51.2	58.4	61.1	95.9	44.5	21.9	58.6	54.4	47.1	60.9	57.4	51.6	43.1	40.6	55.3	64	59.3	47.2	53.2
	BIC	27.4	48.5	65.1	68.9	73.1	50.4	36	63.2	52.5	43.3	59.8	56.8	51.2	42.9	34.6	58.6	63.9	64.5	40.7	52.7
	GLASSO	70.1	76	78.2	78.9	98.4	69.8	58.5	71.8	76.7	79.7	89.6	83.5	73	58	54.9	76.6	93.4	79.6	72.7	76
Leading Eigenvalue (69%)	AIC	58.3	71.1	76.1	75.8	93.2	68.6	50.3	71.7	73.7	65.4	75.8	73.8	69.8	61.4	61.5	76.4	73	78	62.5	70.2
	BIC	31.6	52.7	67.8	78.1	77.4	57.3	39.8	69.8	57.9	46.6	63.8	61.3	56.8	48.9	41.7	64.4	66.2	67.8	47.5	57.7
	GLASSO	74.6	78.4	79.6	82.4	98.8	83.9	52.6	77.2	79.2	80.4	85	82.4	78.4	69.8	65.5	82.4	86.7	79.6	78.3	78.9
Louvain (75.2%)	AIC	51.1	72.8	84.3	90.1	89.2	62.2	72.4	80.9	79.7	62.8	80.5	78.4	73.9	64.4	61.4	81.3	80.5	83.1	65.5	74.3
	BIC	31	55	73.4	91	76	54.4	58.8	76.1	63.5	49.9	69.6	66.9	61.9	52.6	43.7	70	73.7	73.8	51.5	62.8
	GLASSO	80.2	87.2	90.4	95.3	98.7	81.4	85.3	85.9	90	89.9	94.8	92.7	88.4	78.5	70.2	93.4	99.4	89.9	87.4	88.6
PFA (79.4%)	PA	59.1	78.8	88	91.5	75.1	82	81	69.2	85.2	83.6	81.2	81.6	79.9	74.9	64	87.5	86.6	89.5	69.3	79.4
PCA (86.7%)	PA	70.1	87.9	92.3	96.4	98.4	87.1	74.5	78.8	90.5	90.7	94.5	93.1	88.2	71	81.1	88.3	90.6	91.7	81.6	86.7
Spinglass (70.7%)	AIC	46.5	66.7	77.2	83.2	80.9	56.5	69	80.6	71.6	54.7	73.3	71.6	67.9	60.7	59.2	75.1	70.7	77.7	59	68.4
	BIC	30.1	48.9	62	80.2	56.7	38.8	60.5	84.6	56.9	42.4	56.2	54	51.4	46.6	53.2	56.9	48.1	65.4	37.2	52.4
	GLASSO	74.8	82	84.8	86.9	91.8	73.7	76	84.7	80.5	78.1	84.7	83.7	80.9	74.3	59.3	85.3	89.5	88.7	75.8	80.8
Walktrap (73.8%)	AIC	54.5	71.3	80.3	85.8	92	66.1	61.4	73.2	76.6	68.5	80.3	77.6	72.1	60.9	58.1	80.4	80	82	63.6	72.7
	BIC	31.1	54.5	72.1	87.9	76.8	56.2	52.8	72.6	62.6	50.4	69.2	66.1	60.5	50.4	41.5	68.8	73.3	72.6	50.4	61.6
	GLASSO	80.4	85.9	88	93	98.6	83.7	78.3	82.2	88.8	90.4	94.8	91.7	86.3	75.6	67.1	91.9	99.3	88.3	86	87.1

Note. Bolded values represent conditions where 80% or more of the replicated samples were estimated correctly. The algorithms are denoted with their percent correct across conditions in parentheses. PFA = principal factor analysis and PCA = principal component analysis.

Digging into the bias measures, the three lowest MAE was for the PA method and PCA algorithm (0.19) followed by the PA method and PAF algorithm (0.26) and GLASSO method and Louvain algorithm (0.32). The other top PC community detection algorithms (Fast-greedy, Louvain, and Walktrap) were generally on the lower end across network methods in the order of GLASSO (0.32, 0.32, and 0.40, respectively), AIC (0.34, 0.35, and 0.41, respectively), and BIC (0.97, 0.96, and 1.15, respectively). In general, the MAE was much lower for the PA and GLASSO methods than the AIC and BIC methods. When split between number of responses, the AIC, BIC, and PA methods generally had greater values, while the GLASSO method had lower values in the polytomous data relative to the continuous data.

The MBE showed that the AIC method had many of the lowest (Walktrap = 0.11, Leading Eigenvalue = 0.12, and Fast-greedy = 0.18) and highest (Label Propagation = -0.78, Infomap = -1.01, and Edge Betweenness = 1.15) values, which largely corresponded with the each algorithm's PC (i.e., greater PC, lower MBE; Figure 4). The PA methods were among the lowest MBE values with a slight tendency to underfactor (PAF = -0.03 and PCA = -0.12). Of the top accuracy community detection algorithms, there was a general tendency to overfactor (Fast-greedy_{GLASSO} = 0.20, Louvain_{GLASSO} = 0.20, and Walktrap_{GLASSO} = 0.26). The MBE generally increased for the AIC and BIC methods in the polytomous data, while it generally decreased for the GLASSO and PA methods.

In sum, the GLASSO method and Fast-greedy, Louvain, and Walktrap algorithms were among the most accurate and least biased across all conditions. For the continuous data, the PA algorithms were among the most accurate and least biased with the top GLASSO algorithms being comparable. For the polytomous data, the top GLASSO algorithms outperformed all other methods and algorithms with the PA method and PCA algorithm following closely behind.

One peculiar takeaway from Figure 4 is that the BIC method appeared to be less affected by which algorithm was being used, with its performance being relatively flat across the accuracy and bias measures. In contrast, the AIC and GLASSO methods' were affected by which algorithm was being used, which could be essentially split into two groups: higher accuracy and lower bias (Fast-greedy, Louvain, and Walktrap) and lower accuracy and higher bias (Edge Betweenness, Infomap, Label Propagation).

Item Placement

Although accuracy and bias measures are important for determining the overall performance of the algorithms, community detection algorithms for the network methods allow for “deterministic” placement of items in dimensions, specifically the algorithms place items in dimensions without the researcher’s direction. The meaning of deterministic is used loosely because some algorithms (Louvain, Label Propagation, Spinglass) are stochastic and therefore may perform better when item placements are aggregated and summarized (e.g., median) across applications (e.g., consensus clustering approaches; Lancichinetti & Fortunato, 2012). It’s important to remember that the NMI metric is equivalent to accuracy when the number of factors is equal to one.

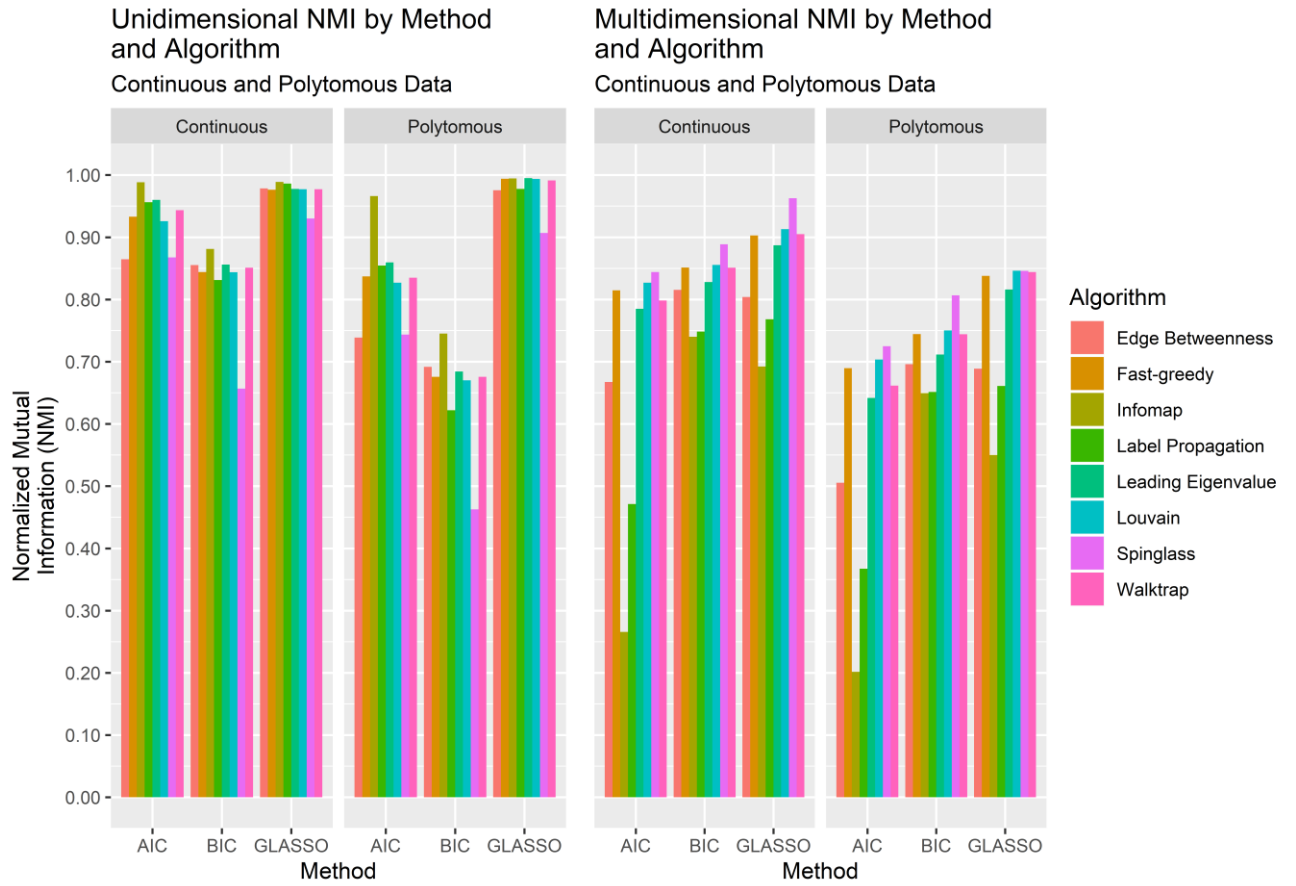


Figure 5. Normalized mutual information broken down by method, algorithm, and number of responses.

Unidimensional structures. In general, most of the algorithms had good performance ($NMI > .80$) regardless of number of responses (Figure 5). The BIC method, however, tended to have the poorest performance and especially when the number of responses were polytomous. The AIC and GLASSO methods tended to have similar patterns of performance for each algorithm; however, the AIC method had lower values for the polytomous data relative to the continuous data, while the GLASSO method had comparable or higher values for the polytomous data relative to the continuous data.

Overall, the GLASSO had the best NMI with several algorithms with values above .98 (in order from greatest to least): Infomap, Leading Eigenvalue, Louvain, Fast-greedy, Walktrap, and Label Propagation.

Multidimensional structures. Relative to the unidimensional structures, the NMI values were much lower across methods except for the BIC method. In contrast, the BIC method generally had better item placement with multidimensional structures (particularly for polytomous data). Consistent with the unidimensional results, most algorithms with the GLASSO method had higher NMI values than all other method and algorithm combinations regardless of the number of responses. Notably, the performance of the AIC method was much lower for multidimensional structures relative to unidimensional structures. Indeed, the BIC method outperformed the AIC on each respective algorithm. Finally, the number of responses had a strong general effect, lowering NMI values about .10 or more across nearly all methods and algorithms.

Summary. Broadly, the GLASSO method had the best item placement performance and demonstrated the highest values of NMI for each respective algorithm. As a general trend across algorithms, the three most accurate and least biased algorithms—Fast-greedy, Louvain, and Walktrap—were also the best performing on the NMI metric. Although this is not surprising, it was certainly not a given because algorithms could hypothetically provide imprecise estimates of the number of dimensions but have more accurate item placements. Overall, the item placement metric provides greater evidence that the GLASSO method, in combination with the Fast-greedy, Louvain, and Walktrap algorithms, is the best performing network method.

Best Algorithms

To provide more nuanced information with condition interactions, I evaluated the accuracy of the top three network algorithms (Louvain, Fast-greedy, and Walktrap) with the GLASSO method and parallel analysis algorithms (Figure 6). Notably, all three network algorithms appear roughly comparable and were largely unaffected by the number of responses (Figure 6). Because of this, I present the results collapsed across number of responses.

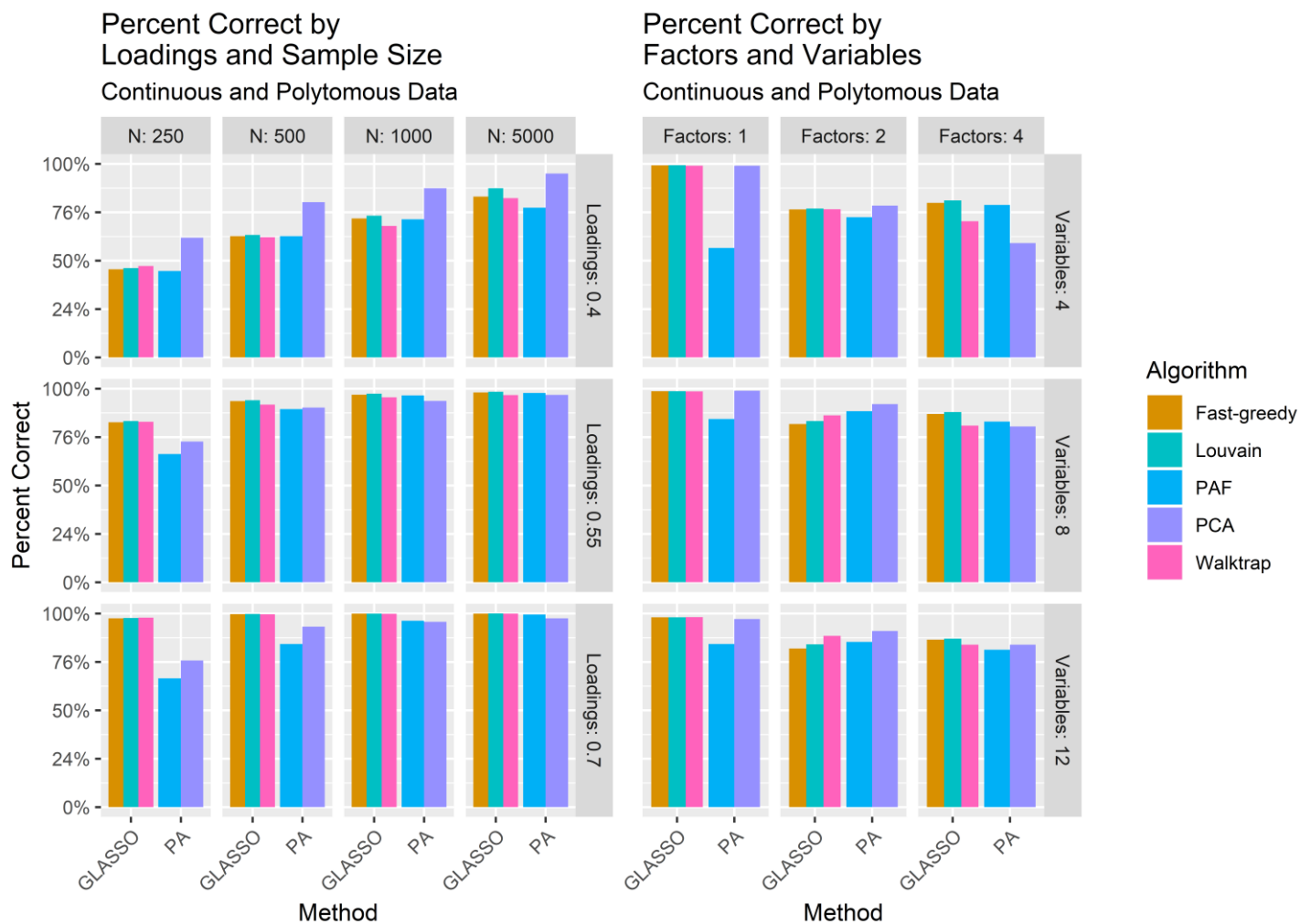


Figure 6. Percent correct broken down by loadings and sample size (left) as well as number of factors and variables (right).

When the percent correct was broken down by loadings and sample size (Figure 6, left), there was a general trend of increased accuracy as loadings and sample size increased. The size of factor loadings appears to have a greater effect on accuracy than the sample size. The PA method and PCA algorithm had the best performance across sample sizes when the factor loadings were small (0.40). The network methods and

algorithms, regardless of sample size, had the best performances when the factor loadings were moderate (.55) or large (.70), replicating previous simulation findings (Golino et al., in press). In general, the GLASSO and PA algorithms' performance are comparable when factor loadings were moderate and large across sample sizes (1000 and 5000, respectively), while the GLASSO algorithms performed better when sample sizes very small and small (250 and 500, respectively).

When the percent correct was broken down by number of factors and variables (Figure 6, right), there was a general trend of increased accuracy as the number of factors decreased. As for the variables, there was a trend for accuracy to decrease as the number of variables increased for the AIC and BIC methods. Conversely, there was a trend for accuracy to increase as the number of variables increased for the GLASSO and PA methods. There was a particularly interesting pattern for the PA method when there were four variables and the number of factors increased, specifically the PCA algorithm had much greater accuracy than the PAF algorithm when there was only one factor, equivalent accuracy when there were two factors, and much lower accuracy when there were four factors (Figure 6).

In general, the GLASSO method appears to be comparable to the PA method across all condition interactions. The network algorithms appeared to have their lowest accuracy relative to the PA method and PCA algorithm when there were low loadings. As for the network algorithms, the Walktrap algorithm appears to have decreased accuracy when there are few variables (4) and many factors (4). In all other interactions, the algorithms performed similarly.

Discussion

This study examined the performance of different network methods and several community detection algorithms to detect underlying latent dimensions. As a comparison, I used the state-of-the-art parallel analysis methods to evaluate whether network methods could be comparable to traditional factor analytic algorithms. In short, I found that some network algorithms could perform comparably to the PA algorithms and this was dependent on the network estimation being used, specifically the Louvain, Fast-greedy, and Walktrap algorithms all performed similarly to the PA algorithms when the GLASSO network estimation method was used. Importantly, this study evaluated two non-regularized network estimation methods and also tested community detection algorithms under traditional psychological conditions (e.g., factor models, ordinal data).

This study was the first to evaluate how different partial correlation network estimation methods performed when identifying dimensions in psychological factor models. Previous work had compared the GLASSO with a correlation-based method, the TMFG, with the GLASSO showing better performance in nearly all conditions (Golino et al., in press). Other work had evaluated the performance of the partial correlation methods used in this study to estimate population network models (Williams & Rast, 2019; Williams et al., 2019). In these studies, the non-regularized partial correlation methods (i.e., AIC and BIC) outperformed the GLASSO on the measure of specificity (avoidance of false positives) in the population network structure. In the context of detecting dimensions, this difference in specificity seemed to benefit the GLASSO where it tended to have better performance for identifying dimensions than both the AIC and

BIC methods. This is likely because the GLASSO was able to consider more edges (or information) in the network, which may have enabled it to better estimate the population factor structure.

As for the community detection algorithms, there has been extensive evaluations of these algorithms across different literatures, but none were specific to psychological factor models. The closest comparison had been with brain network count and correlation structures (Gates et al., 2016). For both count and correlation matrices, the Walktrap algorithm outperformed the other algorithms tests on a measure of item placement. Notably, the Louvain algorithm performed the best when Euclidean Distance was used as a similarity measure. In general, my results largely jibe with their study, showing that the Walktrap and Louvain algorithms were among the best performing algorithms. It's important to note that the Louvain algorithm used in this study (from the *igraph* package in R) may have differed from the one used in their study (from the *Brain Connectivity Toolbox* in Matlab; Rubinov & Sporns, 2010).

One critical finding was that while the Spinglass algorithm was among the best performing algorithms in this study (i.e., Walktrap, Louvain, and Fast-greedy) it was not always able to estimate the number of dimensions in the network. This inability to identify dimensions is likely due to some networks having had unconnected nodes. This was particularly noticeable for the BIC method, which produces the sparsest networks of the three network methods. Across the methods, the performance of the Spinglass algorithm should be tempered with respect to this finding. When anticipating what its

actual performance might be, the AIC method with the Spinglass algorithm had most of the conditions estimated and placed the algorithm among the top methods.

Finally, this study was the first to examine polytomous data with community detection algorithms. As with previous research examining dichotomous data, the differences between the continuous and polytomous data were nuanced but generally showed the same patterns (Golino et al., in press). Overall, the Louvain, Fast-greedy, and Walktrap algorithms were all comparable to the PA methods, particularly when the GLASSO network estimation method was used.

CHAPTER IV

LOADINGS

The evaluation of item quality is fundamental to scale development and validation. Item analyses provide insight into how items relate to one another as well as dimensions of the scale. Item analyses are often used to determine whether items should be removed from the scale because they are not performing as expected (DeVellis, 2017). In contemporary psychometrics, EFA is the most common method applied to obtain this information (Flora & Flake, 2017; Hubley, Zhu, Sasaki, & Gadermann, 2014). EFA presents this information as a factor loading for each item in each dimension, representing an item's association with the dimension.

In most situations, researchers apply EFA with an *oblique* rotation to allow factors to correlate with one another. The output of this analysis includes three factor loading matrices: pattern (unique association between item and factor, controlling for correlations between factors), structure (zero-order correlation between item and factor), and factor (loadings before rotation; Furr, 2017). The pattern matrix is typically used to evaluate items because it provides researchers with the clearest picture of how items “load” onto each individual dimension.

The term “load” in factor analytic jargon is provided by items being regressed on the factors. This gives the substantive interpretation of how well an item represents or measures the latent factor. The main objective in evaluating items is to determine which items have the largest loadings on a single dimension and low loadings on other

dimensions (DeVellis, 2017). Items with this order of loadings is often preferred because it suggests that these items represent a single psychological attribute and in turn a common cause.

From the network perspective, items are evaluated using network measures called *centrality*. Centrality measures quantify the relative position of nodes based on their connections to other nodes in the network. To date, the substantive interpretation of centrality measures has been unclear and subject to debate (e.g., Bringmann et al., 2019). The most common interpretation has been that these measures quantify the relative influence or importance of a node in the network, which suggests increased causal efficacy. Based on this interpretation, many researchers have suggested that more central nodes represent important intervention targets (e.g., symptoms in a psychopathological disorders).

Unfortunately, these interpretations have not held up empirically, with many studies reporting that there is little evidence for the relationship between a node's centrality and its causal efficacy (Dablander & Hinne, 2019). For some researchers, this has led to the development of different network measures that have more straightforward interpretations (e.g., *predictability* or a node's predicted variance from other nodes; Haslbeck & Waldrop, 2018). For others, this has led to a call to get "back to basics" and determine whether these measures are meaningful in a psychological context (Bringmann et al., 2019).

Review of Hallquist, Wright, and Molenaar (2019)

A recent series of simulation studies sought to determine the meaning of centrality measures in relation to factor loadings. In Hallquist et al.'s (2019) study, they compared the most frequently used centrality measures—betweenness, closeness, and node strength—with CFA factor loadings. In their first simulation study, they examined whether there was any correspondence between these centrality measures and factor loadings in unidimensional and multidimensional latent trait models. They setup conditions with 10 variables per factor for models of one, two, and three factors. Factor loadings varied between .4 and .95 and the factors were either orthogonal or moderately correlated (.40). Across the conditions, a sample size of 400 was generated. For the comparison, they fit CFA and GLASSO models to the data.

Their results demonstrated that betweenness (relative number of times a node is used on the shortest path from one node to another) and closeness (distance a node is from the center of the network) centrality were highly correlated with the CFA factor loadings of the one factor model ($r = .74$ and $r = .94$, respectively) but had much lower correlations with these loadings when there was more than one factor (r 's between .31 and .55). In contrast, node strength was significantly correlated with the CFA factor loadings across the models (r 's between .97 and .98). Because of the lack of correspondence of betweenness and closeness centrality with factor loadings, I discuss the rest of the simulations results with node strength only.

In their second simulation study, they examined the effects of common versus specific sources of covariation—that is, the extent to which two indicators on different

factors were related through a shared separate factor (these will be referred to as the *target* indicators). These effects were examined in one of the target indicators and a comparator indicator (i.e., an indicator on the same factor as the respective target indicator). Similar to the first simulation, there were 10 items per factor and a sample size of 400. Different from the first simulation, there was only a condition with two factors and all but one item in their respective factors had a factor loading of .80. The two items that were associated had their correlation vary between $r = 0$ and $r = .64$.

A general finding of this study was that the edge weight (i.e., partial correlation) between the target indicators had a nearly perfect relationship with the extent to which there was a specific association between them ($r = .997$). As for the node strength estimates, there was a moderate main effect of specific-to-shared variance balance and large main effect of indicator type (target and comparator). This suggests that there was a large increase in a node's strength due to the shared separate factor. The comparator indicator's node strength had a small main effect from the specific-to-shared variance balance, suggesting minimal impact from the shared separate factor.

In their third and final simulation study, they examined the effects of multiple latent causes. This study was setup with a two-factor model with eight indicators per factor and the target indicator that loaded onto both factors (i.e., 17 indicators in total). The target indicator had factor loadings on both factors ranging between .20 and .80 in increments of .05. All other loadings were fixed at .80. Similar to their previous simulations, sample sizes of 400 were generated. Like their second simulation, they also examined a comparator indicator. The results of this study revealed that the target

indicator's node strength was an equally weighted combination of Factor 1 and Factor 2 loadings (both r 's = .94). The comparator indicator's node strength was weakly associated with the variation of the target's factor loadings on Factor 1 and Factor 2.

In summary, their simulations demonstrated that the network measure node strength is (a) roughly redundant with CFA factor loadings and (b) affected by different causal sources. These takeaways are important for their own reasons. The first finding suggests that there is a strong connection between node strength and factor loadings, which means that node strength could be used as a potential psychometric tool for item selection in network models. The second finding suggests that the relationship between node strength and factor loadings should be tempered in a way that reflects the unique latent causes in the system. This latter takeaway jibes with the notion that the unique causal components must be identified before network measures can be meaningfully interpreted (Christensen et al., under review; Hallquist et al., 2019).

Present Research

The goal of this simulation study was to extend Hallquist and colleagues' (2019) first simulation study by considering the lessons learned from their second and third simulation. For example, examining how node strength relates to population factor loadings when split by dimensions. This study offers two key additions to their simulations. First, node strength is split between dimensions in order to compensate for the effects of different latent causes that underlie its computation. For this computation, I formalize a standardization of node strength in each dimension that I hereafter refer to as *network loadings*. This term is used to denote the similarity between this formalization

and factor loadings but to also keep the specification that they are derived from the network counterpart.

Second, this study compares the accuracy of network, EFA, and CFA loadings in the estimation of population factor loadings. This contrasts with Hallquist and colleagues' simulation where node strength was correlated with CFA loadings. A direct comparison with the population factor loadings is a better benchmark for whether network models can accurately identify this information and allows for a better comparison of what networks loadings are more "like." On the one hand, CFA loadings typically offer a simple structure where indicators only load on their factor. On the other hand, EFA loadings offer the full complexity of dominant and cross-loadings, which tends to be more useful in scale development contexts. Network models are likely to offer the in-between because some indicators may not connect with indicators in other dimensions, leaving zeros in the matrix.

Method

Data Generation

The data generation approach followed the same approach as in Chapter II's Method.

Psychometric Network Model

Similarly, the same EGA with GLASSO network estimation and Walktrap community detection algorithm in Chapter III's Method was used in this study.

Network Loadings

An important finding of Hallquist and colleagues' (2019) simulations was that node strength represented a combination of dominant and cross-factor loadings. To circumvent this issue, a node's strength can be split between the nodes in each dimension. This can be mathematically written as:

$$NS_i = \sum_{j=1}^n w_{ij},$$
$$NL_{iC_k} = \sum_{j \in C_k}^C NS_{ij},$$

where w_{ij} is the weight (e.g., partial correlation) between node i and j , NS_i is the sum of the node strength for node i across all nodes, and NL_{iC_k} is the weight for node i , which is its sum of all the weights for nodes in dimension C_k . This measure can be standardized using the following formula:

$$Z_{NL_{iC_k}} = \frac{NL_{iC_k}}{\sqrt{\sum_{j \in C_k} NL_{jC_k}}},$$

where the denominator is equal to the square root of the sum of all the weights for nodes in dimension C_k . These standardized network loadings are in the unit of association used in the network, which means the meaning of these network loadings will change based on the association unit used. Importantly, not all nodes are connected to nodes in other dimensions, which means that there will be zeros for some dimensions in the network loading matrix. The network loadings were computed using the `net.loads` function in the *EGAnet* package.

EFA Loadings

For the EFA model, I used the *psych* package's (Revelle, 2018) `fa` function to estimate the factors in the data. Because the number of factors is known, I specified the population number of factors as the number of factors to compute in the EFA. The factor model was estimated using the maximum likelihood for continuous data and weighted least squares for polytomous data. For both types of data, I used the *geomin* oblique rotation from the *GPArotation* package (Bernaards & Jennrich, 2005), which has been shown to have low bias when the factor loadings display a simple structure (i.e., small cross-loadings; Sass & Schmitt, 2010) and have factor loadings closer to CFA (Schmitt & Sass, 2011). Note the cross-loadings in this study were smaller than the simulations performed in Chapter II and III, meaning that the loading structure was closer to a simple structure.

CFA Loadings

For the CFA model, I used the *lavaan* package's (Rosseel, 2012) `cfa` function to estimate factor loadings. The CFA models were specified with the known population structure of the data—that is, the population dimensions with the items placed in their known dimensions. For the continuous data, I used the maximum likelihood estimator; for the polytomous data, I used the weighted least square mean and variance adjusted estimator.

Design

Similar to the population models in Chapter II and III, they were simulated from a multidimensional multivariate normal distribution where factor loadings for each item were generated with $\pm .10$ deviance drawn from a uniform distribution.

In contrast to previous designs, smaller cross-loadings were generated from a random normal distribution with a mean of zero and standard deviation of .05. Moreover, there was only one condition of very large factor loadings (.85). These adjustments in design were made to ensure that variables firmly loaded onto their designated factor and could easily be identified by the Walktrap algorithm.

Two, three, and four factors were simulated to ensure that there were cross-loadings. Four and eight variables per factor were generated to represent conditions commonly found in psychological research and validated scales. Similar to the first two simulations, correlations between factors were orthogonal (.00), small (.30), moderate (.50), and large (.70). Finally, large sample sizes of 1000 and 5000 were generated to ensure that adequate loading estimations could be obtained.

The simulation design of the current study allowed for a mixed factorial design: $3 \times 2 \times 4 \times 2$ (number of factors \times variables per factor \times correlations between factors \times number of responses) for a total of 48 simulated condition combinations.

Statistical Analyses

To compare the performance of the network, EFA, and CFA loadings, I used Spearman's rank-order correlation between each method's loadings and the known

population loadings. Rank-order rather than Pearson's correlation was chosen to have a larger penalty for having loadings that differ in their order from the population loadings.

Results

Across all conditions, the EFA loadings were the most accurate ($\bar{r} = .948$) followed by the network loadings ($\bar{r} = .926$) and CFA loadings ($\bar{r} = .831$). Notably, the type of data did not appear to make a difference: EFA ($\bar{r}_{cont} = .951$ and $\bar{r}_{poly} = .944$), network ($\bar{r}_{cont} = .928$ and $\bar{r}_{poly} = .923$), and CFA ($\bar{r}_{cont} = .835$ and $\bar{r}_{poly} = .827$). When breaking the results down by conditions, a much more detailed pattern emerges (Figure 7).

Comparison of Factor and Network Loadings

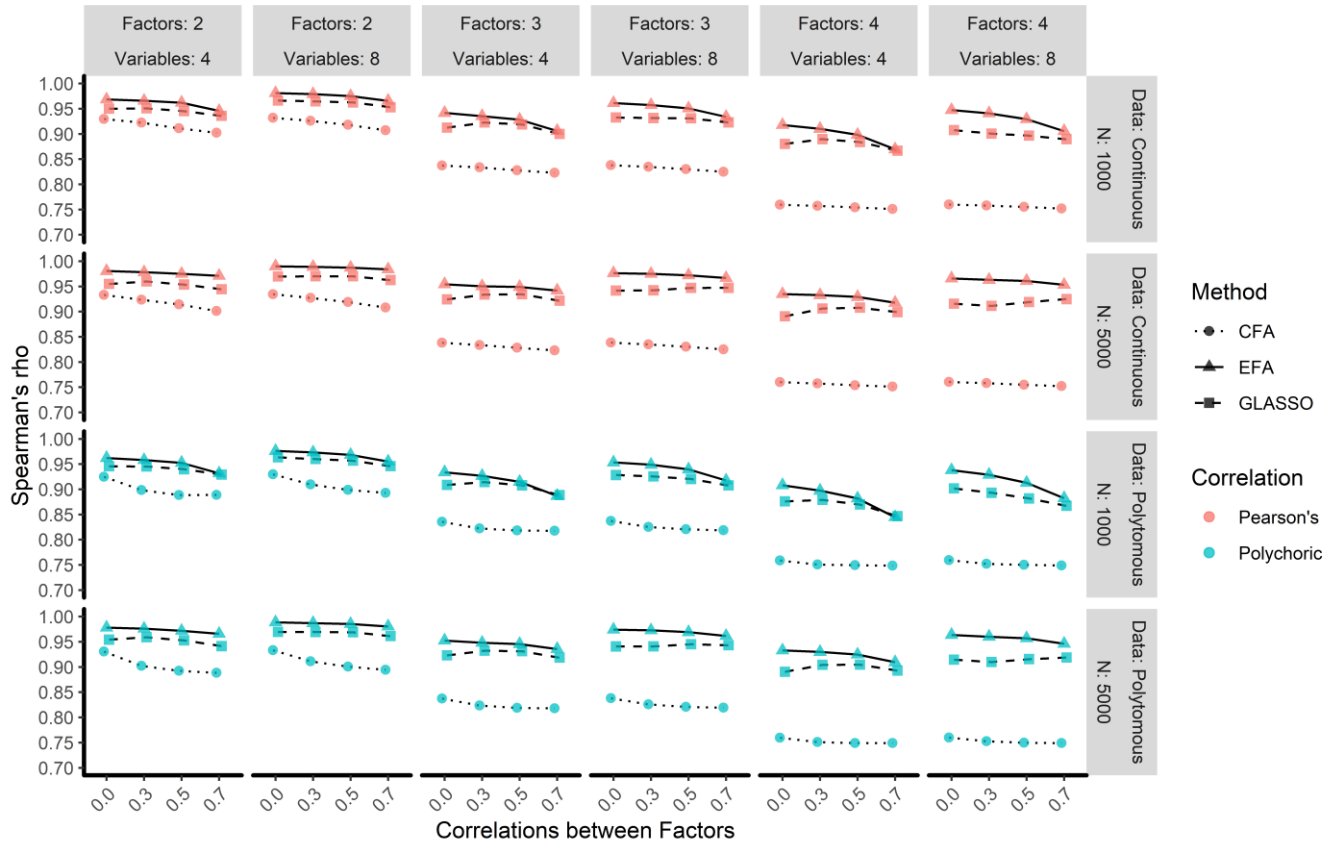


Figure 7. Comparison of factor and network loadings broken down by each condition. Network loadings are represented by the dashed line and square, CFA loadings are represented by the dotted line and circle, and EFA loadings are represented by the solid line and triangle. Continuous data are presented in red and polychoric data are presented in blue. Note that the y-axis begins at .70.

In Figure 7, there were several notable trends to point out. As a general trend, all loading estimation methods were less accurate as the number of factors increased. This is particularly noticeable for the CFA loadings, which were likely affected by its simple structure (i.e., zeros for all non-dominant factor loadings). Another general trend is that the network loadings are right below or comparable to the accuracy of the EFA loadings. The network loadings tended to mostly resemble the EFA loadings when the correlations

between factors was high (.70). In comparison to the EFA loadings, the network loadings appeared to have a relatively lower accuracy when the number of factors increased, which is likely due to the network loadings estimating more zeros in the loading matrix—much like the CFA loadings.

Interestingly, accuracy appeared to increase across loading types when there were a greater number of factors (Figure 7). This trend is likely due to the conditions tested rather than an actual trend of the methods applied, specifically factor loadings were high for the dominant loadings relative to the cross-loadings, which increased the probability that a greater number of indicators would be in the correct rank-order pairs. Sample size did not appear to have much of an effect on the accuracy of the estimates across methods, with a slight increase in accuracy for larger sample sizes.

Discussion

This study sought to derive and evaluate a standardized node strength measure that separated specific contributions of a network's dimension. My study builds on the results and recommendations from Hallquist et al.'s (2019) simulation studies. Factor models with population loadings, for example, were used as a comparison of EFA, network, and CFA loading accuracy rather than cross-comparing measures. My study also analyzed conditions where there were cross-loadings, which added some potential for noise. In large part, my results demonstrate that when node strength is divided between dimensions, they can be shown to accurately recover the population loadings of a factor model.

This result has several implications for the use of network models in assessment. The main implication is that my simulation provides further evidence that node strength is statistically equivalent to factor loadings (Hallquist et al., 2019). Despite this equivalence, it's important to consider their substantive interpretations. As mentioned before, factor loadings refer to how well an indicator measures an underlying common cause. From the network perspective, network loadings are not an indicator of a common cause but rather the coupling of components and the emergence of dimensions in a causal system. In this sense, a node's strength represents its contribution to the emergence of a coherent dimension (or network).

Extending from this implication, network loadings can be used as an equivalent measure of factor loadings, providing many of the same measurement opportunities as other factor models (despite substantive differences). A network loading matrix, for example, can be derived and used for item selection in scale development and validation (DeVellis, 2017). This also opens the door to computing measurement invariance measures such as metric equivalence for network loadings. Finally, network loadings can be used to derive a weighted between-person score for each participant in the model—that is, the network equivalent of factor scores can be derived.

This last implication requires more detailed attention, specifically how should a network score be computed and substantively interpreted? When considering a network of extraversion components, the network itself references the state of the system—that is, the extent to which the network is in an extraverted state, which is determined by the total activation of its components (Christensen et al., under review). From this perspective,

extraversion represents a summary statistic of how components of the network are influenced by one another (Cramer, 2012). Therefore, a network score is more analogous to a formative latent variable (i.e., a weighted composite) than a reflective latent variable (i.e., a common covariance). This substantive explanation suggests that a network score should be computed as a weighted composite, which could be derived from the product of network loadings and each person's corresponding item responses.

For these discussion points, it's important to understand their limitations within the context of my results. First and foremost, the factor loadings were very high and sample sizes were large. In these conditions, the network loadings are more likely to be accurately estimated. In smaller samples and lower population loadings, fewer edges will be estimated in the network, which would lower the accuracy of the loadings estimation (similar to Chapter III's dimensionality results), becoming more like CFA loadings rather than EFA loadings. Moreover, when comparing metric equivalence of samples with different sizes (e.g., $n = 500$ and $n = 5,000$), there is unlikely to be metric equivalence even when there should be because the smaller sample will estimate fewer edges than the larger sample. One potential solution would be to estimate the networks as if they had equivalent sample sizes (i.e., adjusting the GLASSO sample size parameter to be equal), which would allow for a similar number of edges to be estimated.

Future work should evaluate more extensive conditions than the ones in this study, such as smaller sample sizes and different levels of factor loadings, including a condition where factor loadings are variable sizes to better reflect more realistic data conditions. Moreover, larger cross-loadings should be estimated and perhaps adjusted

with the size of correlation between factors (e.g., increasing cross-loadings with size of factor correlations). There should also be a wider comparison of EFA factor loading rotations to examine whether, in certain conditions, network loadings may perform better than rotations that are considered less optimal in those conditions. Similarly, network loadings are largely dependent on the network estimation method (e.g., non-regularized partial correlations networks; Williams et al., 2019), which may alter the results shown in this study.

CHAPTER V

EMPIRICAL EXAMPLE

The three simulations in this paper represent the statistical methods necessary to validate the structure of assessment instruments from the network perspective. These simulations provide evidence for the conceptual framework put forward by Christensen and colleagues (under review). In this framework, the first objective for the validation of any assessment instrument (extant or in development) is to reduce the redundancy of the instrument. After reducing redundancy, the dimensionality of the instrument can be assessed to determine whether the intended structure is identified. Finally, item analyses (e.g., network loadings) can be computed and used to determine the quality of the components in the network. If any items are removed, then dimensionality can be re-assessed.

In accordance with this framework, I provide an empirical example that executes these validation steps. The example is outlined as follows: first, I introduce the node redundancy strategies and guidelines used to decrease the number of components in an instrument. Dimensionality and loadings are straightforward enough that no additional introduction is necessary, beyond their Chapters (III and IV, respectively), to understand their application in the example. Second, I briefly review the personality inventory and demographics of the sample used in the example. Finally, I report and discuss the results of the example.

Node Redundancy Strategies

Once a researcher has their results from the redundancy analysis, they must then use theory about the attribute to guide the identification of redundant items in an assessment instrument. If deciding items should be reduced to a single component, then there are two quantitative strategies that can be used. The first option is to remove all but one item from the questionnaire. When taking this option, there are a few considerations researchers must make. Qualitatively, which item represents the most general case of the attribute? Often items are written with certain situations attached to them (e.g., “I often express my opinions in group meetings”; Lee & Ashton, 2018), which may not apply to all people taking the questionnaire. Therefore, more general items may be better because they do not represent a situation-specific component of an attribute (e.g., “I often express my opinions”). Quantitatively, which item has the most variance? This is a common criterion in traditional psychometrics because greater variation suggests that this item better discriminates between people on the specific attribute (DeVellis, 2017). There may also be cases where one item overlaps with two other items, but the other two items do not overlap themselves (i.e., a mediating item). In these instances, I recommend selecting the mediating item because it sufficiently captures the variance of the other two items to the extent that they are unrelated when controlling for all other items in the network.

The more straightforward option is to combine items into a single variable. This can be done by taking each participant’s sum (or mean) score across redundant items or by estimating a latent variable score (e.g., Epskamp, Rhemtulla, & Borsboom, 2017). Using a latent variable approach is the recommended option because it retains all of the

information in the assessment instrument and maintains the notion of an underlying common cause of the component, offering a more reliable and valid assessment of certain components in the network. Importantly, these components can be reduced to single items with general phrasing when considering item selection or developing a shorter assessment instrument. For the example, I will use the latent variable approach to combine items whose redundancies are due to a theoretical common cause.

Node Redundancy Guidelines

The node redundancy analysis maps the redundancies of each significant pair of connections between nodes (i.e., items). This analysis begins with the item that has the most redundancy with other nodes (i.e., greatest number of significant redundancies) and continues until all redundancies are resolved using one of the strategies discussed above. Each node in this process is evaluated individually and hereafter will be referred to as the *target node*. Importantly, a target node is redundant with other nodes, each which may also have their own redundancies with other nodes. Some of these other nodes may be redundant with the target node, while others may not. The redundancy analysis first identifies nodes with the target node and then iteratively identify nodes that are redundant with those nodes until there are no longer nodes redundant with the identified nodes. This process forms a so-called “redundancy chain” (Figure 8).

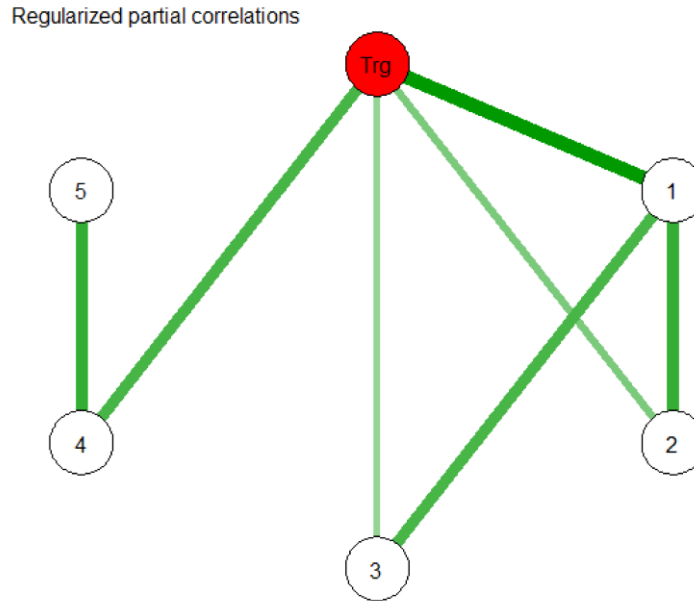


Figure 8. An example of a redundancy chain plot. The red node indicates the target item and the white nodes with numbers correspond to the numbered options. A connection represents significant overlap determined by the redundancy analysis and the thickness of the connection represents the regularized partial correlation between the nodes in the network.

In the redundancy chain (Figure 8), the target node is labelled with “Trg” and depicted in red, while the other nodes are labelled with numeric identifiers. The connections between nodes represent significant redundancy between two nodes. When focusing on the target node, there are connections to Nodes 1, 2, 3, and 4. Notably, Node 5 is not connected to the target node but is connected to Node 4, suggesting that Node 4 has additional redundancies beyond its redundancy with the target node. In this way, there is a “chain” of redundancies from the target node to Node 4 and Node 4 to Node 5.

As a general guideline, there should be particular importance given to *cliques* or fully connected sets of nodes. In Figure 8, there are two 3-cliques (or triangles) with the target item (i.e., Trg – 1 – 2 and Trg – 1 – 3). In the network literature, these triangles

contribute to a measure known as the *clustering coefficient* or the extent to which a node's neighbors are connected to each other. Based on this definition, the clustering coefficient has recently been considered as a measure of redundancy in networks (Costantini et al., 2019; Dinic, Wertag, Tomašević, & Sokolovska, in press). In this same sense, these triangles suggest that these items are likely to have particularly high overlap. Therefore, triangles in these redundancy chain plots can be used as a heuristic for selecting items.

It's important to note, however, the absence of a connection between nodes in the redundancy chain plot may not necessarily mean that two nodes are *not* redundant. The connections only represent nodes that were deemed statistically significant (keeping in mind the results of Chapter II). It's plausible that two nodes could be very similar and yet only one of the two is connected to a third node (e.g., a mediating node). Therefore, the clique heuristic is not a steadfast rule but a general guideline. Theory about the underlying cause of the relations between nodes should be the leading heuristic for whether two nodes are redundant (regardless of statistical redundancy)—that is, are the redundant relationships between two or more nodes due to a common cause (combine to form latent variable) or reciprocal causes and effects (do not combine to remain unique causal components)? In short, the redundancy analysis and clique heuristic provide statistical evidence of redundancy that researchers must weigh with theoretical evidence of cause.

SAPA Inventory

The Synthetic Aperture Personality Assessment (SAPA) inventory was developed by David Condon (2018) for the purpose of moving personality assessment towards a more iterative, transparent, and empirical process. The development of the SAPA inventory followed an empirical approach rather than a theoretical approach by administering “as many items as possible based on administration to as many participants as possible” (Condon, 2018, p. 3). Using the more than 3,000 items available in the International Personality Item Pool (IPIP; Goldberg, 1999; Goldberg et al., 2006), a little more than 600 unique items were selected to cover most of the widely used measures. From this item set, over 34,000 people completed portions of these items over time until all people responded to all items (from December 2013 to February 2017; Condon, 2018).

The SAPA inventory dataset that I will use for my example comes from the “spi” dataset in the *psychTools* package (Revelle, 2019) in R. This dataset includes a 135-item inventory (items were primarily selected from the International Personality Item Pool; ipip.ori.org). These 135 items form an empirically derived structure of 27 personality dimensions. A subset of these items ($n = 70$) form an empirically derived five factor structure that corresponds to the Five Factor Model (FFM; McCrae & Costa, 1987). The instructions were to, “Respond to each item with how accurately the description describes you.” The response options ranged from 1 (“Very inaccurate”) to 6 (“Very accurate”).

This 70-item subset was completed by 4,000 participants over the SAPA project website (sapa-project.org). These participants were collected after the developmental

dataset (from February 2017 to May 2017) and were the first 4000 *complete* cases (not the first 4000 participants; D. Condon, personal communication, January 29, 2020). The sample had a mean age of 26.90 ($SD = 11.49$, range = 11–90) and were well represented for both sex (59.5% female) and education (11.1% graduated high school, 31.8% currently in university, 22% graduated university, and 11.8% held a graduate or professional degree). Race and ethnicity demographics were not provided; however, the data was gathered via the SAPA project website allowing equal opportunity for people of all ages, genders, ethnicities, and socio-economic backgrounds as long as they had access to the internet. Moreover, the exploratory, replication, and confirmatory datasets that were previously collected demonstrated substantial diversity, especially relative to past large-scale personality projects (e.g., Eugene-Springfield Community Sample; Condon, 2018; Goldberg & Saucier, 2016).

One potential sampling bias for this sample was that these participants were included because they completed all 135 items, meaning that participants who did not complete all 135 items during the same time period were not included (regardless of whether they stopped or unintentionally skipped an item; D. Condon, personal communication, January 29, 2020). Despite this potential for representative bias, this sample likely represents a broader and more diverse population than most other self-report research in the personality literature.

There are several reasons for choosing this dataset, but I will elaborate on three specific reasons. First, as just mentioned, the dataset is a large, diverse sample that is open-source, making the analyses performed in this study free for experimentation and

replication. Second, personality inventories are perhaps the most commonly used assessment instruments across psychological research and therefore represent the vast majority of the applications that these analyses target. Finally, the SAPA inventory is structured hierarchically: there are 27 empirically derived lower-order dimensions that can be further collapsed into the prototypical FFM (Condon, 2018). These lower-order dimensions contain substantial redundancy, making the dataset a good example for how the redundancy analysis can be applied and the number of unique components to expect (i.e., around 27).

Results and Discussion

Redundancy

Based on the results from the redundancy simulation (Chapter II), either the weighted topological overlap or partial correlation approach would have been comparable in these data conditions: polytomous data, large sample size, and expectation that there was a large amount of redundancy in the SAPA inventory. I opted to use the weighted topological overlap approach with adaptive alpha because it is a network-derived measure and therefore represents the network psychometric approach.

Following the strategy of combining redundant items with latent variables and the clique heuristic, I reduced the 70-item inventory down to 26 personality components. Interestingly, these 26 components largely reflected the 27 empirically identified lower-order factors found by Condon (2018). This suggests that the redundancy analysis was not only effective but mirrors the empirically defined structure found by other methods. Importantly, these components I identified were driven by statistical heuristics and

theoretical knowledge about the plausible latent causes underlying these redundancies.

The item composition and labels of these components can be found in Table 4.

Table 4. Components Identified in the Node Redundancy Analysis

Dimension	Component (Node Label)	Item Content				
1	Orderly (Ord)	Keep things tidy.	Often forget to put things back in their proper place.	Leave a mess in my room.	Like order.	
1	Motivated (Mtv)	Find it difficult to get down to work.	Need a push to get started.	Start tasks right away.		
1	Perfectionist (Prf)	Want every detail taken care of.	Continue until everything is perfect.			
1	(Shsfmao)	Set high standards for myself and others.				
1	(Nmd)	Neglect my duties.				
1	(Wh.)	Work hard.				
2	Emotional Stability (Ems)	Experience very few emotional highs and lows.	Get overwhelmed by emotions.	Experience my emotions intensely.	Think that my moods don't change more than most peoples do.	
2	Worrier (Wrr)	Worry about things.	Fear for the worst.	Am a worrier.		
2	Irritable* (I(R)	Rarely get irritated.	Am not easily annoyed.	Seldom get mad.		
2	Anxious (Anx)	Would call myself a nervous person.	Panic easily.			
2	Low self-esteem (L..)	Feel a sense of worthlessness or hopelessness.	Dislike myself.			
3	People person (Ppp)	Usually like to spend my free time with people.	Like going out a lot.	Avoid company.	Want to be left alone.	Don't like crowded events.
3	Attention-seeking (At-)	Hate being the center of attention.	Like to attract attention.	Dislike being the center of attention.	Make myself the center of attention.	
3	Laugher (Lgh)	Laugh a lot.	Laugh aloud.			
3	Social-efficacy (Sc-)	Am skilled in handling social situations.	Find it difficult to approach others.			
3	(Eme)	Express myself easily.				
4	Original ideation (Ori)	Am full of ideas.	Am able to come up with new and different ideas.	Am an original thinker.	Love to think up new ways of doing things.	
4	Introspective (Int)	Love to reflect on things.	Try to understand myself.	Spend time reflecting on things.		
4	Self-assessed intelligence (S-i)	Think quickly.	Am quick to understand things.	Can handle a lot of information.		
4	Fantasy (Fnt)	Have a vivid imagination.	Like to get lost in thought.			
5	Concerned for others (Cfo)	Am sensitive to the needs of others.	Feel sympathy for those who are worse off than myself.	Think of others first.	Am concerned about others.	Sympathize with others' feelings. Feel that most people can't be trusted.
5	Sees good in people (Sgip)	Trust what people say.	Believe that people are basically moral.	Trust people to mainly tell the truth.	Believe that others have good intentions.	
5	Manipulative (Mnp)	Use others for my own ends.	Cheat to get ahead.	Tell a lot of lies.		
5	Rule-follower (Rl-)	Rebel against authority.	Try to follow the rules.	Believe that laws should be strictly enforced.		
5	(Ahts)	Am hard to satisfy.				
5	(Ebtoaanmp)	Enjoy being thought of as a normal mainstream person.				

Note. The component labels refer to the theoretical underlying common cause of the redundancies between items. Component labels without a specific name are single items that were not combined into a common cause component. The node labels in Figure 9 are in parentheses. * represents a component's label that should be interpreted in the opposite direction (i.e., reverse coded).

Dimensionality

After the redundancy analysis, the components were analyzed using EGA. The default for EGA is to use the GLASSO network estimation method with the Walktrap community detection algorithm. Based on the dimensionality simulation (Chapter III), it appears that the Louvain algorithm may produce more optimal results. It's important to note, however, that the Walktrap algorithm is among the most accurate and least biased algorithms, especially when used with polytomous data. In light of the results from the simulation, I used the GLASSO network estimation method and Louvain community detection algorithm to estimate the dimensions of the unique components of the SAPA inventory.

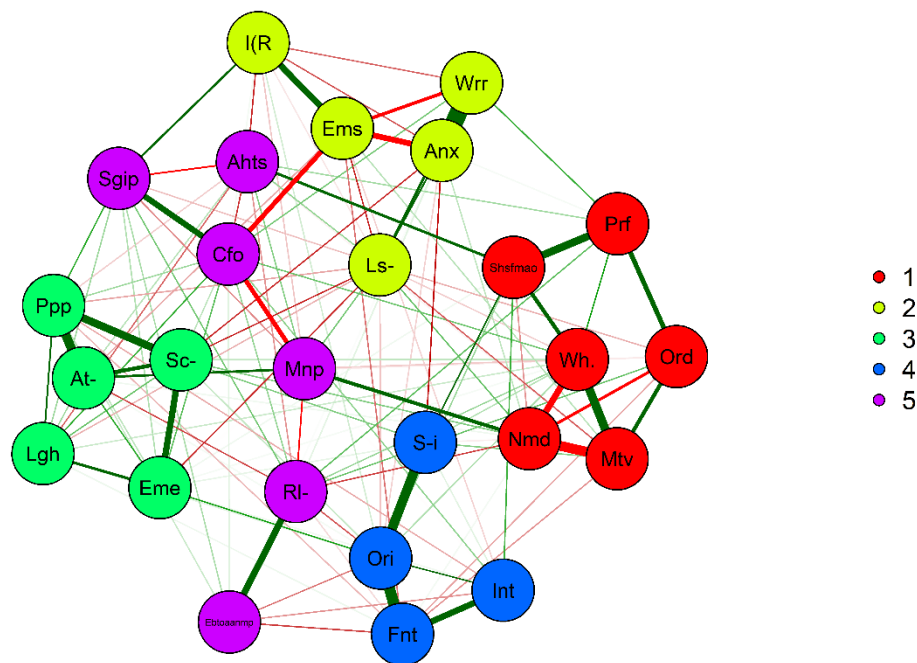


Figure 9. Depiction of the dimensions identified using EGA. The color of the nodes represents the dimensions and the thickness of the lines represent the magnitude of the partial correlations (green = positive; red = negative).

As shown in Figure 9, EGA identified five factors whose item content are displayed in Table 4. When reviewing the item content of these dimensions (Table 4), these factors directly correspond to the FFM: dimensions 1, 2, 3, 4 and 5 reflect conscientiousness, neuroticism, extraversion, openness to experience, and agreeableness, respectively. Although these components were empirically comprised of FFM items, this finding is an empirical validation of the effectiveness of the redundancy and dimensionality analyses.

Loadings

The network loadings were computed using the standardized loadings described in Chapter IV. Below is a table for each personality component in each dimension.

Table 5. Network loadings across the five dimensions identified by EGA.

	1 (Conscientiousness)	2 (Neuroticism)	3 (Extraversion)	4 (Openness to Experience)	5 (Agreeableness)
Work hard.	0.34	-0.011	0.027	0.04	0.063
Neglect my duties.	-0.314	0.047	0.003	0.007	-0.115
Perfectionist	0.241	0.045	0	0	0.052
Orderly	0.222	-0.017	-0.004	-0.024	0.036
Motivated	0.322	-0.05	0.028	0.06	0.003
Set high standards for myself and others.	0.205	0	0.003	0.144	-0.064
Worrier	0.04	0.385	0	0	0.034
Anxious	0.015	0.405	-0.048	-0.076	0.006
Low self-esteem	-0.091	0.174	-0.135	-0.015	-0.065
Irritable (R)	0	-0.178	0	0.012	0.12
Emotional stability	0.015	-0.32	-0.02	-0.049	-0.085
People person	0	-0.019	0.326	-0.041	0.088
Attention-seeking	-0.007	0.019	0.284	0.003	-0.116
Social-efficacy	0.038	-0.094	0.341	0.048	0.014
Laugher	0.01	-0.017	0.183	0.012	0.096
Express myself easily.	0.005	-0.045	0.24	0.07	0.051
Original ideation	0.037	-0.032	0.074	0.409	-0.066
Fantasy	-0.044	0.03	0.02	0.302	-0.048
Introspective	0.04	0.01	0.022	0.194	0.046
Self-assessed intelligence	0.095	-0.053	0.033	0.17	0
Concerned for others	0.03	0.119	0.119	0.029	0.248
Manipulative	-0.087	0.042	0.066	-0.001	-0.186
Sees good in people	0	-0.076	0.088	0	0.223
Am hard to satisfy.	0.083	0.069	-0.062	0	-0.146
Enjoy being thought of as a normal mainstream person.	0	0.001	0.012	-0.115	0.126
Rule-follower	0.123	0.011	-0.042	-0.054	0.23

Note. Grey boxes indicate the loadings of the dimension that correspond to each component's respective dimension. Bold values indicate the largest loading for each component. Components labeled with reverse coding are denoted with (R).

To verify the network loadings were in proper orientation with traditional factor loadings, I used Spearman's correlation between the two. The Spearman's correlation (r

= 0.87) mirrored the simulation results and suggesting that the network loadings are largely redundant with the traditional factor loadings.

When examining the network loadings matrix, there were a few things worth noting. First, the network loadings were much smaller than the loadings of a traditional factor loading matrix. The largest loading is 0.409 for the *original ideation* component in the openness to experience dimension. By traditional factor analysis standards, this is a weak factor loading. This difference in the magnitude of the loadings is due to the association measure underlying the computation of the loading—that is, partial correlations vs. zero-order correlations. The network loadings thus represent partial correlation loadings, meaning that 0.409 is actually a very large loading.

Second, the network loading matrix has particularly small cross-loadings, including some loadings that are zero. Many of the small cross-loadings are small not just by traditional factor analysis standards but also partial correlation standards. This is because of the network estimation where many pairwise correlations are shrunk to zero, leaving many nodes not connected to other nodes. Therefore, if a node (component) is not connected to any nodes in another dimension, there is no loading for that node in the dimension. This shrinkage also affects the size of the cross-loadings by making most cross-dimension connections small, resulting in lower loadings. Much like standard factor analysis, it's often useful to remove small loadings from the matrix to make the loading matrix more interpretable. Below is a table removing loadings less than or equal to .10.

Table 6. Network loadings across the five dimensions identified by EGA with small loadings (less than or equal .10) removed.

	1 (Conscientiousness)	2 (Neuroticism)	3 (Extraversion)	4 (Openness to Experience)	5 (Agreeableness)
Work hard.	0.34				
Neglect my duties.	-0.314				-0.115
Perfectionist	0.241				
Orderly	0.222				
Motivated	0.322				
Set high standards for myself and others.	0.205			0.144	
Worrier		0.385			
Anxious		0.405			
Low self-esteem		0.174	-0.135		
Irritable (R)		-0.178			0.12
Emotional stability		-0.32			
People person			0.326		
Attention-seeking			0.284		-0.116
Social-efficacy			0.341		
Laugher			0.183		
Express myself easily.			0.24		
Original ideation				0.409	
Fantasy				0.302	
Introspective				0.194	
Self-assessed intelligence				0.17	
Concerned for others		0.119	0.119		0.248
Manipulative					-0.186
Sees good in people					0.223
Am hard to satisfy.					-0.146
Enjoy being thought of as a normal mainstream person.				-0.115	0.126
Rule-follower	0.123				0.23

Note. Grey boxes indicate the loadings of the dimension that correspond to each component's respective dimension. Components labeled with reverse coding are denoted with (R).

Finally, when looking at Table 6, the loading matrix becomes much clearer and the patterns of which components are most associated with each dimension is obvious

(and much closer to a simple structure). The component of *low self-esteem*, for example, was negatively associated with the extraversion dimension. The *concerned for others* component was positively related to both neuroticism and extraversion. One peculiar cross-loading is the component *set high standards for myself and others* with openness to experience. From the loading matrix, it's difficult to discern why this component would be related to openness to experience. The network, however, provides greater insight into this relation, specifically the *set high standards for myself and others* (Shsfmao) component is connected to the *self-assessed intelligence* (S-i) and *introspective* (Int) components (Figure 9).

Summary

This example strings together the three simulations presented in this dissertation, demonstrating their respective contributions to assessment validation. The SAPA inventory represented an optimal dataset for the example because it offered a large sample, substantial redundancy between items, and had an empirically derived hierarchical structure. This hierarchical structure offered an a priori expectation of the results, enabling an objective criterion for the effectiveness of the analyses. In short, the redundancy analysis identified 26 unique components in the SAPA personality network, which largely corresponded to the 27 lower-order dimensions identified in previous empirical work (Condon, 2018). The dimensionality analysis identified 5 dimensions from the components that corresponded to the FFM. Finally, the network loadings were shown to be redundant with traditional factor analysis loadings when estimating five factors. Overall, these network-driven analyses for assessment form a theoretically

(simulations) and empirically (SAPA inventory) supported approach for the validation of assessment instruments.

CHAPTER VI

CONCLUSIONS

This dissertation sought to systematically and empirically investigate the conceptual framework for the validation of assessment instruments proposed by Christensen, Golino, and Silvia (under review). Three simulation studies were performed to evaluate components of detecting node redundancy, identifying dimensionality, and computing network loadings. An empirical example that demonstrated how these three analyses can be applied to real-world data. Taken together, these approaches were validated by the simulation and empirical results.

For the node redundancy and network loadings, novel approaches were first conceptually developed and then evaluated in simulations. In the node redundancy simulation, the weighted topological overlap and partial correlation approaches for node redundancy worked best when paired with the adaptive alpha multiple comparison correction. One approach did not appear to be superior to the other; however, the partial correlation approach boasted slightly better performance on the sensitivity and specificity measures. This finding supports both perspectives of psychometric networks and latent variable models. The weighted topological overlap measure provides a redundancy approach that aligns with the network perspective, while the partial correlation approach aligns with the latent variable perspective.

The implications of the redundancy analysis should be far reaching for network analysts in psychology. Identifying unique components of psychological attributes is

essential for understanding the processes that underlie them as well as valid measurement of the attribute itself (Hallquist et al., 2019). Assessment instruments in personality, for example, are often redundant, which may make for more reliable measures but may also decrease the validity of the measurement (McCrae & Möttus, 2019). Reducing redundancy allows researchers to assess personality traits more broadly, often without losing reliability (McCrae, 2015), thereby maximizing both efficiency and information gathered from participants (McCrae & Möttus, 2019). Therefore, reducing redundancy should not just have a role in psychometric network assessment but the development of assessment instruments as whole.

The dimensionality simulation provided the most comprehensive psychometric evaluation of community detection algorithms for the estimate dimensions from factor structures to date. This simulation evaluated several open-source community detection algorithms in the *igraph* package in R, finding that some algorithms work better than others when paired with the current state-of-the-art network estimation algorithm, the GLASSO. These results shed light on current practices and offer avenues for the way forward. The most common approach for dimensionality from the network perspective has been EGA, which uses the GLASSO network estimation method and Walktrap community detection algorithm (Golino & Epskamp, 2017; Golino et al., in press).

This simulation, for example, was the first to evaluate the EGA approach in polytomous data, and the results mirror previous simulation studies that examined continuous and dichotomous data (Golino et al., in press). Notably, the dimensionality simulation differed from the previous by having a broader distribution of cross-loadings.

The correspondence between the results suggests that EGA is not severely affected by larger cross-loadings. Moreover, polytomous data was evaluated for the first time, which also demonstrated that EGA was not affected by the number of response options. This stands in contrast to parallel analysis, which was designed for and works better with continuous data (Garrido et al., 2013; Horn, 1965).

In regard to the community detection algorithms, there is good evidence that the Louvain and Fast-greedy algorithm are worthwhile considerations for adaption into the EGA approach. Because the two algorithms are relatively redundant and demonstrate similar performance, preference for the Louvain algorithm should be given because it also provides hierarchical or “multi-level” structuring of dimensions. Such hierarchical structuring would be important for determining different levels of taxonomies that exist in assessment instruments and particularly in personality questionnaires (Christensen et al., under review). Moreover, it also provides another method for the results of EGA to be compared to such that the best fitting or most theoretically consistent model can be chosen based on the results (e.g., Golino et al., under review).

Finally, in the network loadings simulation, the adapted node strength measure was derived from previous simulation evidence showing that node strength is relatively redundant with CFA factor loadings (Hallquist et al., 2019). My adapted measure split node strength between dimensions identified by EGA and standardized each dimension’s values. This approach provided accurate recovery of the ordering of factor loadings in the simulation (determined by Spearman’s correlation) and was comparable to factor analysis loadings in the empirical example. This suggests that network loadings are not only

accurate, but they are relatively redundant to factor loadings. This result opens up several avenues for future work related to measurement invariance and network scores.

A key point moving forward will be to establish norms for what constitutes a small, moderate, and large network loading. It seems fair to suggest that effect sizes for multiple regression may hold for network loadings; however, the f^2 metric is likely to be more confusing for practical researchers than not (Cohen, 1992). Instead, using effect sizes that *typically* translate from these f^2 might be more interpretable; specifically, effect sizes of .10, .30, and .50 corresponding to small, moderate, and large effect sizes, respectively. Although this issue requires further examination, I suspect that these guidelines are reasonable enough for researchers to find them useful (e.g., Table 6).

In sum, this dissertation aimed to move towards an expanded role of psychometric network models in psychometric assessment. Based on the three simulation studies and empirical example, it appears that network models are not just a novel measurement perspective but rather an effective approach for the validation of assessment instruments. Some researchers may question the novelty these methods and ask what they provide over and above traditional psychometric approaches. These researchers have a valid point: the redundancy analysis could be performed using more traditional metrics, while factor analysis and loadings have long been established in traditional psychometrics. To this point, the additional information that these methods provide may appear to be minimal.

The methods in this dissertation, however, were not introduced to reinvent the wheel but rather to gather evidence for psychometric applications from the network

perspective. It is, after all, the substantive interpretation of classical test theory that differentiates itself from modern test theory (Borsboom, Mellenbergh, & van Heerden, 2004). It is therefore not a matter of statistical equivalency (van Bork et al., 2019) but a matter of validity: how and why do observed variables co-occur and emerge as psychological attributes? The novelty is therefore in the perspective that psychometric network models provide (e.g., reducing redundancy in assessment instruments; Christensen et al., under review). To date, these models have lacked the tools to validate assessment instruments from their perspective. This dissertation takes one step towards that goal.

REFERENCES

- Barabási, A.-L. (2012). The network takeover. *Nature Physics*, 8, 14–16.
- 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, 289–300.
- Bernaards, C. A., & Jennrich, R. I. (2005). Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement*, 65, 676–696.
- Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008, P10008.
- Bollmann, S., Heene, M., Küchenhoff, H., & Bühner, M. (2015). *What can the real world do for simulation studies? A comparison of exploratory methods*. (Technical Report 181). Retrieved from Department of Statistics, University of Munich <https://epub.ub.uni-muenchen.de/24518>
- Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology*, 64, 1089–1108.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, 110, 203–219.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111, 1061–1071.

- Bringmann, L. F., Elmer, T., Epskamp, S., Krause, R. W., Schoch, D., Wichers, M., . . . Snippe, E. (2019). What do centrality measures measure in psychology networks? *Journal of Abnormal Psychology, 128*, 892–903.
- Bringmann, L. F., & Eronen, M. I. (2018). Don't blame the model: Reconsidering the network approach to psychopathology. *Psychological Review, 125*, 606–615.
- Champely, S. (2018). pwr: Basic functions for power analysis. R package version 1.2-2.
- Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika, 95*, 759–771.
- Christensen, A. P., Golino, H. F., & Silvia, P. J. (under review). A psychometric network perspective on the validity and validation of personality trait questionnaires. *PsyArXiv*.
- Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E, 70*, 066111.
- Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.
- Condon, D. M. (2018). The SAPA personality inventory: An empirically-derived, hierarchically-organized self-report personality assessment model. *PsyArXiv*.
- Costantini, G., Richetin, J., Preti, E., Casini, E., Epskamp, S., & Perugini, M. (2019). Stability and variability of personality networks. A tutorial on recent developments in network psychometrics. *Personality and Individual Differences, 136*, 68–78.
- Cramer, A. O. J. (2012). Why the item “23+1” is not in a depression questionnaire: Validity from a network perspective. *Measurement: Interdisciplinary Research &*

Perspective, 10, 50–54.

- Cramer, A. O. J., van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., . . . Borsboom, D. (2012). Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality*, 26, 414–431.
- Csárdi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, 1695, 1–9.
- Dablander, F., & Hinne, M. (2019). Node centrality measures are a poor substitute for causal inference. *Scientific Reports*, 9, 6846.
- Danon, L., Díaz-Guilera, A., Duch, J., & Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005, P09008.
- De Beurs, D., Fried, E. I., Wetherall, K., Cleare, S., O'Connor, D. B., Ferguson, E., . . . O'Connor, R. C. (2019). Exploring the psychology of suicidal ideation: A theory driven network analysis. *Behaviour Research and Therapy*, 120, 103419.
- Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, 64, 1–34.
- DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Thousand Oaks, CA: SAGE Publications.
- Dinic, B. M., Wertag, A., Tomašević, A., & Sokolovska, V. (in press). Centrality and redundancy of the Dark Tetrad traits. *Personality and Individual Differences*.
- Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D.

- (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, 48, 1–18.
- Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*, 23, 617–634.
- Epskamp, S., Maris, G., Waldorp, L. J., & Borsboom, D. (2018a). Network psychometrics. In P. Irwing, D. Hughes, & T. Booth (Eds.), *The Wiley handbook of psychometric testing, 2 volume set: A multidisciplinary reference on survey, scale and test development*. New York, NY: Wiley.
- Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika*, 82, 904–927.
- Epskamp, S., Waldorp, L. J., Möttus, R., & Borsboom, D. (2018b). The Gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research*, 53, 453–480.
- Fan, Y., Li, M., Zhang, P., Wu, J., & Di, Z. (2007). Accuracy and precision of methods for community identification in weighted networks. *Physica A: Statistical Mechanics and its Applications*, 377, 363–372.
- Flora, D. B., & Flake, J. K. (2017). The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. *Canadian Journal of Behavioural Science*, 49, 78–88.
- Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486, 75–174.

- Foygel, R., & Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, & A. Culotta (Eds.), *Advances in neural information processing systems 23* (pp. 604–612). Vancouver, CA: Neural Information Processing Systems.
- Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35–41.
- Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *PsyArXiv*.
- Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9, 432–441.
- Furr, R. M. (2017). *Psychometrics: An introduction* (3rd ed.). Thousand Oaks, CA: SAGE Publications.
- Garcia-Garzón, E., Abad, F. J., & Garrido, L. E. (2019). Improving bi-factor exploratory modeling. *Methodology*, 15, 45–55.
- Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at Horn’s parallel analysis with ordinal variables. *Psychological Methods*, 18, 454–474.
- Gates, K. M., Henry, T., Steinley, D., & Fair, D. A. (2016). A Monte Carlo evaluation of weighted community detection algorithms. *Frontiers in Neuroinformatics*, 10, 45.
- Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99, 7821–7826.
- Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality*

Psychology in Europe, 7, 7–28.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40, 84–96.

Goldberg, L. R., & Saucier, G. (2016). The Eugene-Springfield community sample: Information available from the research participants (Tech. Rep. No. 56-1). *Eugene, Oregon: Oregon Research Institute*.

Golino, H. F., & Christensen, A. P. (2020). EGAnet: Exploratory Graph Analysis – A framework for estimating the number of dimensions in multivariate data using network psychometrics. R package version 0.9.2.

Golino, H. F., & Demetriou, A. (2017). Estimating the dimensionality of intelligence like data using Exploratory Graph Analysis. *Intelligence*, 62, 54–70.

Golino, H. F., & Epskamp, S. (2017). Exploratory Graph Analysis: A new approach for estimating the number of dimensions in psychological research. *PLoS ONE*, 12, e0174035.

Golino, H. F., Moulder, R., Shi, D., Christensen, A. P., Garrido, L. E., Nieto, M. D., ... Boker, S. M. (under review). Entropy fit indices: New fit measures for assessing the structure and dimensionality of multiple latent variables. *PsyArXiv*.

Golino, H. F., Shi, D., Christensen, A. P., Garrido, L. E., Nieto, M. D., Sadana, R., . . . Martinez-Molina, A. (in press). 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., Wright, A. C. G., & Molenaar, P. C. M. (2019). Problems with centrality measures in psychopathology symptom networks: Why network psychometrics cannot escape psychometric theory. *Multivariate Behavioral Research*.
- Haslbeck, J. M., & Waldorp, L. J. (2018). How well do network models predict observations? On the importance of predictability in network models. *Behavior Research Methods*, 50, 853–861.
- Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, 66, 393–416.
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179–185
- Hubley, A. M., Zhu, S. M., Sasaki, A., & Gadermann, A. M. (2014). Synthesis of validation practices in two assessment journals: Psychological Assessment and the European Journal of Psychological Assessment. In B. Zumbo & E. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 193–213). Cham, CH: Springer.
- Kruis, J., & Maris, G. (2016). Three representations of the Ising model. *Scientific Reports*, 6, srep34175.
- Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E*, 80, 056117.
- Lancichinetti, A., & Fortunato, S. (2012). Consensus clustering in complex networks.

- Scientific Reports*, 2, 336.
- Lauritzen, S. L. (1996). *Graphical models*. Oxford, UK: Clarendon Press.
- Lee, K., & Ashton, M. C. (2018). Psychometric properties of the HEXACO-100. *Assessment*, 25, 543–556.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3, 635–694.
- Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L. J., . . .
- Maris, G. (2018). An introduction to network psychometrics: Relating Ising network models to item response theory models. *Multivariate Behavioral Research*, 53, 15–35.
- Massara, G. P., Di Matteo, T., & Aste, T. (2017). Network filtering for big data: Triangulated maximally filtered graph. *Journal of Complex Networks*, 5, 161–178.
- McCrae, R. R. (2015). A more nuanced view of reliability: Specificity in the trait hierarchy. *Personality and Social Psychology Review*, 19, 97–112.
- McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52, 81–90.
- McCrae, R. R., & Möttus, R. (2019). What personality scales measure: A new psychometrics and its implications for theory and assessment. *Current Directions in Psychological Science*, 28, 415–420.

- Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103, 8577–8582.
- Newman, M. E. J. (2010). *Networks: An introduction*. Oxford, UK: Oxford University Press.
- Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69, 026113.
- Nowick, K., Gernat, T., Almaas, E., & Stubbs, L. (2009). Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proceedings of the National Academy of Sciences*, 106, 22358–22363.
- Pérez, M. E., & Pericchi, L. R. (2014). Changing statistical significance with the amount of information: The adaptive α significance level. *Statistics & Probability Letters*, 85, 20–24.
- Pons, P., & Latapy, M. (2006). Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, 10, 191–218.
- R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76, 036106.
- Reichardt, J., & Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E*, 74, 016110.
- Revelle, W. (2018). psych: Procedures for personality and psychological research. R

- package version 1.9.12.
- Revelle, W. (2019). psychTools: Tools to accompany the ‘psych’ package for psychological research. R package version 1.9.12.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software*, 48, 1–36.
- Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105, 1118–1123.
- Rubinov, M., & Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52, 1059–1069.
- Sass, D. A., & Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research*, 45, 73–103.
- Schmitt, T. A., & Sass, D. A. (2011). Rotation criteria and hypothesis testing for exploratory factor analysis: Implications for factor pattern loadings and interfactor correlations. *Educational and Psychological Measurement*, 71, 95–113.
- Schmittmann, V. D., Cramer, A. O. J., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology*, 31, 43–53.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.
- van Bork, R., Rhemtulla, M., Waldorp, L. J., Kruis, J., Rezvanifar, S., & Borsboom, D. (2019). Latent variable models and networks: Statistical equivalence and

- testability. *Multivariate Behavioral Research*.
- van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*, 842–861.
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York, NY: Springer.
- Ward, J. H. (1963). Hierarchical grouping to optimize on objective function. *Journal of the American Statistical Association*, *58*, 236–244.
- Williams, D. R. (2019). *GGMnonreg: Estimate non-regularized Gaussian graphical models*. R package version 1.0.0.
- Williams, D. R., & Rast, P. (2019). Back to the basics: Rethinking partial correlation network methodology. *British Journal of Mathematical and Statistical Psychology*.
- Williams, D. R., Rhemtulla, M., Wysocki, A. C., & Rast, P. (2019). On nonregularized estimation of psychological networks. *Multivariate Behavioral Research*, *54*, 719–750.
- Yang, Z., Algesheimer, R., & Tessone, C. J. (2016). A comparative analysis of community detection algorithms on artificial networks. *Journal of Computational and Theoretical Nanoscience*, *6*, 30750.
- Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*,

1-45.