MULTIVARIATE BEHAVIORAL RESEARCH
2021, VOL. 56, NO. 2, 243–248

OPEN ACCESS

On the Importance of Estimating Parameter Uncertainty in Network Psychometrics: A Response to Forbes et al. (2019)

Eiko I. Fried[a], Claudia D. van Borkulo[b], and Sacha Epskamp[c]

[a]Department of Clinical Psychology, Leiden University, Leiden, Netherlands; [b]Psychological Methods, Universiteit van Amsterdam Faculteit der Maatschappij- en Gedragswetenschappen, Amsterdam, Netherlands; [c]Department of Psychological Methods, Universiteit van Amsterdam, Amsterdam, Netherlands

ABSTRACT

In their recent paper, Forbes et al. (2019; FWMK) evaluate the replicability of network models in two studies. They identify considerable replicability issues, concluding that "current 'state-of-the-art' methods in the psychopathology network literature [...] are not well-suited to analyzing the structure of the relationships between individual symptoms". Such strong claims require strong evidence, which the authors do not provide. FWMK identify low replicability by analyzing point estimates of networks; contrast low replicability with results of two statistical tests that indicate higher replicability, and conclude that these tests are problematic. We make four points. First, statistical tests are superior to the visual inspection of point estimates, because tests take into account sampling variability. Second, FWMK misinterpret the statistical tests in several important ways. Third, FWMK did not follow established recommendations when estimating networks in their first study, underestimating replicability. Fourth, FWMK draw conclusions about methodology, which does not follow from investigations of data, and requires investigations of methodology. Overall, we show that the "poor replicability "observed by FWMK occurs due to sampling variability and use of suboptimal methods. We conclude by discussing important recent simulation work that guides researchers to use models appropriate for their data, such as nonregularized estimation routines.

KEYWORDS

gaussian graphical model; network model; regularization; replicability

In their paper entitled "Quantifying the reliability and replicability of psychopathology network characteristics", Forbes et al. (2019)—from here on FWMK—conducted two studies. First, they estimated Gaussian Graphical Models (GGMs) of 16 depression and anxiety symptoms in two waves of data from an observational study ($n = 403$) one week apart. Second, they re-analyzed GGMs of 16 posttraumatic stress disorder symptoms across 4 datasets reported by Fried et al. (2018).

In both studies, FWMK highlight replicability issues of network models, and conclude that the "low reliability of partial correlations and high rates of false negatives related to LASSO regularization in such data suggest that the current 'state-of-the-art' methods in the psychopathology network literature [...] are not well-suited to analyzing the structure of the relationships between individual symptoms of mental illness" (p. 14).

To assess replicability, FWMK compare *point estimates* of network parameters, and contrast their results of low replicability with the results of two *statistical tests* that indicate higher replicability. The two tests are part of two R packages, which were developed with the express purpose so that researchers stop overinterpreting point estimates when conducting comparisons. The first package is *bootnet* (Epskamp et al., 2018), which performs nonparametric and subsampling bootstraps to assess the stability and accuracy of estimated networks. The second package is the *Network Comparison Test* (NCT; van Borkulo et al., 2017), which uses permutations to test if two samples feature different underlying network structures.

We agree with FWMK that there are numerous challenges in this emerging field, both of substantive and statistical nature. However, we disagree with some of the conclusions the authors draw, and focus on four issues below.

© 2020 The Author(s).
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

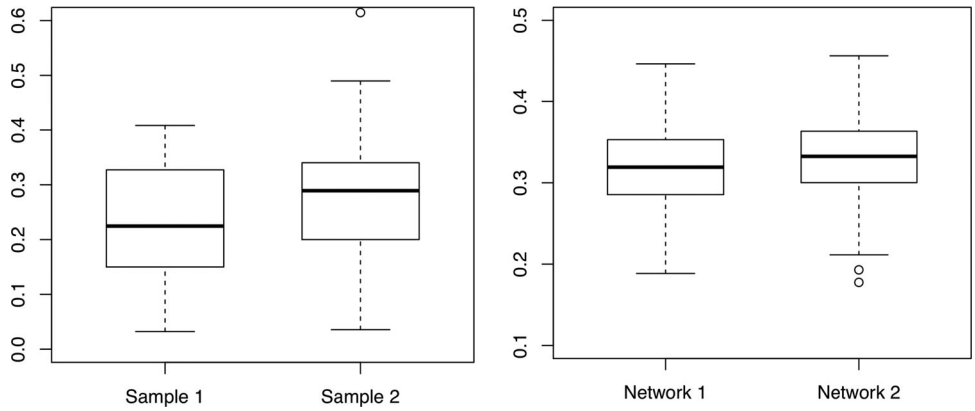

Figure 1. Sampling variability. Left: Neuroticism levels across two samples. Right: Bootstrapped edge weights across two networks (PHQ1—PHQ2 in Figure 2).

Sampling variability

Statistical tests aim to avoid overinterpretation of point estimates. Suppose we want to compare standardized neuroticism levels across two small samples of 30 participants each, with means .24 (SD = 1.1) and .28 (SD = 1.2). Point estimates clearly differ from each other, but a *t*-test indicates that this difference is not significant; $t(57.9)=1.6$, $p = 0.12$ (Figure 1, left).

Or suppose that point estimates of the factor loadings of an item x1 differ slightly across two factor models estimated in two samples, but measurement invariance tests indicate that the null hypothesis that the models are the same cannot be rejected (Meredith, 1993). Researchers would rely on the results of the statistical tests—not the point estimates—to draw inferences. This also applies to network models: point estimates of two specific edges might be different (Figure 1, right), but whether this difference is meaningful can only be determined in the light of sampling variability, which requires statistical tests to find out if a difference in two parameter estimates is more extreme than one would expect under the null hypothesis. *Bootnet* and *NCT* were developed with the purpose to conduct such tests—*bootnet* for comparing edge weights within one network, *NCT* for comparing edge weights across different networks.

In their paper, FWMK analyze and visualize differences in point estimates in detail, and show that statistical tests provided by *bootnet* and *NCT* arrive at different conclusions than the authors' inferences of point estimates. For example, they state: "all NCT results indicated that the depression and anxiety symptom networks had no significant differences when *in fact* they had a multitude of differences" (p. 15, our highlight). This inference is no different than concluding that the *t*-test for neuroticism in Figure 1 reaches a non-significant result when *in fact* point estimates of neuroticism differ across samples (Figure 1)—it ignores sampling variability. Of note, FWMK only apply this logic to parameters derived from network models, not to other statistical parameters. For instance, they state that polychoric correlation matrices of the longitudinal data "could be constrained to equality without affecting model fit ([…] $p = 0.666$)". In other words, while the point estimates of the correlations were not exactly identical across samples, a statistical test provided evidence that this is likely the result of sampling variability, similar to the neuroticism example in Figure 1. This interpretation of point estimates contrasts with the authors' interpretation of point estimates of network models in the remainder of the paper. While we focus this commentary on the two waves of depression and anxiety data, the problem of overinterpreting point estimates also applies to the analysis of the PTSD datasets.

Misinterpretation of bootnet and NCT

Three conclusions of FWMK regarding *bootnet* and *NCT* do not follow from their investigation. First, a central claim of FWMK is that "existing suite of methods tended to suggest that the networks were accurately estimated" (p. 12), and that "the interpretation guidelines for bootnet results err toward indicating stability and interpretability in networks" (p. 15). This is not true: *bootnet* results clearly indicate lack of stable estimates for study 1. The *CS* coefficient—a metric *bootnet* provides researchers an idea about the stability of the order of centrality estimates—was 0.13 for the depression and anxiety networks, implying that the centrality order is unstable and should thus not be interpreted (Epskamp et al., 2018). FWMK write: "the CS-coefficient was below the minimum recommended cutoff [for accurate estimation] at both

Table 1. Replication of depression and anxiety networks based on polychoric and Spearman correlations.

Replication metric proposed by FWMK	GGMs based on polychoric correlations	GGMs based on Spearman correlations
Edges of N1 replicated with same sign in N2	55 (70.5%)	**65 (81.3%)**
Edges of N1 that switch sign in N2	4 (5.1%)	**0 (0%)**
Absent edges of N1 replicated in N2	23 (54.8%)	**28 (70%)**
Bridging edges of N1 replicated with same sign in N2	23 (56.1%)	**23 (67.6%)**
Bridging edges of N1 switched sign in N2	3 (7.3%)	**0 (0%)**
Absent bridging edges of N1 replicated in N2	14 (63.6%)	**22 (75.9%)**

Note: Higher replicability marked bold. GGM: Gaussian Graphical Model; N1: network estimated on first time point in depression and anxiety data; N2: network estimated on second time point. Total number of edges per network: N1 (Spearman) = 80, N1 (polychoric) = 78, N2 (Spearman) = 78, N2 (polychoric) = 42.

waves (CS(0.7) = .13), which notably represents the only clear guideline available for interpreting *bootnet* output" (p. 9). In other words, FWMK 1) identified only one clear benchmark for accurate estimation in *bootnet*, the CS-coefficient; 2) demonstrate that this one benchmark is below the recommended threshold for accurate estimation; and 3) conclude nonetheless that *bootnet* guidelines suggest networks were estimated accurately. Thus, their conclusion that *bootnet* errs toward indicating stability and interpretability does not follow from the evidence FWMK present.

Second, FWMK correctly state in the manuscript that "the authors of the [*bootnet*] package emphasize that the CIs do not represent significance tests" (p. 3). However, the authors then continue to interpret *bootnet* results in precisely this way. They consider an edge to be "bootnet-accurate"—a novel concept invented by FWMK that has not been used before in the literature—if its bootstrapped CI does not contain zero, irrespective of how large the bootstrapped CI is. This definition is inconsistent with common definitions of parameter accuracy: a parameter estimate is accurate when it has small confidence regions, *regardless* of whether these confidence regions contain zero. For example, a standardized parameter that is reliably estimated to be non-zero but fluctuates between 0.1 and 0.8 in bootstrap samples indicates the estimate is not accurate, but falls under the definition of accurate according to FWWK. A related concern is that regularization already pulls parameters to zero. Bootstrapping to test the null-hypothesis that an edge weight is different from zero after using regularization leads to a considerable drop in statistical power to detect if an edge weight is nonzero, and the probability increases that an estimated edge weight is exactly zero, which in turn leads to a loss of statistical power to detect an edge to be nonzero using such a confidence region (i.e. increased type 2 error rate). It is not surprising that when comparing many such low powered tests, many edges will not be considered different from zero in both datasets—this is expected by chance alone. FWMK seem aware of both problems, as is evident from the quote above and footnote 5, but used "bootnet-accurate" to inform their most important conclusions nonetheless.

Third, with regard to the *NCT*, FWMK state that "The simulations in van Borkulo et al. (under review) indicated that the *NCT* omnibus test should be adequately powered to compare these networks" (p. 15). This is not correct: the simulation study investigated continuous, independent data, not ordinal, dependent (i.e., longitudinal) data as used by the authors. It is therefore unclear how much power FWMK have to detect differences. In this case, the authors obtain a non-significant *p*-value, and argue that there is something wrong with the method because it arrives at different conclusions than the comparison of parameter point estimates. We refer back to our discussion above on sampling variation for the logic of this argument, or to measurement invariance in structural equation models (Meredith, 1993). Comparing factor models across two samples is not done by visualizing differences in point estimates of factor loadings, but by using statistical tests to see if parameters differ more than would be expected under sampling variability.

Problems of polychoric correlations in skewed ordinal data

The results of study 1, including the switching of signs in edges, are partly due to the specific estimation algorithm used: regularized GGM estimation based on polychoric correlations. As described in the tutorial paper cited by FMWK (Epskamp & Fried, 2018), polychoric correlations can lead to biased estimates in case of skewed ordinal data, especially in smaller samples: "When the sample size is relatively low, some cells in the item by item frequency table can be low or even zero […]. The estimation of polychoric correlations is […] biased whenever an expected frequency is too small (i.e., below 10; Olsson, 1979). Low frequencies can thus lead to biased polychoric correlations, which can compound to large biases in the estimated partial correlations.

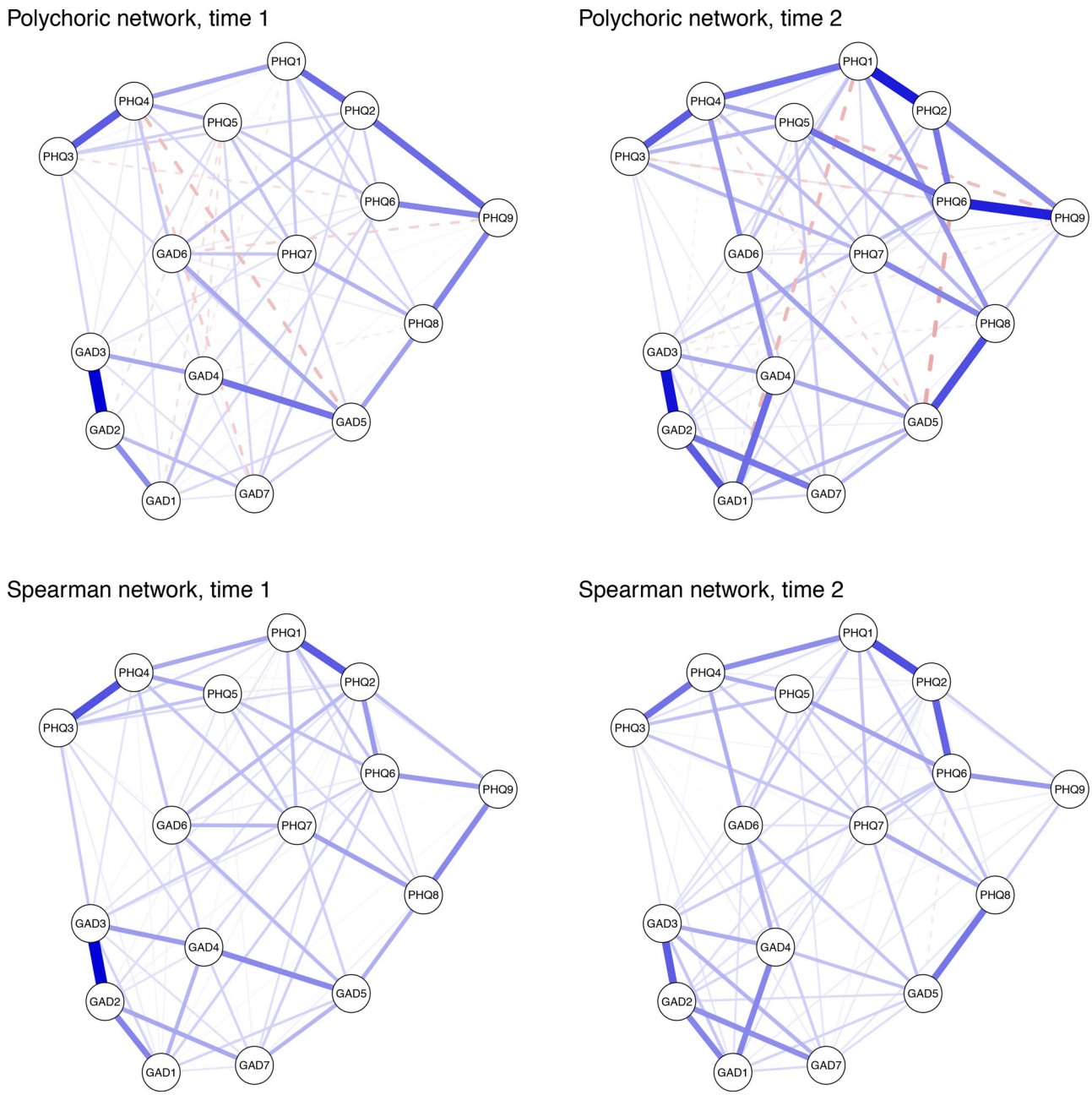

Figure 2. Networks estimated on timepoints 1 and 2 from the depression and anxiety data used by FMWK. Top row: GGMs estimated by FWMK, based on polychoric correlations. Bottom row: GGMs based on Spearman correlations.

Another situation in which one might obtain low frequencies is when the scores are highly skewed (Rigdon & Ferguson, 1991), which unfortunately often is the case in psychopathology data. Again, the network based on polychoric correlations should be compared with a network based on Spearman correlations" (p. 14).

Such estimation problems can be expected in the data of the FMWK: for timepoint 1, all 16 items are skewed (mean skewness = 1, range 0.44 to 2.59; mean kurtosis = 0.54, range −0.82 to 6.33), with an average item mean of 0.86 (range 0.29 to 1.31) on a scale from 0 to 4. As expected, such data shows floor effects, with a correlation of 0.86 between item means and standard deviations. Timepoint 2 shows even more pronounced skew and floor effects. Following the recommendation of using Spearman correlations to obtain more stable estimates (Epskamp & Fried, 2018), replicability is consistently higher, on 6 out of 6 metrics calculated by FWMK (Table 1). This highlights that the results from FMWK appear to be specific to the estimation method used rather than generalize to network estimation methods in general (Figure 2).

Replicability is a substantive, not a statistical phenomenon

Finally, we want to highlight that conclusions about methodology cannot follow from *investigations of data*—they require *investigations of methodology*, which can be done via simulation studies or mathematical proofs (Borsboom et al., 2017, 2018). This means that replications or non-replications of substantive findings cannot be informative about the quality of statistical models. One can fairly criticize researchers for using *t*-tests incorrectly, but one would not criticize *t*-tests *as a statistical tool* because large-scale collaborative projects in psychology that primarily relied on *t*-tests have shown limited replicability (Camerer et al., 2018; Open Science Collaboration, 2015). This is the conclusion FWMK draw about network models: "Poor replicability underpins our concern surrounding the use of these methods" (p. 1).

Conclusion

We show that the "limited reliability of the detailed characteristics of networks" (p. 1) observed by FWMK occurs in part due to sampling variability, and in part because polychoric correlations can be unstable in skewed ordinal data, especially in small samples.

Network models are high-dimensional, multivariate models with many parameters. Exploratory search in this parameter space will come with sampling variability and differences in the performance across specific algorithms. Prior work has discussed these aspects in some detail, and put forward ways to address sampling variability using statistical tests (Epskamp et al., 2018; Epskamp et al., 2016; Fried & Cramer, 2017; Williams & Rast, 2018; Williams et al., 2019). Recent methodological studies have led to an increased understanding of the performance of specific network estimation methods under different conditions. For instance, the regularization algorithm commonly used was specifically designed for sparse underlying network structures. Under dense structures, regularization leads to a higher false positive rate than was previously known (Williams & Rast, 2018; Williams et al., 2019). We welcome such simulation studies that are important methodological contributions to the literature, as they strengthen the knowledge about methodology. Network psychometrics is no different in this regard than psychometrics in general.

Overall, such methodological insights encourage researchers to use models appropriate for their data. For network models, researchers may, under specific circumstances, consider alternative estimation routines that do not rely on regularization, such as non-regularized estimation procedures (e.g., Isvoranu et al., 2019). But rather than to conclude this is a problem for network psychometrics, we conclude that researchers should be aware of the assumptions inherent in the methods such as regularization, and choose estimation methods that are most appropriate for their data and research question.

Finally, it is unclear if e.g. 75% replicated edges imply comparably good or bad model performance, as we are not aware of simulation studies on the expected number of replicated present and absent edges given conditions such as sample size, skewness of data, and number of variables. That is, even if data for two samples come from the same true model, it is unclear how many edges can be expected to replicate—the same way one would not expect for all factor loadings to replicate across two samples in all possible conditions, even if the data generating mechanism is the same factor model across both. We have implemented the function `replicationSimulator` in the *bootnet* package to facilitate research on the topic, which allows researchers to conduct such simulation studies across a range of network estimation methods and compare their performance. Results may differ substantially per estimation method, and the optimal method may depend on the particular needs of the researcher.

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

References

Borsboom, D., Fried, E. I., Epskamp, S., Waldorp, L. J., van Borkulo, C. D., van der Maas, H. L. J., & Cramer, A. O. J. (2017). False alarm? A comprehensive reanalysis of "evidence that psychopathology symptom networks have limited replicability" by Forbes, Wright, Markon, and Krueger (2017). *Journal of Abnormal Psychology*, *126*(7), 989–999. doi:10.1037/abn0000306

Borsboom, D., Robinaugh, D. J., Rhemtulla, M., Cramer, A. O. J., The Psychosystems Group. (2018). Robustness and replicability of psychopathology networks Network. *World Psychiatry*, *17*(2), 143–144. doi:10.1002/wps.20515

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. doi:10.1038/s41562-018-0399-z

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A Tutorial Paper. *Behavior Research Methods*, *50*(1), 195–212. doi:10.3758/s13428-017-0862-1

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

Epskamp, S., Kruis, J., & Marsman, M. (2016). Estimating psychopathological networks: be careful what you wish for. *ArXiv*, *1604.08045*.

Forbes, M., Wright, A. G. C., Markon, K. E., & Krueger, R. (2019). Quantifying the reliability and replicability of psychopathology network characteristics. *Multivariate Behavioral Research*, 1–19. doi:10.1080/00273171.2019.1616526

Fried, E. I., & Cramer, A. O. J. (2017). Moving forward: challenges and directions for psychopathological network theory and methodology. *Perspectives on Psychological Science*, *12*(6), 999–1020. doi:10.1177/1745691617705892

Fried, E. I., Eidhof, M. B., Palic, S., Costantini, G., Huisman-van Dijk, H. M., Bockting, C. L. H., Engelhard, I., Armour, C., Nielsen, A. B. S., & Karstoft, K.-I. (2018). Replicability and generalizability of Posttraumatic Stress Disorder (PTSD) Networks: A cross-cultural multisite study of PTSD symptoms in four trauma patient samples. *Clinical Psychological Science*, *6*(3), 335–351. doi:10.1177/2167702617745092

Isvoranu, A., Guloksuz, S., Epskamp, S., Os, J., Van Borsboom, D. & G. Investigators, (2019). Toward incorporating genetic risk scores into symptom networks of psychosis. *Psychological Medicine*, 1–8.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. Retrieved from http://link.springer.com/article/10.1007/BF02294825 doi:10.1007/BF02294825

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*, 443–460.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. 10.1126/science.aac4716

Rigdon, E. E., & Ferguson, C. E., Jr. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Re-search*, *28*, 491–497.

van Borkulo, C. D., Boschloo, L., Kossakowski, J. J., Tio, P., Schoevers, R. A., Borsboom, D., Waldorp, L. (2017). *Comparing Network Structures on Three Aspects*. doi:10.13140/RG.2.2.29455.38569

Williams, D., & Rast, P. (2018). Back to the basics : Rethinking partial correlation network methodology. *British Journal of Mathematical and Statistical Psychology*, 1–15.

Williams, D., Rhemtulla, M., Wysocki, A. C., & Rast, P. (2019). On nonregularized estimation of psychological networks. *Multivariate Behavioral Research*, *54*(5), 719–750.