Network Replicability & Generalizability: Exploring the Effects of Sampling Variability, Scale Variability, and Node Reliability

Arianne Herrera-Bennett[1] and Mijke Rhemtulla[1]

[1]University of California, Davis

Author Note

Arianne Herrera-Bennett. https://orcid.org/0000-0002-3207-4034

Mijke Rhemtulla. https://orcid.org/0000-0003-2572-2424

We have no known conflict of interest to disclose.

Correspondence concerning this article should be addressed to Arianne Herrera-Bennett, Department of Psychology, University of California, Davis, 1 Shields Avenue, Davis, CA, 95616, USA. Email: aherrerabennett@ucdavis.edu

Abstract

Work surrounding the replicability and generalizability of network models has increased in recent years, prompting debate as to whether network properties can be expected to be consistent across samples. To date, certain methodological practices may have contributed to observed inconsistencies, including the common use of single-item indicators to estimate nodes, and use of non-identical measurement tools. The current study used a resampling approach to systematically disentangle the effects of sampling variability from scale variability when assessing network replicability. Additionally, we explored the extent to which consistencies in network characteristics were improved when precision in node estimation was increased. Overall, scale variability produced less stability in network properties than sampling variability, however under more optimal measurement conditions (i.e. larger sample, greater node precision), discrepancies were markedly reduced. Findings also importantly underscored the value of improving node reliability: Use of multi-item indicators led to denser networks, higher network sensitivity, greater estimates of global strength, and greater levels of consistency in network properties (e.g., edge weights, centrality scores). Taken together, variability in network properties across samples may be less indicative of a lack of replicability, but may arise from poor measurement precision, and/or may reflect properties of the underlying true network model or scale-specific properties. All data and syntax are openly available online (https://osf.io/m37q2/).

Keywords: *network models, replicability, measurement error, sampling variability*

In recent years, the network analysis approach to multivariate psychology has received increased attention (e.g., see Robinaugh et al., 2020), with growing emphasis on research regarding the replicability and generalizability of network models (e.g., Borsboom et al., 2017; Forbes et al., 2017a, 2017b, 2019; Fried et al., 2018; Funkhouser et al., 2020; Jones et al., 2021; van Borkulo et al., 2017; Williams, 2018, 2020; Williams et al., 2020). Inconsistencies in the literature have prompted debate as to the extent to which one can expect network properties to replicate across samples. Although this debate has primarily stemmed from the psychopathology literature, concerns raised are not exclusive to this domain. Network models have been applied to a range of complex phenomena in order to approximate the structural relations of psychological constructs, including attitudes and beliefs (e.g., Dalege et al., 2017; Rodriguez et al., 2019; Zwicker et al., 2020), cognitive processes (e.g., Golino & Epskamp, 2017; Van Der Maas et al., 2017), classroom dynamics (e.g., Abacioglu et al., 2019; Huitsing & Veenstra, 2012), and clinical disorders (e.g., Borsboom et al., 2017; Fried et al., 2018; Funkhouser et al., 2020). In networks, nodes serve to represent a coherent unit of study (e.g., symptoms, behaviors, facets, or some other variable of interest), whereas edges correspond to the estimated relationships, links, or interactions between these measured units.

Use of different metrics to assess replicability has been discussed as one factor contributing to observed inconsistencies between network models (Forbes et al., 2019). For instance, when assessing the constructs of depression, anxiety, and posttraumatic stress disorder (PTSD), Forbes et al. (2019) found that certain bootstrapping methods (e.g., case-dropping approach; Epskamp et al., 2018) and permutation tests (e.g., invariance tests for comparing global strength and network structure; van Borkulo et al., 2017) each pointed toward stability and consistency in network properties across samples; consistent with conclusions offered by

Borsboom et al. (2017) and Fried et al. (2018). In contrast, when authors tallied the number of edges that were estimated consistently (i.e., present and with the same sign) across pairs of networks, these proportions were smaller. In the case of the PTSD data, which consisted of four independent samples each estimating a 16-node network, nearly all edges (i.e., 114 of the 120 possible edges) were estimated as present in at least one of the samples; however, only 39 (or 34.2%) of these were estimated consistently in all four networks. In light of such findings, Forbes et al. (2019) argued that the use of global summary statistics (e.g., total connectivity, global strength scores, or overall correlations between edge weights or node centrality scores) may mask pervasive discrepancies between estimated networks.

Developing and quantifying realistic expectations about consistency in network properties across samples is an important step in evaluating the replicability and generalizability of network models. While some have claimed that inconsistencies in network characteristics between samples undermine the reliability and utility of current network methods and replicability metrics (Forbes et al., 2017a, 2017b, 2019), others have argued that "observed differences across data sets may not only signify meaningful differences between samples (i.e., nonreplication) but they also may signify random sampling variation, poor reliability in measurement, or a variety of other explanations" (Jones et al., 2021, p. 2). In direct response to Forbes et al. (2019), Williams (2020) argued that the utility of descriptive methods, such as the tally approach above, will ultimately hinge on the degree to which *observed* replicability rates can be considered or judged in relation to *expected* levels of replicability for any particular metric. With respect to the PTSD example, the number 34.2% should be interpreted relative to the expected percent of consistent edges across four networks randomly sampled from the same population. Without, however, some *expected* baseline to compare *observed* rates against,

making sense of reported replicability indices becomes much more difficult. To further complicate matters, when it comes to networks specifically, the greater the number of true non-zero edges (e.g., due to the size or sparsity of a network), the more challenging or unrealistic it would be to expect that all edges replicate. Borrowing from Armour et al.'s (2017) data, Williams (2020) estimated only a 66% probability that more than 34% of the edges in a 20-node PTSD network replicate.

The network replicability debate underscores the need to understand how different sources of variability contribute to discrepancies in network properties. In the absence of identical samples and perfectly reliable measures, a core challenge in comparing any set of networks, regardless of domain, is accounting for and distinguishing between factors that introduce noise and variance into the estimation of nodes and edges (e.g., sampling variability, measurement error, contextual variability, population characteristics, etc.). Only then, can one accurately interpret and explain observed discrepancies between sample findings. Failure to do so leaves one open to conflating expected deviations (e.g., due to normal sampling variability) with genuine differences (Jones et al., 2021).

The current project aims to provide an overview of the differences in network properties that can be expected to arise, given the presence of different sources and degrees of variability. Specifically, we make use of a unique data set that allows us to disentangle the effects of sampling variability (i.e., networks estimated on independent samples) from scale variability (i.e., networks estimated using non-identical measurement tools) when assessing network replicability. One core interest of our study is to explore the extent to which consistencies in network characteristics are improved when precision in node estimation is increased (i.e., when node reliability increases). A second key interest is to explore the idea of network

generalizability when it comes to the same construct measured using different scales. To date, the majority of the network literature has focused on assessing network replicability, whereas less work exists on evaluating network generalizability, that is, the extent to which network characteristics are consistent across – and thus theoretically applicable to – different settings or conditions. One exception is a study by Funkhouser et al. (2020) who found that network properties were more similar across samples within the same (e.g., clinical) population than across (e.g., clinical and nonclinical) populations. The current study seeks to investigate the extent to which networks generalize across different measurement tools, by estimating networks on the same set of nodes with non-identical items (i.e., different scales), while varying whether the same sample or a different sample is used. In this way, we hope to more systematically explore the degree to which the use of different scales (or 'scale variability') contributes to discrepancies between network properties, and in turn, the extent to which these discrepancies are potentially diminished when precision in node estimation (or 'node reliability') is increased.

Challenges with Network Models: Gap in the Research

As network replicability is specifically concerned, a few key challenges have been highlighted within the literature.

First, given the predominant use of single-item indicators to measure nodes, the issue of poor measurement reliability is particularly relevant in the context of estimating cross-sectional networks, and has been raised as a nontrivial limitation (e.g., Forbes et al., 2017a, 2019; Fried & Cramer, 2017; Funkhouser et al., 2020). In response, some authors have advocated for the use of multiple questions to assess each variable (e.g., the Inventory of Depression and Anxiety

Symptoms scale (IDAS); Watson et al., 2007) in order to improve the reliability of node measurements (Fried & Cramer, 2017; Funkhouser et al., 2020).

Second, the use of different instruments or scales to measure the same construct (e.g., Hamilton Rating Scale for Depression (HRSD) vs. Beck Depression Inventory (BDI)) poses conceptual and practical constraints on drawing comparisons across samples: Heterogeneity of scale properties, such as low content overlap (see e.g., Fried, 2017), brings into question the interchangeability of different measures, whereas discrepancies in the number of variables measured necessarily imposes different network structures (Epskamp et al., 2018). Such dissimilarities or variation when it comes to the measurement and estimation of network variables limits the extent to which comparisons between networks are interpretable and observed discrepancies informative.

Finally, sampling variability has been raised as an added challenge inherent to conditional dependence models, such as network models that estimate the pairwise relationships between nodes (i.e., edge weights) via partial correlations (Williams, 2020). Thus, when assessing network replicability, it is important to first assess and/or quantify the degree to which sampling variability should realistically contribute to discrepancies between network properties, especially given that effects of sampling variability have been shown to be more pronounced in networks that make use of ordinal data (with fewer categories) as compared to continuous data (Williams, 2020).

Current Study: Overview

To date, the extent to which different sources of variability are expected to individually or jointly contribute to the replicability and generalizability of network models (or lack thereof)

remains unclear. In light of the three limitations identified above, the current study adopted a novel approach to investigating network replicability, specifically by applying a resampling technique to real-world data. In keeping with suggestions from the network literature (see Forbes et al., 2019; Jones et al., 2021), this method allowed us to systematically vary parameters of interest (i.e., sampling variability, scale variability, and node reliability), whilst also closely mirroring the properties of empirical data. Moreover, what makes our study original is our choice of data set (see details below). Until now, previous studies have primarily explored the effects of sampling variability on network replicability (i.e., how networks that are estimated using the same measurement tool compare across different samples or subsamples). The current data set allows us to extend this research to include additional sources of variability, namely: the effects of estimating networks on identical versus independent samples of varying size (i.e., sampling variability), the use of identical versus different scales (henceforth referred to as 'scale variability'), and the impact of single-item versus multi-item measures (henceforth referred to as 'node reliability').

With respect to network comparisons, we expect to find that the greater the overlap in sample and scale characteristics, the greater the consistency in network properties. Moreover, given that smaller samples should produce larger standard errors around network estimates (e.g., node values, edge weights, centrality measures, etc.), we expect that when two networks are estimated on independent samples, discrepancies between observed network properties should decrease under conditions with smaller sampling error (i.e., larger samples) and with smaller measurement error (i.e., greater node reliability). Similarly, discrepancies between networks, produced as a result of measuring network variables using different measurement tools, are also expected to be attenuated given higher levels of precision in node estimation. The goal of this

study is not to confirm these predictions but to characterize these effects when estimated in real data.

Methods

Database

We made use of the Eugene-Springfield Community Sample (ESCS; Goldberg & Saucier, 2016), an open-source Harvard Dataverse data set, which consists of data collected across the same sample of individuals, over a period of two decades, on a variety of multi-item personality questionnaires. Included amid these questionnaires are the NEO Personality Inventory-Revised (Costa & McCrae, 1992; from here on referred to as NEO) and the International Personality Item Pool (IPIP; Goldberg, 1999; from here on referred to as IPIP), two widespread scales used to measure the five overarching personality dimensions (i.e., Neuroticism, Extraversion, Openness, Conscientiousness, and Agreeableness). We selected the NEO and IPIP scales as they seemed particularly well-suited to investigating the effects of scale variability and node reliability. First, the personality questionnaires can be considered two proxies for the same set of constructs: Given the proprietary nature of the NEO, the IPIP scale was specifically developed as an open-source alternative to the NEO scale, and as such was constructed with a similar style of items which shared similar properties (Goldberg, 1999; more details to follow). Second, each dimension is theoretically comprised of a set of more fine-grained constituent elements (or “facets”), each of which is captured via a multi-item subscale. Use of these tools thus allows us to model personality dimensions as networks, whereby nodes serve to represent constituent facets. We assume that the conceptual distinction between facets should translate to lower degrees of topographical overlap between nodes in the network (see

also Funkhouser et al., 2020; Jones, 2018). Finally, given that both the NEO and IPIP scales are comprised of multi-item scales for each facet, we can systematically vary the level of measurement precision (i.e., node reliability) by choosing a single item to represent each facet or summing up to eight items.

As described, each questionnaire includes five large scales to measure each of five major constructs; each of these scales are then further subdivided into six subscales to measure the six facets of each respective dimension. For instance, the Neuroticism dimension is made up of the six following facets: Anxiety, Anger, Depression, Self-consciousness, Impulsiveness, and Vulnerability (see Appendix for items and reliability coefficients for each subscale). For simplicity's sake, the current study focused exclusively on networks modeling the Neuroticism dimension, measured via the NEO (8 items per facet) and IPIP (10 items per facet) instruments. For each of the Neuroticism scales, reliabilities (omega coefficients) for the facet subscales were comparable across both instruments (ranging from 0.74 to 0.90). Mean correlation between facet subscale scores for our sample was 0.77 (ranging from r_p = 0.74 to 0.81).

Sample Demographics

The ESCS sample consists of participant data collected across a series of questionnaires from 1993 to 2003. In 1993, the sample consisted of 98.4% Caucasian participants (56.9% female), ranging in age from 18 to 85, and holding varying levels of highest education attained (1.2% not graduated high school, 8.9% high school degree, 54.6% vocational or college degree, 35.2% post-college education). For the current study, the largest available sample with complete data on both the NEO and IPIP questionnaires were included (N = 424, after excluding cases with missing observations).

Resampling Approach and Experimental Conditions

In order to create a set of resampling conditions for network comparisons, in which different sources of variability were present, we started by sampling pairs of data frames from the full ESCS sample, in three different ways (referred to henceforth as the three "resampling conditions").

I. In the 'sampling variability' resampling condition, we held the measurement tool fixed but varied the sample, that is, we randomly sampled a pair of data frames, where each data frame contained observations from independent subsamples, but were measured using identical scale items. For instance, for single-item indicators, the Anxiety facet could be measured using the same NEO scale item (e.g., "I am not a worrier") in both data frames.

II. In the 'scale variability' resampling condition, we held the sample fixed but varied the measurement tool, that is, we randomly sampled a pair of data frames, where each data frame contained observations from the same subsample of individuals, but were measured via non-identical items from two different scales. For example, one data frame would measure Anxiety using a NEO scale item (e.g., "I am not a worrier") whereas the other data frame would use an IPIP scale item (e.g., "Am not easily bothered by things").

III. Finally, in the 'sampling and scale variability' resampling condition, we varied both the sample as well as the measurement tool, that is, we randomly sampled a pair of data frames, where each data frame contained observations from independent subsamples, and were measured using the two different scales.

Additionally, across all three resampling conditions, we introduced five levels of node reliability by varying the number of items (i.e., 1, 2, 3, 5, or 8 items) used to estimate each node in the network. While there exist methods that integrate the use of latent variable models and network models (e.g., Epskamp et al., 2017), we simply created sum scores for multi-item indicators prior to estimating the network. Finally, we also manipulated sample size by sampling different proportions of the full sample, namely 20%, 50%, 80%, or 100% of the total number of cases (i.e., n = 84, n = 212, n = 339, or N = 424). In order to avoid overlapping cases between samples, resampling conditions that compared networks across independent samples (i.e., resampling conditions I and III) were necessarily restricted to n's of size 84 or 212.

Taken together – across all three resampling conditions, node reliability levels, and sample size levels – our design included 40 conditions (i.e., a total of 40 combinations of characteristics for pairs of data frames sampled). For each of the 40 conditions, we generated 50 pairs of networks ("replications"); we limited the design to 50 replications per condition in order to reduce computational burden, which was high due to the network estimation approach.

Network Estimation

For each pair of data frames, the two networks were first estimated separately, then compared on a series of metrics (detailed below). We used the 'ggmModSelect' algorithm (Epskamp et al., 2012) to estimate networks. This algorithm begins by estimating a series of 100 networks using the graphical lasso ('glasso'; Friedman et al., 2008) with a range of penalty parameters. These 100 starting networks differ in which edges are estimated to be absent and which are present. Holding these structures of edge presence fixed, each of the 100 networks is re-estimated without the penalty parameter, to obtain non-regularized edge weights. These re-

fitted networks are compared using the Bayesian Information Criterion (BIC), and the optimal (lowest BIC) model is selected. Next, edges are added or removed, one by one, until the BIC can no longer be improved (Epskamp et al., 2012). The final estimated network (i.e., the weighted adjacency matrix) contains a combination of missing edges and non-zero edges, with non-zero edge weights representing partial correlations. Network edges are thus taken to represent the conditional dependencies, i.e. direct causal associations, between any pair of nodes in the network. All analyses were run in R, using the bootnet package (Epskamp et al., 2018) and the "estimateNetwork" function with default = ggmModSelect amd stepwise = TRUE. Across all network estimations, the data were treated as continuous (corMethod = "cor").

Network Comparisons

Consistency between network properties has been assessed within the literature using a variety of metrics. We focused on the set of metrics described below, when comparing between pairs of networks, across all three resampling conditions. Taken together, these metrics should provide a descriptive overview of the variability in network properties that one might expect given the presence of sampling variability and/or scale variability, and at varying degrees of node reliability.

It should be noted that permutation tests, like the *Network Comparison Test* (*NCT*, van Borkulo et al., 2017), can be used to statistically test for invariance in network properties. Due to the nature of our resampling design, however, such invariance tests have limited interpretability. This is because our resampling approach involves randomly resampling from within the same population sample, and therefore necessitates that across replications, samples drawn will share overlapping observations. Moreover, when varying node reliability, the larger the number of

items being sampled (e.g., multiple-item indicators with 5 or 8 items), the greater the level of overlap in items sampled across replications. For these reasons, when presenting results of our resampling conditions, we focus on reporting descriptive statistics, and interpreting observed trends without appealing to significance tests.

Adjacency Matrices

First, we compared unweighted adjacency matrices, for each pair of networks, by tallying the number of corresponding edges that were consistently estimated as absent in both networks, as present in both networks, or different in each network (i.e., absent in one network, but present in the other, or vice-versa). The greater the proportion of consistent edges, the more similar the network skeletons or structures. This metric allows us to compare network estimation outcomes (i.e., which edges are estimated as present vs. absent), but does not consider the relative size of corresponding edge weights.

Global Strength

Next, we assessed differences in global strength, computed as the sum of all absolute edge weights in the network (van Borkulo et al., 2017). Let V represent the set of nodes in networks (graphs) G_1 and G_2. Let A_1 and A_2 represent the weighted adjacency matrices for G_1 and G_2, respectively. Global strength (for G_1) was defined as

$$S_{G1} = \sum_{i,j \epsilon V} |A_{1ij}|,$$

and difference in global strength (between G_1 and G_2) was defined in terms of distance D

$$D(G_1, G_2) = |S_{G1} - S_{G2}|.$$

Global strength has been used as a measure of overall network connectivity (van Borkulo et al., 2017). The smaller the discrepancy in global strength scores, the greater the similarity in connectivity. Because global strength is computed as an absolute sum score, two networks can yield similar connectivity scores despite differences in network structures and edge directions.

Edge Weight Correlations

Overall similarity between edge weights was assessed with the Pearson correlation (r_p) of the set of corresponding edge weights (i.e. 15 raw partial correlation values) in each pair of networks, averaged across replications.

Centrality Correlations

Overall similarity in node centralities was measured by computing the Spearman-rank correlation (r_s) between the raw Expected Influence scores (Robinaugh et al., 2016) for each pair of networks, and averaged across replications. Expected Influence has been described as a means to identify influential nodes. If one assumes that edges represent equal and bidirectional causal effects (Borsboom & Cramer, 2013), then Expected Influence accounts for both the activating and deactivating influence a node is expected to have on its immediate neighbors (i.e., the nodes it shares an edge with), and in turn, the rest of the network (Robinaugh et al., 2016). Similar to Strength, a node's Expected Influence is measured as the summed weight of all edges connected to neighboring nodes; unlike Strength, however, Expected Influence distinguishes between positive and negative edges, rather than taking the absolute values of edges.

Network Generalizability

In order to explore more closely whether we can expect network properties to generalize across different measurement tools (i.e., NEO vs. IPIP scale), we first estimated and compared whole-sample networks (described below). Next, we drew additional comparisons using a tallied frequency approach, within the 'scale variability' resampling condition, at n = 212 (50% the sample size of the full sample) and at three levels of node reliability (1-item, 3-item, 8-item).

Whole-sample Networks

Whole-sample networks for the NEO scale and IPIP scale were estimated on the full available sample (N = 424) and the full set of scale items (i.e. 8-item and 10-item indicators, respectively), using the same network estimation approach outlined above. These networks can be seen as representing the population from which all our resampled replications were drawn. Comparison between whole-sample networks illustrates the extent to which network properties can be expected to differ when the same construct is estimated using non-identical measurement tools, under relatively optimal measurement circumstances; that is, when estimated on the same sample of individuals (i.e. no sampling variability), at a sample size twice as large as the observed subsample replications, and using scales with high node reliability.

In addition to descriptive statistics (global strength, number of present and absent edges, density, largest edge weight, most central node) and correlational analyses (edge weights, node centralities), we ran the *NCT* (1,000 iterations), to test the invariance in global strength and edge weights. The *NCT* also includes an invariance test for comparison of network structures. Given, however, that it operationalizes network structure invariance as the hypothesis that all edge weights are identical across networks, invariance can be violated as a function of one significant

edge difference, or alternatively multiple significant edge differences. The outcome of the network structure invariance test alone, therefore, cannot reveal the full scope of differences that may exist between networks. As such, we report only on specific edge differences when comparing whole-sample networks. In order to assess accuracy of edge estimation, we computed the 95% bootstrapped confidence intervals (CIs) for each network (see Epskamp et al., 2018). Finally, we also computed the *correlation of stability coefficient* (*CS*-coefficient) for each of the networks' Expected Influence estimates, based on 1,000 bootstrapped samples. *CS*-coefficients are meant to serve as an index of centrality stability, and represent the maximum proportion of sampled cases that can be dropped in order to retain a correlation of 0.7 (with 95% probability) between original sample and subsample centrality estimates (Epskamp et al., 2018).

Tallied Frequencies

To illustrate the variability in local network properties, we tallied the frequencies across replications for two edge measures (i.e., the largest edge in each network, and the maximum edge difference observed between pairs of networks) and one node measure (i.e. the most central node in each network). For instance, we assessed the extent to which the same node was ranked as most central by tallying the number of times (across all 50 replications) it was ranked first in the NEO network versus the IPIP network. Use of these tallied frequencies is meant first and foremost to provide an impression of the expected variability one can expect when it comes to more local network properties, given the use of different instruments, and under varying measurement conditions.

Results

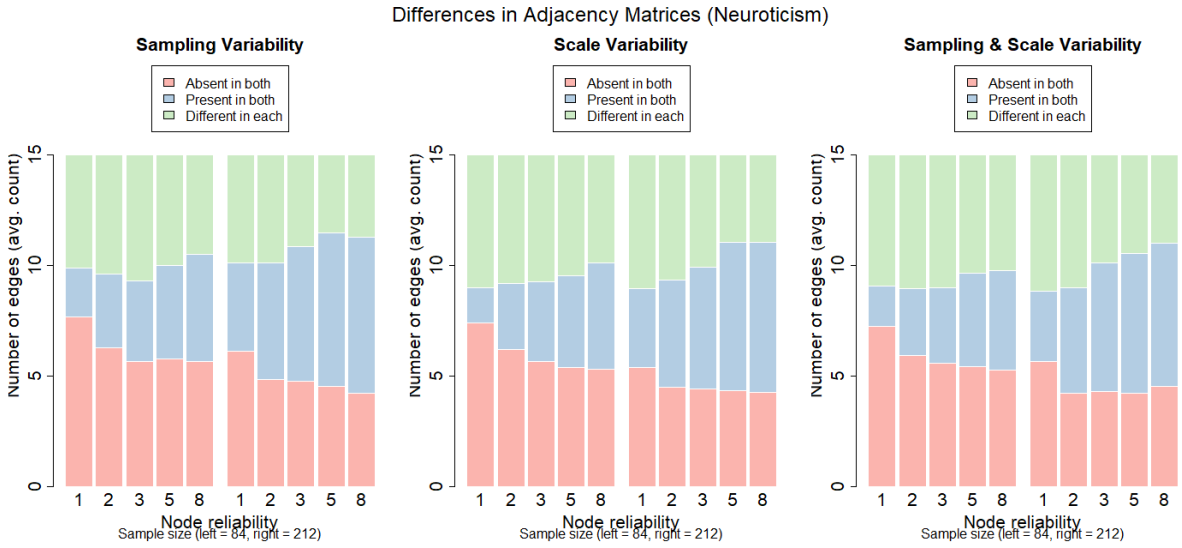
First, we present the results of the three resampling conditions, which explored the individual and joint effects of sampling variability and scale variability, at varying degrees of node reliability. For simplicity, we only display the effects of sampling variability for networks measured on the NEO scale (all statistics were also computed using the IPIP scale, and trends were comparable). This is followed by a more focused look at the effects of scale variability in order to assess the extent to which network characteristics generalize across instruments.

Effects of Resampling Conditions

Results from our three resampling conditions revealed some clear trends with respect to the effects of improved measurement conditions (i.e., larger sample sizes and increased node reliability) on consistency in network properties. In particular, greater precision in node estimation led to denser networks (see Figure 1). Figure 1 demonstrates the results of comparing adjacency matrices across pairs of networks: Plots display the proportion of edges that were consistently estimated as present or absent (in both networks), or different (in each network). Trends in Figure 1 show that when node reliability (or sample size) increased, more edges in the network were estimated, on average, as non-zero. Put differently, when all nodes in the network were more reliably estimated, greater network sensitivity was observed. This finding is fairly intuitive: The more precise an instrument is at estimating the components of the network (i.e., node values), the better it should be at approximating the respective interactions between these nodes (i.e., edge values). Moreover, when using non-identical items to estimate each network, the likelihood that the same set of network edges were estimated as non-zero was greater given increased precision of each of the instruments (i.e. at higher levels of node reliability).

Figure 1

Differences in Adjacency Matrices: Edges Estimated as Present versus Absent

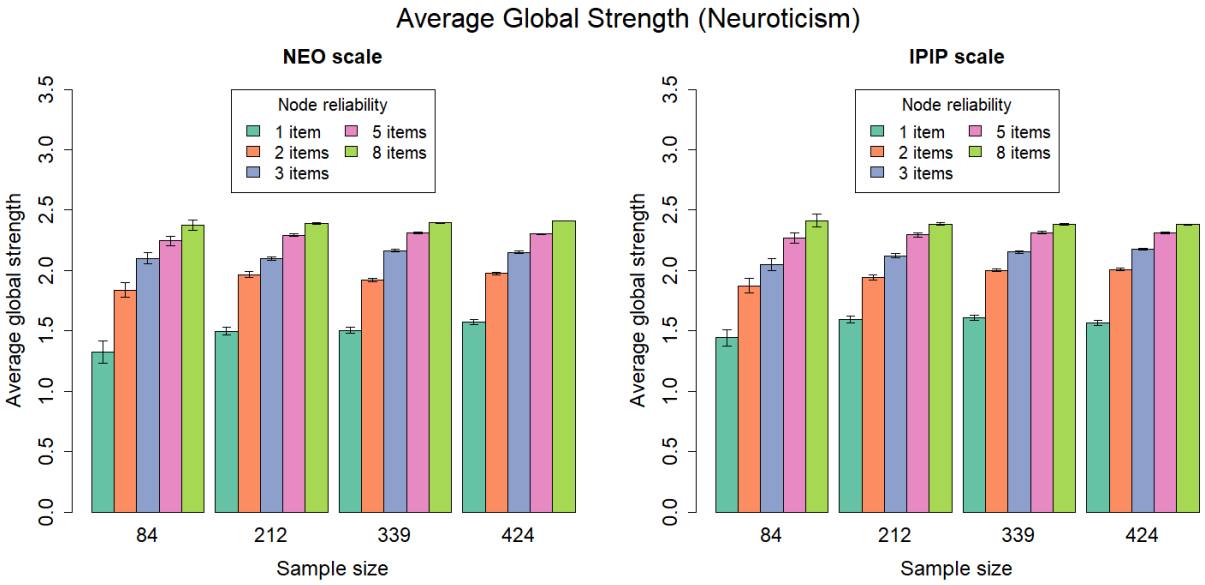


Note. Effects of sample size and node reliability on differences in adjacency matrices, across resampling conditions: sampling variability (left), scale variability (middle), sampling and scale variability (right). Results are displayed in the form of stacked bar charts, such that each bar height represents the total number of possible network edges (i.e., 15), and coloured bar segments break down the relative proportions of total edges that were consistently estimated as absent (red bars) or present (blue bars), or different (green bars).

When assessing network connectivity (i.e., global strength), estimates increased under higher levels of node reliability, but were relatively unaffected by sample size (see Figure 2). Figure 2 plots the global strength score, averaged across replications, for NEO networks (left plot) and IPIP networks (right plot), across all levels of sample size and node reliability. Increased global strength, under higher levels of node reliability, is consistent with the fact that more edges are estimated as present in networks with greater node reliability (see Figure 1).

Figure 2

Average Global Strength

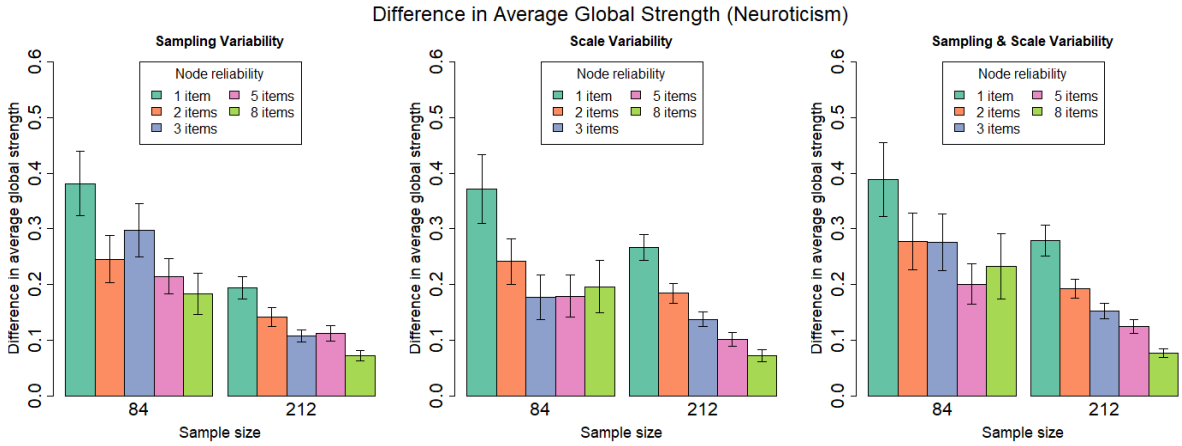


Note. Effects of sample size and node reliability on average global strength of networks. Neuroticism networks were estimated via use of the NEO scale (left) and the IPIP scale (right). Error bars represent 95% confidence intervals around the mean.

Differences in network connectivity, across conditions, are plotted in Figure 3: Higher bars represent greater discrepancies in global strength across pairs of networks. Overall, differences in global strength were more pronounced with smaller sample sizes and with lower node reliability (Figure 3), with marginally greater discrepancies when both sampling and scale variability was present (right plot), as compared to sampling variability (left plot) or scale variability (middle plot) alone. Differences between resampling conditions, however, essentially disappear under more optimal measurement conditions (e.g., n = 212, 8-item indicators).

Figure 3

Differences in Average Global Strength



Note. Effects of sample size and node reliability on differences in average global strength, across resampling conditions: sampling variability (left), scale variability (middle), sampling and scale variability (right). Error bars represent 95% confidence intervals around the mean.

Finally, findings indicated that correspondence between edge weights (Figure 4) and centrality scores (Figure 5) were stronger under more optimal measurement conditions. Figures 4 and 5, respectively, display Pearson and Spearman correlation coefficients, across conditions: Higher bars represent greater strength in correlation between edge and centrality measures. Across both figures, correlation coefficients for 1-item indicator conditions were particularly weak, especially in the 'scale variability' and 'sampling and scale variability' resampling conditions. In general, the overall variability or range of edge correlations (and centrality correlations) was notably more extreme for these two resampling conditions. For example, the range of edge correlations was markedly wider when scale variability (r_p = .09 to .57 at n = 84) or both scale and sampling variability (r_p = .17 to .49 at n = 84) were present (see Figure 4, middle and right plots). In contrast, observed edge correlations ranged from r_p = .30 to .50 at n =

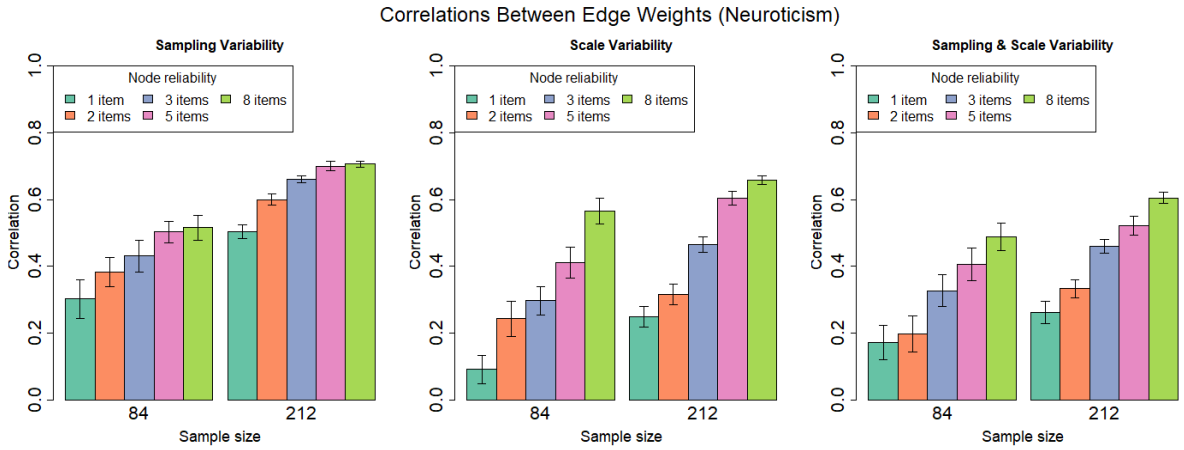
84 in the sampling variability condition (see Figure 4, left plot), where each node was measured with the same item or set of items across samples. A similar pattern was observed for centrality correlations (see Figure 5).

Looking closely at the centrality correlation values (see also supplementary Table S5), another pattern is evident: Across all three resampling conditions, strength of correlations were comparable when networks were estimated with single-item indicators and n = 212 cases (r's = 0.53, 0.47, and 0.31, resp.), or when networks were estimated with 2-item indicators and n = 84 cases (r's = 0.51, 0.46, and 0.39, resp.). In other words, improving node reliability by using one additional indicator yielded the same improvement in correlation scores as increasing the sample size by 2.5 times. This finding underscores the value of improving node estimation in order to increase the likelihood that network properties are observed as consistent across samples.

Under more optimal measurement conditions (i.e., n = 212, and use of 8-item indicators), average correlations between edge weights were more similar across resampling conditions (i.e., r_p = .71, r_p = .66, and r_p = .61, resp.), however still strongest in the absence of scale variability. Similarly, under improved measurement conditions (n = 212, 8-item indicators), average correlations between Expected Influence scores were comparable in size across resampling conditions (i.e., r_s = .85, r_s = .89, and r_s = .86, resp.). Taken together, findings suggest that variability in the measurement tool produced greater disparities in network properties (i.e. edge weight and centrality correlations), as compared to variability in the sample, especially at lower levels of node reliability. Such disparities, however, were attenuated under conditions with reduced measurement error. It follows that improvements in node reliability are thus particularly worthwhile when comparing networks that measure the same construct via different scales.

Figure 4

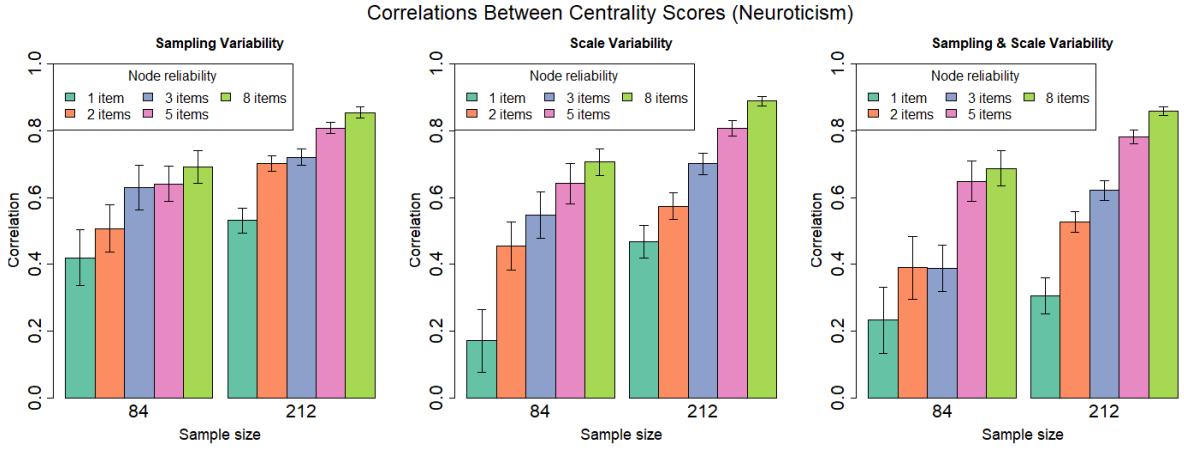
Correlations Between Edge Weights



Note. Effects of sample size and node reliability on correlations between edge weights, across resampling conditions: sampling variability (left), scale variability (middle), sampling and scale variability (right). Error bars represent 95% confidence intervals around the mean.

Figure 5

Correlations Between Centrality Scores



Note. Effects of sample size and node reliability on rank correlations between centrality (i.e. Expected Influence) scores, across resampling conditions: sampling variability (left), scale variability (middle), sampling and scale variability (right). Error bars represent 95% confidence intervals around the mean.

Network Generalizability

One focal interest of the current study was to afford a characterization of the differences likely to arise when two networks of the same construct are estimated via different measurement tools. First, we estimated whole-sample networks for the NEO and IPIP scales. Whole-sample networks can be assumed to provide a better approximation of the 'true' Neuroticism network in the population, but also shed light on scale-specific distinctions in how each instrument captures the facets of Neuroticism. Next, we assessed the variability of three local network properties (largest edge, maximum edge difference, and most central node), by tallying the frequencies of these outcomes across subsample networks (i.e., across 50 replications estimated on *n*'s of 212). This allowed us to evaluate the stability of properties across replications, comparing differences between NEO and IPIP scales, and assess whether this variability was greater under conditions with lower node reliability.

Whole-Sample Networks

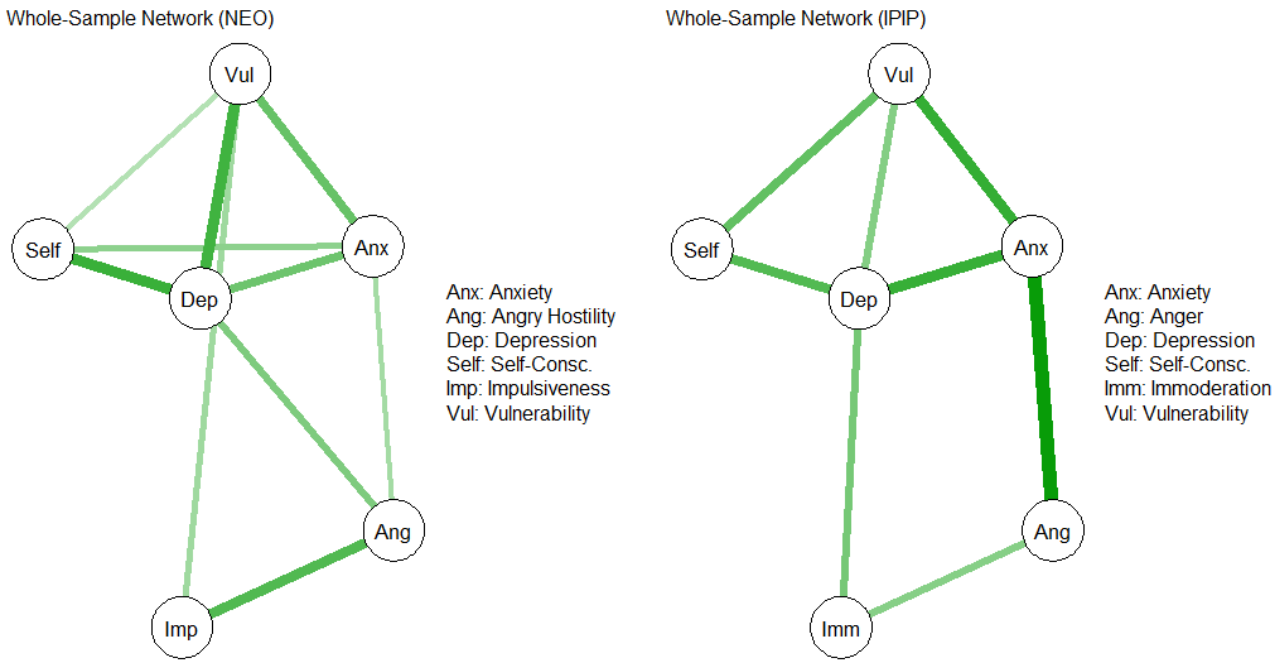
To get the closest possible estimate of the similarities and differences between the population networks of the NEO and IPIP scales, we estimated both networks on the full sample of $N = 424$ participants and their full set of scale items (i.e., 8-item and 10-item indicators, resp.) (see Figure 6). Overall, assessment of global properties illustrates a strong correspondence between NEO and IPIP population networks. Differences, however, become more evident when evaluating discrepancies at the level of specific edge and node properties.

Whole-sample Networks: Descriptive Statistics. Table 1 provides a descriptive summary of the network characteristics for each of the whole-sample networks. Differences in adjacency matrices are evident from both Table 1 and Figure 6: Network plots display a greater

number of edges estimated as present in the NEO network as compared to the IPIP network (i.e., 10 edges and 8 edges, resp.). In turn, network density was greater for the NEO scale (0.67) than for the IPIP scale (0.53). Interestingly, despite differences in network densities, measures of global strength were nearly identical across NEO and IPIP networks (2.41 and 2.40, resp.). This is because, on average, the denser NEO network had an overall smaller mean edge strength than the sparser IPIP network (see Table 1). In this way, differences at the level of adjacency matrices did not produce observed discrepancies in network connectivity, that is, when network connectivity was operationalized as a sum score of all absolute edge weights in the network (van Borkulo et al., 2017).

Figure 6

Whole-Sample Networks



Note. Non-regularized GGM networks for the Neuroticism network, estimated on the full sample (N = 424), using the NEO scale (left) vs. the IPIP scale (right). Solid-green lines represent positive edge weights; dashed-red lines represent negative edge weights. Networks were plotted using the average of the two Fruchterman-Reingold (or "spring") layouts in *qgraph* (max edge weight set to 0.45; Epskamp et al., 2012). Largest edge weights for the NEO and IPIP networks are, resp., between DEP and SELF (edge weight = 0.35), and ANX and ANG (edge weight = 0.44).

Table 1

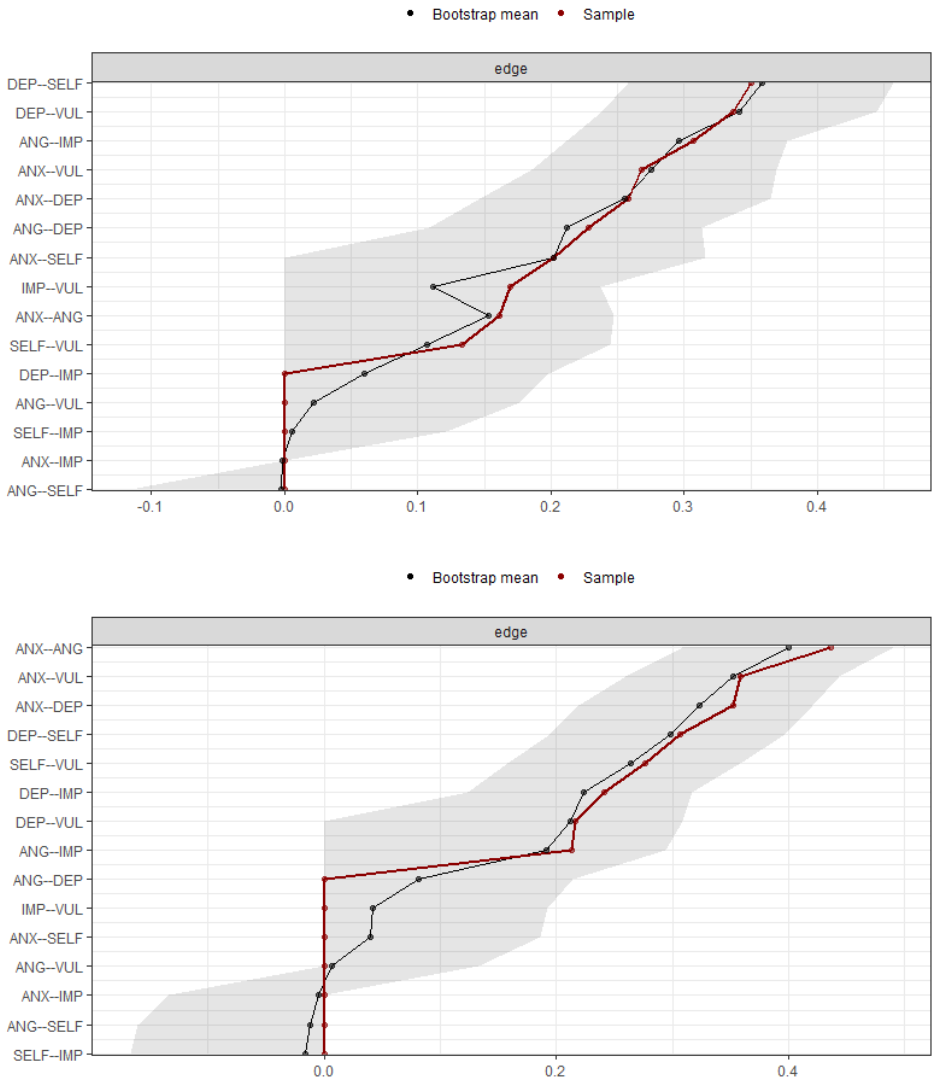
Whole-Sample Networks: Descriptive Overview of Network Characteristics

		NEO	IPIP
Network characteristic		8-items	10-items
Global strength		2.41	2.40
Average edge strength: M (SD)		0.24 (0.08)	0.30 (0.08)
Number of non-zero edges		10	8
Number of zero edges		5	7
Network density		0.67	0.53
Largest edge weight		DEP–SELF (0.35)	ANX–ANG (0.44)
Strength of most different edge (ANX–ANG)		0.16	0.44
CS-coefficient for Expected Influence [a]		0.75	0.75
Centrality ranks (Expected Influence raw score)			
	1st	DEP (1.17)	ANX (1.15)
	2nd	VUL (0.91)	DEP (1.12)
	3rd	ANX (0.89)	VUL (0.85)
	4th	ANG (0.70)	ANG (0.65)
	5th	SELF (0.68)	SELF (0.58)
	6th	IMP (0.48)	IMP (0.45)

Note. All values were computed on the full sample of $N = 424$. ANX = anxiety, ANG = anger, DEP = depression, SELF = self-consciousness, IMP = impulsiveness, and VUL = vulnerability.
a. *CS*-coefficients are an index of centrality stability, and represent the maximum proportion of sampled cases that can be dropped in order to retain a correlation of 0.7 (with 95% probability) between original sample and subsample centrality estimates (Epskamp et al., 2018).

Figure 7

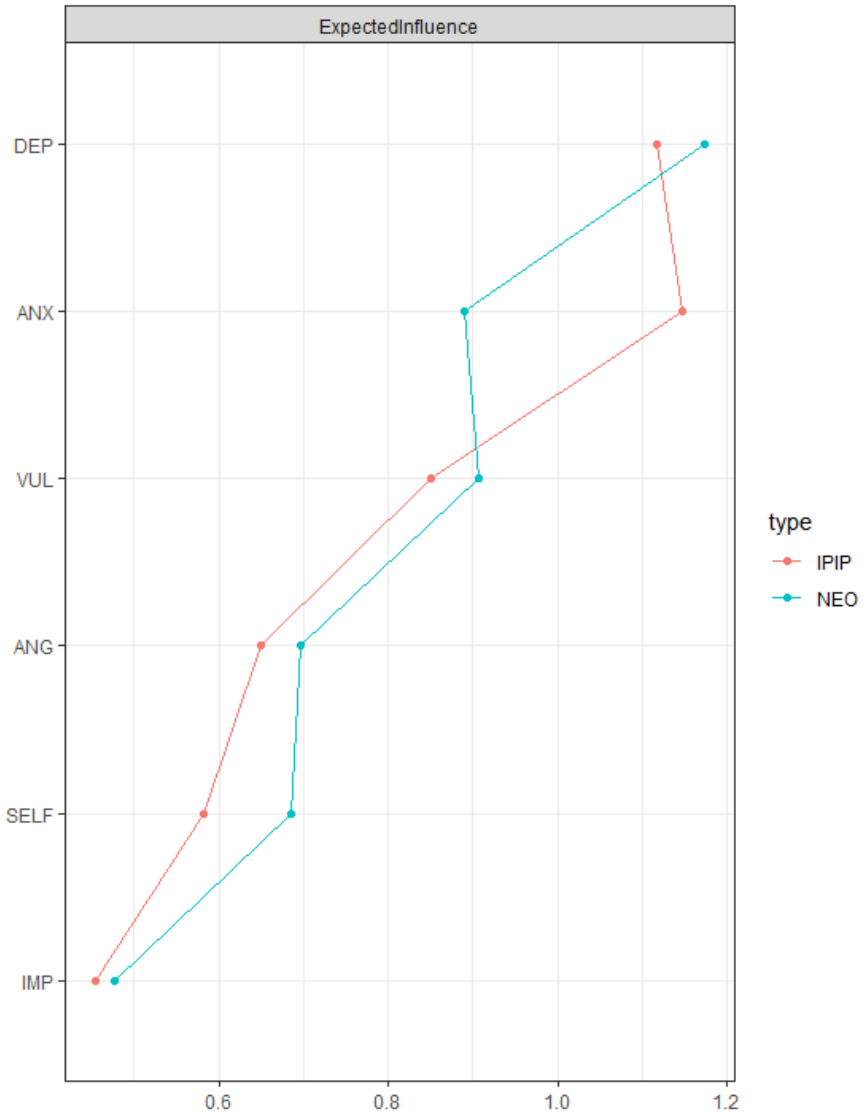
Whole-Sample Networks: Accuracy of Edge Weight Estimates



Note. Edge weight estimates are plotted for NEO (top panel) and IPIP (bottom panel) whole-sample networks. All 15 edges are listed along the y-axis, in descending order of edge weights (indicated by the solid red line). Grey horizontal bars represent the 95% confidence intervals (CIs) for the estimates. Edges whose bootstrapped CIs contain zero should be interpreted with caution (see Epskamp et al., 2018). ANX = anxiety, ANG = anger, DEP = depression, SELF = self-consciousness, IMP = impulsiveness, and VUL = vulnerability.

Figure 8

Whole-Sample Networks: Correspondence in Expected Influence Scores



Note. Expected Influence scores for whole-sample networks are plotted in terms of raw values along the x-axis, with nodes ordered by (descending) rank along the y-axis. Blue line represents estimates derived from the NEO scale, and red line represents estimates derived from the IPIP scale. All values were computed on the same sample of subjects (i.e. full sample of N = 424). ANX = anxiety, ANG = anger, DEP = depression, SELF = self-consciousness, IMP = impulsiveness, and VUL = vulnerability. Most central node, as estimated by the NEO and IPIP scales, respectively, are Depression and Anxiety.

Whole-sample Networks: *NCT* Statistics. To complement descriptive statistics, the *NCT* was run to test for the invariance of global strength, and the invariance of individual edge weights. Unsurprisingly, given the similarity in global strength scores, the *NCT* failed to reject the test for invariance in global strength (difference = 0.01, $p = .731$). With respect to edge differences, edge weights for each of the respective networks are plotted in order of strength (see Figure 7). The strongest edge was not consistent across both networks. In the NEO network, the edge linking the Depression and Self-consciousness facets was estimated as having the strongest weight; this weight was marginally weaker in the IPIP network (edge weights of 0.35 and 0.31, respectively). In the IPIP network, the strongest edge was between Anxiety and Anger (weight of 0.44). In contrast, this edge weight was markedly smaller in the NEO network (weight of 0.16, with bootstrapped 95% CIs including zero), and was the second smallest non-zero edge in the entire network. *NCT* was run to test the invariance across all 15 individual edge weights, with Bonferroni adjustment for testing of multiple edges. In the majority of cases, (i.e., 12 out of the 15 network edges), the *NCT* failed to reject the hypothesis of invariance in edge strengths. Three individual edges, however, were significantly different between networks. The greatest difference of edge strengths was between Anxiety and Anger (difference = 0.28, $p < .001$). Additionally, edges linking Depression and Impulsiveness (difference = 0.24, $p < .001$), and Anger and Depression (difference = 0.23, $p = .03$), were also significantly different. In the case of these latter two edges (i.e., Depression–Impulsivity; Anger–Depression), edges were estimated as present in one network, but absent in the other (see Figure 7). Non-zero edges are sometimes taken as evidence for direct causal associations between nodes, though this practice has been criticized (Dablander & Hinne, 2018, 2019). The findings outlined here would imply

distinctions in how Neuroticism facets are inferred to causally relate, depending on the scale used.

Whole-sample Networks: Correlational Analyses. The correlation between edge weights was strong (r_p = .74), and comparable in size to the mean correlation between corresponding subscale scores (r_p = .77). We discuss potential implications in the Discussion. The correlation between centrality scores was also quite strong (r_s = .83); see Figure 8 for visual display of the correspondence between raw centrality scores. Both NEO and IPIP networks yielded *CS*-coefficients of 0.75, demonstrating high levels of stability (see also Figure S1 in Supplementary Information). Finally, with respect to node properties, rank order of nodes, from most central to least central, was not consistent across the two scales (see Table1): Use of the NEO scale determined Depression as the most central node, whereas use of the IPIP scale ranked Anxiety as most central (see also Figure 8).

Tallied Frequencies

Next, we used a tally approach to better characterize the variability in network properties across replications, when two networks are estimated via different measurement tools. We focused on tallying the frequencies across three measures: i) the largest edge in each network; ii) the maximum edge difference observed between pairs of networks; and iii) the most central node in each network. We computed each of these outcomes across all 50 replications in the 'scale variability' resampling condition, at n = 212, and at three levels of node reliability (1-item, 3-item, and 8-item indicators). In this way, we explored not only the variability of these specific network characteristics across replications, but also how discrepancies potentially diminished with improved node reliability and compared to whole-sample networks.

Study findings support the idea that differences in network characteristics, produced as result of using two distinct scales, are likely to be reduced given improved precision in node measurement. Moreover, tallied frequencies of edge and node properties underscore how differences between samples may be less reflective of nonreplication, but rather consistent with properties of the population network and/or scale-specific differences.

Table 2 summarizes the selective output from the 'scale variability' condition, displaying average estimates for networks measured via the NEO scale versus the IPIP scale. Visual inspection of Table 2 reaffirms some of the trends already discussed above: At each level of node reliability, average global strength was fairly comparable between NEO and IPIP networks. As node reliability improved, average global strength (i.e. sum of absolute edge weights) increased, consistent with the fact that greater node precision led to greater network density. Meanwhile, average edge strength (of non-zero edges) remained fairly stable, reflecting simultaneous increases in the strength of existing edges and the addition of new, weaker edges. Increased precision in node estimation also led to greater overall correspondence in edge weights and centrality scores. Table 2 helps illustrate the merit of using multi-item indicators over single-item indicators. Even a small increase in the number of items used to estimate each node (i.e. from one to three items), led to substantial differences in correlations observed: correlation between edge weights improved from r_p = .25 to .47, and correlation between centrality scores from r_s = .47 to .70.

Assessment of more local network characteristics (e.g., specific edge and node comparisons) helped underscore the existence of scale-specific differences inherent to the NEO and IPIP instruments. First, observed frequencies for the top three most central nodes (Table 2) showed that, overall, higher degrees of node reliability tended to produce greater levels of

consistency across replications. For instance, when networks were estimated via the NEO scale using 1-item indicators, only 29 out of 50 replications ranked Depression as the most central node. In contrast, proportions of replications that yielded the same outcome were increasingly higher in networks based on 3-item and 8-item indicators (i.e., 36 and 49 out of 50 replications, respectively). Comparison between NEO versus IPIP scales reflected differences in patterns of centrality proportions. Specifically, high node reliability (i.e., 8-item indicators) led to almost perfect agreement across replications (49 out of 50) for networks based on the NEO scale. In contrast, centrality scores for IPIP networks remained fairly divided: 31 out of 50 replications determined that the node with greatest Expected Influence in the network was Depression, whereas 17 out of 50 replications determined it to be Anxiety. These differences between scales are (in retrospect) not surprising given the pattern in centrality scores for each of the whole-sample networks: As visible in Figure 8, the NEO whole-sample network displays a clear rank order when it comes to the Expected Influence of nodes, whereas for the IPIP scale two nodes (Anxiety and Depression) are essentially tied for first. This outcome highlights the role that population features play in replicability outcomes: Variability in network properties, across different samples, should be considered in light of the underlying true network model and scale-specific properties.

A similar pattern was observed for the consistency across replications for the largest edge. At the highest level of node reliability (i.e., 8-item indicators), there was fairly high agreement across replications (35 out of 50) that the Anxiety-Anger edge was largest in networks based on the IPIP scale; in a much smaller proportion of replications, the Anxiety-Depression edge and Anxiety-Vulnerability edge (6 out of 50 each) were estimated as largest. This again reflects the edge weights of the IPIP whole-sample network: the size of the largest edge weight

(Anxiety-Anger = 0.44) was visibly larger than the second (Anxiety-Vulnerability = 0.36) and third (Anxiety-Depression = 0.35) largest edges. In comparison, for networks based on the NEO scale, proportions of replications were more evenly split: The links between Depression and Vulnerability, and between Depression and Self-consciousness, were estimated as having the strongest associations in the majority of replications (26 and 22 out of 50 replications, resp.). This aligns with the NEO whole-sample network which had two edges (Depression-Vulnerability = 0.34 and Depression-Self-consciousness = 0.35), nearly identical in size, that were strongest in the network (see Figure 6). Finally, maximum edge difference was more consistent across replications at higher levels of node reliability, however still fairly low (max 16 out of 50 replications) overall.

Taken together, findings imply that when centrality scores or edge weights are similar in size or strength within a network, such as two “equally” influential nodes or two “equally” strong associations between network variables, it should be reasonable to expect greater variability in such metrics across replications. Moreover, whether node or edge properties are observed to be similar in size/strength may importantly be a function of the measurement instrument itself, and how units within the network are conceptualized and operationalized. In other words, variability in network properties across replications may be less reflective of nonreplication, but rather consistent with properties of the underlying true network model or scale-specific properties.

Table 2

Scale Variability Condition: Descriptive Overview of Network Characteristics

Network characteristic Mean (SD)	1-item		3-items		8-items	
	NEO	IPIP	NEO	IPIP	NEO	IPIP
Average global strength	1.50 (0.24)	1.59 (0.20)	2.10 (0.13)	2.12 (0.12)	2.39 (0.05)	2.39 (0.10)
Average edge strength [a]	0.24 (0.03)	0.24 (0.02)	0.26 (0.02)	0.27 (0.03)	0.27 (0.02)	0.28 (0.03)
Mean no. of non-zero edges (out of a possible 15)	6.38 (1.31)	6.80 (1.09)	8.02 (1.02)	8.04 (0.97)	8.90 (0.68)	8.62 (0.92)
Mean no. of zero edges (out of a possible 15)	8.62 (1.31)	8.20 (1.09)	6.98 (1.02)	6.96 (0.97)	6.10 (0.68)	6.38 (0.92)
Avg. network density	0.43	0.45	0.53	0.54	0.59	0.57
Avg. correlation between edge weights	r_p = .25 (.23)		r_p = .47 (.17)		r_p = .66 (.09)	
Avg. correlation between centrality scores	r_s = .47 (.36)		r_s = .70 (.24)		r_s = .89 (.10)	
Centrality ranks: Expected Influence (frequencies, out of 50, listed in parentheses)						
Most central node (i.e. ranked 1st most often)	DEP (29)	DEP (23)	DEP (36)	DEP (32)	DEP (49)	DEP (31)
Second most central node (i.e. ranked 1st second most often)	ANX (10)	ANX (13)	VUL (8)	ANX (13)	VUL (1)	ANX (17)
Third most central node (i.e. ranked 1st third most often)	SELF (6)	VUL (9)	ANX (5)	VUL (5)	n/a	VUL (2)
Largest edge weights (frequencies, out of 50, listed in parentheses)						
Largest edge	ANX–DEP (15)	ANG–DEP DEP–VUL (9)	DEP–VUL (19)	ANX–ANG (12)	DEP–VUL (26)	ANX–ANG (35)

Second largest edge	DEP–SELF (9)	ANX–ANG (8)	ANX–DEP (13)	DEP–VUL (10)	DEP–SELF (22)	ANX–DEP ANX–VUL (6)
Third largest edge	ANG–DEP (6)	ANX–VUL (7)	DEP–SELF (12)	ANX–DEP (8)	ANX–DEP (2)	DEP–SELF (2)
Maximum edge differences (frequencies, out of 50, listed in parentheses)						
Most different edge	ANX–ANG (7)		ANX–ANG (16)		ANX–ANG (15)	
Second most different edge	ANG–DEP DEP–SELF ANX–VUL DEP–VUL (6)		ANG–IMP SELF–VUL (6)		ANX–SELF (10)	
Third most different edge	ANX–DEP (4)		DEP–SELF (5)		DEP–IMP (6)	

Note. All values were computed on a sample of n = 212. ANX = anxiety, ANG = anger, DEP = depression, SELF = self-consciousness, IMP = impulsiveness, and VUL = vulnerability. r_p and r_s denote Pearson and Spearman-rank correlation coefficients, respectively.

a. Average edge strength was computed as the mean of absolute edge weights across all non-zero edges in the network.

Discussion

The current study investigated how different sources of variability (i.e., sampling variability and scale variability) individually and jointly contribute to observed discrepancies in network properties, under poorer versus more optimal measurement conditions (i.e., larger samples and higher levels of node reliability). Our work revealed some clear patterns of findings: As precision in node estimation increased, denser networks were observed, i.e., more edges in the network were estimated as non-zero, indicative of higher network sensitivity. Accordingly,

increased node reliability led to greater estimates of global strength across individual networks, and smaller levels of discrepancy between global strength estimates for pairs of networks. Improvements in node precision also led to greater correlations between edge weights and between centrality scores, which was even noticeable when moving from use of single-item indicators to 2- or 3-item indicators. Findings also indicated that the addition of one indicator can improve consistency in centrality scores, across replications, to the same degree as increasing the sample size by 2.5 times. Findings underscore the value of improving node estimation to increase the consistency of network properties across samples.

Comparisons across resampling conditions highlighted the relative impact of sampling variability versus scale variability. Specifically, correlations between edge weights and between centrality scores were weakest in the 'scale variability' and 'sampling and scale variability' conditions, indicating that variance in the use of items led to greater discrepancies between network replications than comparisons across independent samples. Even under improved measurement conditions (i.e., 8-item indicators and n = 212), edge correlations were still weaker in the 'scale variability' condition (r_p = .66) than the 'sampling variability' condition (r_p = .71), and weakest when both sampling variability and scale variability were present (r_p = .61). Comparing these latter two values, it is clear that introducing the use of two different scales to measure the same construct led to added discrepancies in network properties beyond those produced from sampling variability alone. If we construe nonreplication as departures from invariance beyond those consistent with expected sampling variability (Jones et al., 2021; Williams et al., 2020), then results indicate that how we measure a construct can importantly impact the likelihood that it will replicate in future samples.

Evaluation of more local edge and node properties, within the 'scale variability' resampling condition, shed more light on the notion and implications of scale-specific differences. Observed variability in outcomes across replications (represented as tallied frequencies in Table 2) appeared to be consistent with patterns observed in the whole-sample networks. While, overall, consistency across replications increased with improved node reliability, not all properties reached consensus across replications, even under good measurement conditions. This was especially visible in cases where the two strongest edges in a network were roughly equal (i.e., links between Depression and Vulnerability, and Depression and Self-consciousness, in the NEO whole-sample network), and when two nodes were observed to be equally ranked as most influential (i.e., Anxiety and Depression nodes in the IPIP whole-sample network). In these cases, even when networks were based on 8-item indicators (n = 212), proportions of replications were still relatively split between both largest edges (for the NEO networks) and between both influential nodes (for the IPIP networks). It follows that variability across replications may not necessarily be indicative of unstable properties, but rather may stem from differences in how a construct has been conceptualized and operationalized into a specific set of items. More broadly, if we assume, for any network, that a true network model exists, then this same idea applies: Variability across replications may simply stem from the nature in which nodes relate at the population level.

To elucidate this point, we can consider more closely elements of the NEO and IPIP Neuroticism scales. While we assumed that the NEO and IPIP scales were viable proxies of the same construct (mean correlation between subscale scores r_p = 0.77), with good levels of reliability for their respective subscales (omega coefficients ranging from .74 to .90), it is possible that these two instruments tapped into different yet real features of the Neuroticism

construct and constituent facets. Take, for instance, the Depression subscales which correlated highest amongst the facets (r_p = 0.81), and had the highest reliability coefficients (0.87 and 0.90 for NEO and IPIP, respectively): In the NEO scale, some items may have tapped more into the element of shame or guilt (e.g., “I have sometimes a deep sense of guilt or sinfulness”, “I tend to blame myself when anything goes wrong”), whereas the IPIP items may have attempted to capture the idea of self-worth (e.g., “Have a low opinion of myself”, “Dislike myself”). Given the complex nature of a facet such as Depression, it is not surprising that content varies between scales or even between items on the same scale (Fried, 2017). Thus, when comparing networks across two different scales, improving precision in node measurement may not produce more convergent results in cases where two reliable scales are capturing different aspects of the same construct. Additional research might benefit from comparing network properties of the same construct across different tools, under good measurement conditions (i.e., large samples and high node reliability), in order to get a better sense of which network properties are stable regardless of instrument used, versus those which may be more scale-dependent.

What was interesting to observe was that the mean correlation between corresponding subscale scores (r_p = .77) was comparable in size to the observed correlation between edge weights (r_p = .74). The former coefficient is the (average) zero-order correlation measuring the relationship between individuals’ facet-level mean scores when measured on the NEO scale and on the IPIP scale. It is thus representative of the degree to which each of the instruments are theoretically measuring the same set of constructs (in this case, the six Neuroticism facets or nodes in the network). The latter coefficient, on the other hand, is the (average) zero-order correlation between network edge weights, representative of the extent to which the observed conditional relationships between facets are consistent when modeled via each of the respective

scales. From these observations, we might infer that the degree to which two instruments reliably capture the same construct will be accordingly reflected in the extent to which network properties (in this case, edge weights) are consistent. This interpretation alone would be too simplistic to account for the effects of scale variability on network replicability; that said, it may hint at the extent to which network properties may generalize across measurement tools.

Another point worth addressing is how different replicability metrics allow for different information to be inferred, and may in turn appear to lead to inconsistent results. One example was the observed differences in adjacency matrices for comparisons between whole-sample NEO and IPIP networks (i.e., 10 non-zero edges and 8 non-zero edges, resp.). Despite differences in network densities, measures of global strength were nearly identical across networks (2.41 and 2.40, resp.), due to the fact that the denser network had an overall smaller mean edge strength than the sparser network. In this way, differences regarding which edges were estimated as present or absent did not produce observed discrepancies in network connectivity, that is, when network connectivity was operationalized as a sum score of all absolute edge weights in the network (van Borkulo et al., 2017). Another example is the strong CS-coefficient for both NEO and IPIP whole-sample networks, and the strong correlation between network centrality scores (r_s = .83). Taken together, these metrics would suggest strong stability and correspondence between centrality ranks. What is worth noting, however, is where the discrepancies between node centralities occur: Both networks identify the same set of nodes as most central (i.e. Depression, Vulnerability, Anxiety) and as least central (i.e. Anger, Self-consciousness, Impulsivity), however only the three least central nodes are matched in order. If we observed the converse, that is, only the top three nodes were matched in order, the size of the correlation would not change. Moreover, if the general correspondence of most central and least central

nodes were maintained across networks, but neither the rank order of the top three or last three nodes were perfectly aligned, then the overall Spearman correlation between networks would drop from 0.83 to 0.66. Given that there is an interest in the literature in identifying the nodes with the highest, rather than lowest, centrality (e.g., Borsboom & Cramer, 2013; Elliott et al., 2020), and less emphasis on the overall correspondence across all node centralities, then global metrics (such as rank correlations or stability coefficients) may not be as informative as comparisons between specific nodes.

These findings resonate, in part, with concerns raised by Forbes et al. (2019), in that use of global summary statistics may mask differences in local characteristics of networks. That said, this may ultimately rest on the extent to which the conceptualizations – and operationalizations – of network properties (such as network connectivity) are ambiguous or flexible across researchers. Observed variability of network properties across replications may stem from differences in measurements and interpretations, rather than instability across samples. This will accordingly impact inferences drawn about the replicability of networks across different samples or measurement tools.

Conclusion

A key take-away from our study is that improving measurement at the level of estimating network nodes should improve network replicability. Findings suggest that networks based on single-item indicators, even at large samples, are likely to produce very unstable estimates, and therefore should be expected to have low replicability. Additionally, when two networks of the same construct are being assessed via non-identical measurement tools, greater levels of reliability for each scale should lead to fewer discrepancies between observed networks. Put

simply, networks built on more reliable scales should generalize better. One practical recommendation to improve network replicability would be to use multi-item indicators and sum scores to represent variables in the network. Similarly, in cases where one can assess *a priori* the reliability of one's measurement tool at the level of each node (rather than across all items in the scale), this would be a fruitful consideration to factor into one's research plan prior to data collection, given how poor observed replicability indices were at low levels of node reliability.

A broader concept raised was the role the structure of the network model, at the population level, when considering the expected replicability of network properties. In cases where two (or more) edges, or two (or more) nodes, share similar levels of strength or centrality, then replicability metrics or descriptive comparisons that appeal to rank orders (e.g., largest edge or most central node) will necessarily expect to fluctuate more than networks which have a clearer and more systematic pattern of ranking. Here, both substantive expertise, as well as accounting for how a specific instrument conceptualizes and operationalizes scale items, may provide clarity on where variability can be expected to arise. Finally, when assessing network replicability, care should be taken when considering both the informational value as well as the constraints on inference that each replicability metric carries.

Disclosure statement

No potential competing interests are reported by the authors.

Data availability statement

The data (i.e. data set, syntax, supplemental online materials) that support the findings of this study are openly available on the OSF platform (https://osf.io/m37q2/).

References

Abacioglu, C. S., Isvoranu, A.-M., Verkuyten, M., Thijs, J., & Epskamp, S. (2019). Exploring multicultural classroom dynamics: A network analysis. *Journal of School Psychology*, *74*, 90–105. https://doi.org/10.1016/j.jsp.2019.02.003

Armour, C., Fried, E. I., Deserno, M. K., Tsai, J., & Pietrzak, R. H. (2017). A network analysis of DSM-5 posttraumatic stress disorder symptoms and correlates in U.S. military veterans. *Journal of Anxiety Disorders*, *45*, 49–59. https://doi.org/10.1016/j.janxdis.2016.11.008

Borsboom, D., & Cramer, A. O. J. (2013). Network Analysis: An Integrative Approach to the Structure of Psychopathology. *Annual Review of Clinical Psychology*, *9*(1), 91–121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

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. https://doi.org/10.1037/abn0000306

Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment*, *4*(1), 5–13.

Dablander, F., & Hinne, M. (2018). *Node Centrality Measures are a poor substitute for Causal Inference* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/nue4z

Dablander, F., & Hinne, M. (2019). Node centrality measures are a poor substitute for causal inference. *Scientific Reports*, *9*(1), 6846. https://doi.org/10.1038/s41598-019-43033-9

Dalege, J., Borsboom, D., van Harreveld, F., Waldorp, L. J., & van der Maas, H. L. J. (2017). Network Structure Explains the Impact of Attitudes on Voting Decisions. *Scientific Reports*, *7*(1), 4909. https://doi.org/10.1038/s41598-017-05048-y

Elliott, H., Jones, P. J., & Schmidt, U. (2020). Central Symptoms Predict Posttreatment Outcomes and Clinical Impairment in Anorexia Nervosa: A Network Analysis. *Clinical Psychological Science*, *8*(1), 139–154. https://doi.org/10.1177/2167702619865958

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

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*(4). https://doi.org/10.18637/jss.v048.i04

Forbes, M. K., Wright, A. G. C., Markon, K. E., & Krueger, R. F. (2017a). Evidence that psychopathology symptom networks have limited replicability. *Journal of Abnormal Psychology*, *126*(7), 969–988. https://doi.org/10.1037/abn0000276

Forbes, M. K., Wright, A. G. C., Markon, K. E., & Krueger, R. F. (2017b). Further evidence that psychopathology networks have limited replicability and utility: Response to Borsboom et al. And Steinley et al. *Journal of Abnormal Psychology*, *126*(7), 1011–1016. https://doi.org/10.1037/abn0000313

Forbes, M. K., Wright, A. G. C., Markon, K. E., & Krueger, R. F. (2019). Quantifying the Reliability and Replicability of Psychopathology Network Characteristics. *Multivariate Behavioral Research*, 1–19. https://doi.org/10.1080/00273171.2019.1616526

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191–197. https://doi.org/10.1016/j.jad.2016.10.019

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. https://doi.org/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. https://doi.org/10.1177/2167702617745092

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

Funkhouser, C. J., Correa, K. A., Gorka, S. M., Nelson, B. D., Phan, K. L., & Shankman, S. A. (2020). The replicability and generalizability of internalizing symptom networks across five samples. *Journal of Abnormal Psychology*, *129*(2), 191–203. https://doi.org/10.1037/abn0000496

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*(1), 7–28.

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

Huitsing, G., & Veenstra, R. (2012). Bullying in Classrooms: Participant Roles From a Social Network Perspective: Social Networks and Participant Roles in Bullying. *Aggressive Behavior*, *38*(6), 494–509. https://doi.org/10.1002/ab.21438

Jones, P. J. (2018). *Networktools: Assorted tools for identifying important nodes in networks. R package version 1.1. 0* (1.1.0) [Computer software]. https://CRAN. Rproject. org/package= networktools

Jones, P. J., Williams, D. R., & McNally, R. J. (2021). Sampling Variability Is Not Nonreplication: A Bayesian Reanalysis of Forbes, Wright, Markon, and Krueger. *Multivariate Behavioral Research*, *56*(2), 249–255. https://doi.org/10.1080/00273171.2020.1797460

Robinaugh, D. J., Hoekstra, R. H. A., Toner, E. R., & Borsboom, D. (2020). The network approach to psychopathology: A review of the literature 2008–2018 and an agenda for future research. *Psychological Medicine*, *50*(3), 353–366. https://doi.org/10.1017/S0033291719003404

Robinaugh, D. J., Millner, A. J., & McNally, R. J. (2016). Identifying highly influential nodes in the complicated grief network. *Journal of Abnormal Psychology*, *125*(6), 747–757. https://doi.org/10.1037/abn0000181

Rodriguez, A. L., Stephens, D. P., Brewe, E., Ramarao, I., & Madhivanan, P. (2019). A Network Analysis of Domestic Violence Beliefs Among Young Adults in India. *Journal of Interpersonal Violence*, 088626051988992. https://doi.org/10.1177/0886260519889923

van Borkulo, C. D., Bork, R. V., Boschloo, L., Kossakowski, J., Tio, P., Schoevers, R., Borsboom, D., & Waldorp, L. (2017). *Comparing network structures on three aspects: A permutation test*. https://doi.org/10.13140/RG.2.2.29455.38569

van der Maas, H. L. J., Kan, K.-J., Marsman, M., & Stevenson, C. E. (2017). Network Models for Cognitive Development and Intelligence. *Journal of Intelligence*, *5*(2), 16. https://doi.org/10.3390/jintelligence5020016

Watson, D., O'Hara, M. W., Simms, L. J., Kotov, R., Chmielewski, M., McDade-Montez, E. A., Gamez, W., & Stuart, S. (2007). Development and Validation of the Inventory of Depression and Anxiety Symptoms (IDAS). *Psychological Assessment*, *19*(3), 253–268. https://doi.org/10.1037/1040-3590.19.3.253

Williams, D. R. (2018). *Bayesian Estimation for Gaussian Graphical Models: Structure Learning, Predictability, and Network Comparisons* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/x8dpr

Williams, D. R. (2020). *Learning to Live with Sampling Variability: Expected Replicability in Partial Correlation Networks* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/fb4sa

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

Zwicker, M. V., Nohlen, H. U., Dalege, J., Gruter, G.-J. M., & van Harreveld, F. (2020). Applying an attitude network approach to consumer behaviour towards plastic. *Journal of Environmental Psychology*, *69*, 101433. https://doi.org/10.1016/j.jenvp.2020.101433

Appendix

NEO instrument (8 items per facet)	IPIP instrument (10 items per facet)
Anxiety (N1) (alpha = 0.84, omega = 0.84)	**Anxiety (N1) (alpha = 0.85, omega = 0.85)**
I am not a worrier. [R]	Worry about things.
I often feel tense and jittery.	Am relaxed most of the time. [R]
I have fewer fears than most people. [R]	Am afraid of many things.
I am easily frightened.	Fear for the worst.
I rarely feel fearful or anxious. [R]	Don't worry about things that have already happened. [R]
Frightening thoughts sometimes come into my head.	Am not easily disturbed by events. [R]
I'm seldom apprehensive about the future. [R]	Adapt easily to new situations. [R]
I often worry about things that might go wrong.	Get stressed out easily.
	Get caught up in my problems.
	Am not easily bothered by things. [R]
Angry Hostility (N2) (alpha = 0.81, omega = 0.81)	**Anger (N2) (alpha = 0.89, omega = 0.90)**
I'm an even-tempered person. [R]	Lose my temper.
Even minor annoyances can be frustrating to me.	Get irritated easily.
It takes a lot to get me mad. [R]	Seldom get mad. [R]
I often get angry at the way people treat me.	Rarely complain. [R]
I often get disgusted with people I have to deal with.	Keep my cool. [R]
At times I have felt bitter and resentful.	Get upset easily.
I am known as hot blooded and quick tempered.	Am often in a bad mood.
I am not considered a touchy or temperamental person. [R]	Am not easily annoyed. [R]
	Rarely get irritated. [R]
	Get angry easily.

Depression (N3) **(alpha = 0.86, omega = 0.87)**	**Depression (N3)** **(alpha = 0.90, omega = 0.90)**
I am seldom sad or depressed. [R]	Seldom feel blue. [R]
I have a low opinion of myself.	Have a low opinion of myself.
Sometimes things look pretty bleak and hopeless to me.	Feel desperate.
I rarely feel lonely or blue. [R]	Have frequent mood swings.
Sometimes I feel completely worthless.	Am often down in the dumps.
I have sometimes experienced a deep sense of guilt or sinfulness.	Feel comfortable with myself. [R]
I tend to blame myself when anything goes wrong.	Dislike myself.
Too often, when things go wrong, I get discouraged and feel like giving up.	Feel that my life lacks direction.
	Often feel blue.
	Am very pleased with myself. [R]
Self-consciousness (N4) **(alpha = 0.76, omega = 0.77)**	**Self-consciousness (N4)** **(alpha = 0.81, omega = 0.82)**
It doesn't embarrass me too much if people ridicule and tease me. [R]	Am not embarrassed easily. [R]
I often feel inferior to others.	Am easily intimidated.
In dealing with other people, I always dread making a social blunder.	Am afraid that I will do the wrong thing.
I seldom feel self-conscious when I'm around people. [R]	Am able to stand up for myself. [R]
At times I have been so ashamed I just wanted to hide. [R]	Stumble over my words.
I feel comfortable in the presence of my bosses or other authorities. [R]	Find it difficult to approach others.
If I have said or done the wrong thing to someone, I can hardly bear to face them again.	Am afraid to draw attention to myself.
When people I know do foolish things, I get embarrassed for them.	Only feel comfortable with friends.
	Am comfortable in unfamiliar situations. [R]
	Am not bothered by difficult social situations. [R]

Impulsiveness (N5) **(alpha = 0.74, omega = 0.74)**	**Immoderation (N5)** **(alpha = 0.79, omega = 0.79)**
I have trouble resisting my cravings.	Am able to control my cravings. [R]
I have little difficulty resisting temptation. [R]	Easily resist temptations. [R]
I rarely overindulge in anything. [R]	Rarely overindulge. [R]
When I am having my favorite foods, I tend to eat too much.	Often eat too much.
Sometimes I do things on impulse that I later regret.	Do things that I later regret.
I sometimes eat myself sick.	Never splurge. [R]
I am always able to keep my feelings under control. [R]	Don't know why I do some of the things I do.
I seldom give in to my impulses. [R]	Love to eat.
	Go on binges.
	Never spend more than I can afford. [R]

Vulnerability (N6) **(alpha = 0.78, omega = 0.78)**	**Vulnerability (N6)** **(alpha = 0.83, omega = 0.83)**
I'm pretty stable emotionally. [R]	Get overwhelmed by emotions.
I feel I am capable of coping with most of my problems. [R]	Know how to cope. [R]
I keep a cool head in emergencies. [R]	Am calm even in tense situations. [R]
I often feel helpless and want someone else to solve problems.	Can handle complex problems. [R]
It's often hard for me to make up my mind.	Can't make up my mind.
I can handle myself pretty well in a crisis. [R]	Panic easily.
When everything seems to be going wrong, I can still make good decisions. [R]	Readily overcome setbacks. [R]
When I'm under a great deal of stress, sometimes I feel like I'm going to pieces	Feel that I'm unable to deal with things.
	Become overwhelmed by events.
	Remain calm under pressure. [R]

Note. Reverse-items are denoted by [R].