Journal of Abnormal Psychology
2017, Vol. 126, No. 7, 989–999

0021-843X/17/$12.00 http://dx.doi.org/10.1037/abn0000306

COMMENTARY

False Alarm? A Comprehensive Reanalysis of "Evidence That Psychopathology Symptom Networks Have Limited Replicability" by Forbes, Wright, Markon, and Krueger (2017)

Denny Borsboom, Eiko I. Fried, Sacha Epskamp, Lourens J. Waldorp, Claudia D. van Borkulo, and Han L. J. van der Maas
University of Amsterdam

Angélique O. J. Cramer
Tilburg University

Forbes, Wright, Markon, and Krueger (2017) stated that "psychopathology networks have limited replicability" (p. 1011) and that "popular network analysis methods produce unreliable results" (p. 1011). These conclusions are based on an assessment of the replicability of four different network models for symptoms of major depression and generalized anxiety across two samples; in addition, Forbes et al. analyzed the stability of the network models within the samples using split-halves. Our reanalysis of the same data with the same methods led to results directly opposed to theirs: All network models replicated very well across the two data sets and across the split-halves. We trace the differences between Forbes et al.'s results and our own to the fact that they did not appear to accurately implement all network models and used debatable metrics to assess replicability. In particular, they deviated from existing estimation routines for relative importance networks, did not acknowledge the fact that the skip structure used in the interviews strongly distorted correlations between symptoms, and incorrectly assumed that network structures and metrics should be the same not only across the different samples but also across the different network models used. In addition to a comprehensive reanalysis of the data, we end with a discussion of best practices concerning future research into the replicability of psychometric networks.

General Scientific Summary
This commentary presents a reanalysis of the data presented in the target article by Forbes, Wright, Markon, and Krueger (2017) that shows that, contrary to their conclusions, network models replicate well.

Keywords: network analysis, replication, psychopathology networks, methodology, psychometrics

Supplemental materials: http://dx.doi.org/10.1037/abn0000306.supp

Network modeling is quickly gaining ground as a promising way of understanding psychopathological phenomena. As both the theoretical framework and the statistical modeling routines have seen rapid development over the past few years, recent articles have begun to take stock of what has been achieved and to evaluate which new directions psychopathological network research should take (Fried & Cramer, 2017; Fried et al., 2017). The reproducibility of network research ranks firmly among the top priorities: As Epskamp, Borsboom, and Fried (2017) stated, "the current replication crisis in psychology stresses the crucial importance of obtaining robust results, and we want the emerging field of psychopathological networks to start off on the right foot" (p. 989).

Denny Borsboom, Eiko I. Fried, Sacha Epskamp, Lourens J. Waldorp, Claudia D. van Borkulo, and Han L. J. van der Maas, Department of Psychology, University of Amsterdam; Angélique O. J. Cramer, Department of Methodology and Statistics, Tilburg University.

Denny Borsboom, Eiko Fried, Claudia van Borkulo, and Lourens Waldorp are supported by European Research Council Consolidator Grant 647209. Angélique Cramer is supported by Veni Grant 451-14-002 awarded by the Netherlands Organisation for Scientific Research. We thank Richard McNally, Jeroen Vermunt, Claudi Bockting, and Helma van den Berg for their comments on a draft of this article and Ria Hoekstra for her help in gathering and processing data used in this article.

Correspondence concerning this article should be addressed to Denny Borsboom, Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129-B, 1018 WT Amsterdam, the Netherlands. E-mail: dennyborsboom@gmail.com

Similarly, replicability was recently highlighted as one of the five core challenges that the psychopathological network discipline is facing (Fried & Cramer, 2017).

Thus, the importance of assessing stability and replicability of network structures stands beyond doubt. Upon reading Forbes, Wright, Markon, and Krueger's (2017) conclusions, therefore, our immediate reaction was one of concern about some of the network analysis methodologies currently in use, a response we expect many readers to share, especially because Forbes et al. did not tread lightly in their assessment of psychopathology networks. Even though their analysis was limited to just two data sets, they did not hesitate to draw general conclusions and state that "popular network analysis methods produce unreliable results" (General Scientific Summary, p. 1011), have "poor replicability" (p. 1011) and "limited utility" (p. 1011), so that "novel results originating from psychopathology networks should be held to higher standards of evidence before they are ready for dissemination or implementation in the field" (p. 1011).

However, after we had acquired access to the data sets Forbes et al. (2017) analyzed and had used the appropriate network analyses, we found that many of the numerical results from our statistical analyses turned out vastly different from those of Forbes et al. and supported the exact opposite of their conclusion: Psychopathology networks replicate very well. We were able to trace the diverging results to a number of inaccuracies in their analyses. First, contrary to their claims, Forbes et al. did not accurately implement state-of-the-art network analyses, as we show later. Second, their methodology for assessing replication uses debatable measures of replicability. Third, the correlation matrices used by Forbes et al. are distorted due to the presence of a skip structure in the interview.

In the present commentary, we illustrate how these issues led Forbes et al. (2017) to underestimate the quality of network methodology. In addition, we discuss best practices to most effectively conduct research into the reproducibility of psychopathology networks.

Evidence That Psychopathology Networks Replicate Well

When we set out to reproduce Forbes et al.'s (2017) results using the same analyses on the same National Comorbidity Survey Replication (NCS-R) and National Survey of Mental Health and Wellbeing (NSMHWB) data and split-halves,[1] we found that networks replicated well. Table 1 shows a summary of these results for Ising models, relative importance networks, and directed acyclic graphs (DAGs). We do not report results for association networks, first because Forbes et al. did not challenge the replicability of association networks and second because we encountered major issues with the correlation matrices that we discuss in the next section. In addition to the replicability metrics used by Forbes et al., we report metrics to facilitate assessment of the degree to which networks replicate.[2] The most intuitive and important of these metrics, in our view, is the correlation between the network connections in the NCS-R and NSMHWB data sets. This correlation measures the correspondence between the strength of network connections found in both data sets. If the correlation equals one, network connections in the networks are perfectly linearly related across samples, meaning that the networks have essentially the same structure; if it equals zero, the networks have no detectable linear correspondence; if it equals minus one, the networks are exact opposites.

Table 1 shows that the correlations between network connection strengths are all well above .9, indicating that the networks found in the data sets under consideration are highly similar. Figure 1 shows this high correspondence between the network structures by representing them using the same layout; this is advisable because even when plotting two exactly identical networks with different layouts, it is impossible to tell visually how similar networks are. Our split-half analyses, using the same splits as used by Forbes et al. (2017), show comparable results: All parametric network models show correlations between network connection parameters of well over .9.[3] We shortly discuss these results, after which we turn to the question why Forbes et al. reached conclusions opposite from ours.

The Ising Model

The Ising model (van Borkulo et al., 2014) is arguably the most important of the models fitted by Forbes et al. (2017), because it represents state-of-the-art regularized network model estimation for pairwise Markov random fields (PMRFs; Epskamp, 2017) in dichotomous data. Tallying all networks that are reported in the literature at the moment of writing this comment, 62% used a variant of the PMRF, and this percentage is growing quickly because the PMRF has become the default network modeling technique. It is complemented by robustness analyses in *bootnet* (Epskamp et al., 2017) as well as statistical tests for network invariance (van Borkulo et al., 2016), which are powerful tools in assessing network estimation quality and testing the equivalence of network models in different populations, as we illustrate in this comment.

As Forbes et al. (2017) themselves noted, and as Figure 1 (left panels) shows, estimated Ising networks are nearly identical: Node threshold parameters correlate .93 across the data sets, whereas network connection parameters (edge weights) show a correlation of .95 (Spearman correlations equal .85 and .88, respectively). Even though the absolute position of nodes in centrality orders is not invariant, as also reported by Forbes et al., their relative positions are strongly aligned: The centrality metrics of strength, betweenness, and closeness correlate .94, .94, and .76, respectively, across the two data sets. The only sign of nonreplication concerns the presence of three weak negative edges in the NCS-R data that were absent in the NSMHWB data; however, this difference across samples was not statistically significant (as we will show later in this paper). Split-half analyses, as reported in Appendix B of the online supplemental materials, show similar results and indicate high stability of the Ising model.

[1] We thank Forbes, Wright, Markon, and Krueger (2017) for providing us with the exact splits of the data used in the split-half analyses.

[2] All analyses we report were performed using R Version 3.3.1 and the relevant packages on platform ×86_64-w64-mingw32. All code is available at https://osf.io/akywf, with the exception of the NSMHWB data set, which is not publicly accessible; an instructive summary of our analyses with a subset of sample code can be consulted in Appendix A of the online supplemental materials.

[3] Results of the split-half analyses are included in Appendix B of the online supplemental materials.

Table 1
Replication Results of Comparing the Networks for the NCS-R and NSMHWB Data

Variable	Ising models		Relative importance networks (censored)		Relative importance networks (uncensored)		DAGs	
	NCS-R	NSMHWB	NCS-R	NSMHWB	NCS-R	NSMHWB	NCS-R	NSMHWB
Network characteristics[a]								
No. of edges (% of possible)	80 (52.3)	79 (51.6)	118 (38.6)	124 (40.5)	306 (100)	306 (100)	34 (22.2)	33 (21.6)
Density (as in Forbes et al.)	1.08	1.17	.13	.12	.06	.06		
Quality of replication								
Correlation between all edges	.95		.98		.99			
Correlation for nonzero edges	.97		.98		.99			
Jaccard index[b]	.77		.92		1.00		.68	
Change in edge weights (%)[a]	30.4		8.3		22.2			
Replicated edges (%)[a]	69 (86.3)		116 (98.3)		306 (100)		27 (79.4)	
Nonreplicated edges (%)[a]	11 (13.8)		2 (1.7)		0 (0)		7 (20.6)	
Edges unique to replication set (%)[a]	10 (12.7)		8 (6.5)		0 (0)		6 (18.2)	
Node centrality correlations								
Strength/outstrength/outdegree	.94		.94		.98		.87	
Instrength/indegree			.76				.62	
Closeness	.76				.98		1.00	
Betweenness	.94		.84		.92		.79	
Most central nodes[c]								
Strength/outstrength/outdegree	even	even	depr	depr	depr	depr	depr	depr, inte
Instrength/indegree			inte	weig	tie (15 nodes)	mFat	tie (4 nodes)	irri
Closeness	depr	mFat			mFat	mSle	anxi	anxi
Betweenness	depr	even	ctrl	even	gFat	gFat	edge	depr
	Correlation (τ_b)	Matches (%)	Correlation (τ_b)	Matches (%)	Correlation (τ_b)	Matches (%)	Correlation (τ_b)	Matches (%)
Rank-order correspondence[a]								
Strength/outstrength/outdegree	.69	3 (16.7)	.82	9 (50)	.8	4 (22.2)	.75	14 (77.8)
Instrength/indegree			.39	2 (11.1)			.57	16 (88.9)
Closeness	.71	3 (16.7)			.87	6 (33.3)	1.00	18 (100)
Betweenness	.77	11 (61.1)	.84	14 (77.8)	.57	9 (50)	.66	10 (55.6)

Note. In addition to the metrics discussed by Forbes, Wright, Markon, and Krueger (2017; see their Table 2 for detailed explanations), this table reports Pearson correlations between network parameters in the two samples (all > .9), replication statistics for censored and uncensored relative importance networks as implemented in accordance with Robinaugh, LeBlanc, Vuletich, and McNally (2014), and most central nodes for different centrality measures (see Table 1 of Forbes et al. for node abbreviations). NCS-R = National Comorbidity Survey Replication; NSMHWB = National Survey of Mental Health and Wellbeing; even = anxiety about > 1 event; depr = depressed mood; inte = loss of interest; weig = weight problems; mFat = fatigue; irri = irritability; mSle = sleep problems; anxi = chronic anxiety/worry; ctrl = no control over anxiety; gFat = fatigue; edge = feeling on edge.
[a] Computed following the methodology of Forbes et al. [b] The proportion of shared edges relative to the total number of edges in both networks (shared and nonshared). [c] Computed following the methodology of Forbes et al. but for single centrality measures.

Moving beyond descriptive measures, and in contrast to Forbes et al. (2017), we used the Network Comparison Test (NCT) to statistically evaluate the similarity of the Ising models estimated on the NCS-R and NSMHWB data using permutation testing (van Borkulo et al., 2016). The NCT results also indicate that the network structures of the NCS-R and NSMHW replicate very well. First, a test for invariance of network structures, which tests the null hypothesis that all edges are precisely identical across the samples, was not significant ($M = 2.66$, $p = .121$). Second, testing for the invariance of individual edges revealed that none of the edges differed significantly across the two data sets. Thus, despite the high power to detect differences given the two large samples ($N \sim 9{,}000$ per sample), we could not reject the null hypothesis that the NCS-R and NSMHWB networks are precisely identical at the level of the populations from which these samples were drawn.

Relative Importance Networks

As shown in Figure 1 (middle panels), relative importance networks, which were estimated by exactly following the original methodology described in Robinaugh, LeBlanc, Vuletich, and McNally (2014), replicated even better than did the Ising models. Uncensored relative importance networks featured a correlation of .99 between the estimated edge weights in the two data sets, as well as between split-halves of the same data sets (see Appendix B of the online supplemental materials). These findings deviate significantly from those of Forbes et al. (2017); we explain this divergence in the next section.

DAG Analysis

Replication results for DAGs were good, although not as excellent as were the results for the Ising models and relative importance networks. This is not surprising, because DAGs require stronger assumptions,[4] which are less likely to be met in these data. As Table 1 and Figure 1 show, 27 out of 34 DAG edges replicated from the NCS-R to the NSMHWB data set (79.4%), which indicates that the results do converge. In addition, in- and outdegree of nodes featured correlations of .62 and .87, respectively. Visual inspection of Figure 1 (right panels) shows that the same bridge symptoms, which connect major depressive episode (MDE) to generalized anxiety disorder (GAD), are identified in the

[4] For example, DAG analysis assumes that the causal graph contains no cycles and that there are no independence relations in the data that are not a function of the causal relations coded in the DAG (faithfulness; see, e.g., Pearl, 2009, for an extensive treatment).

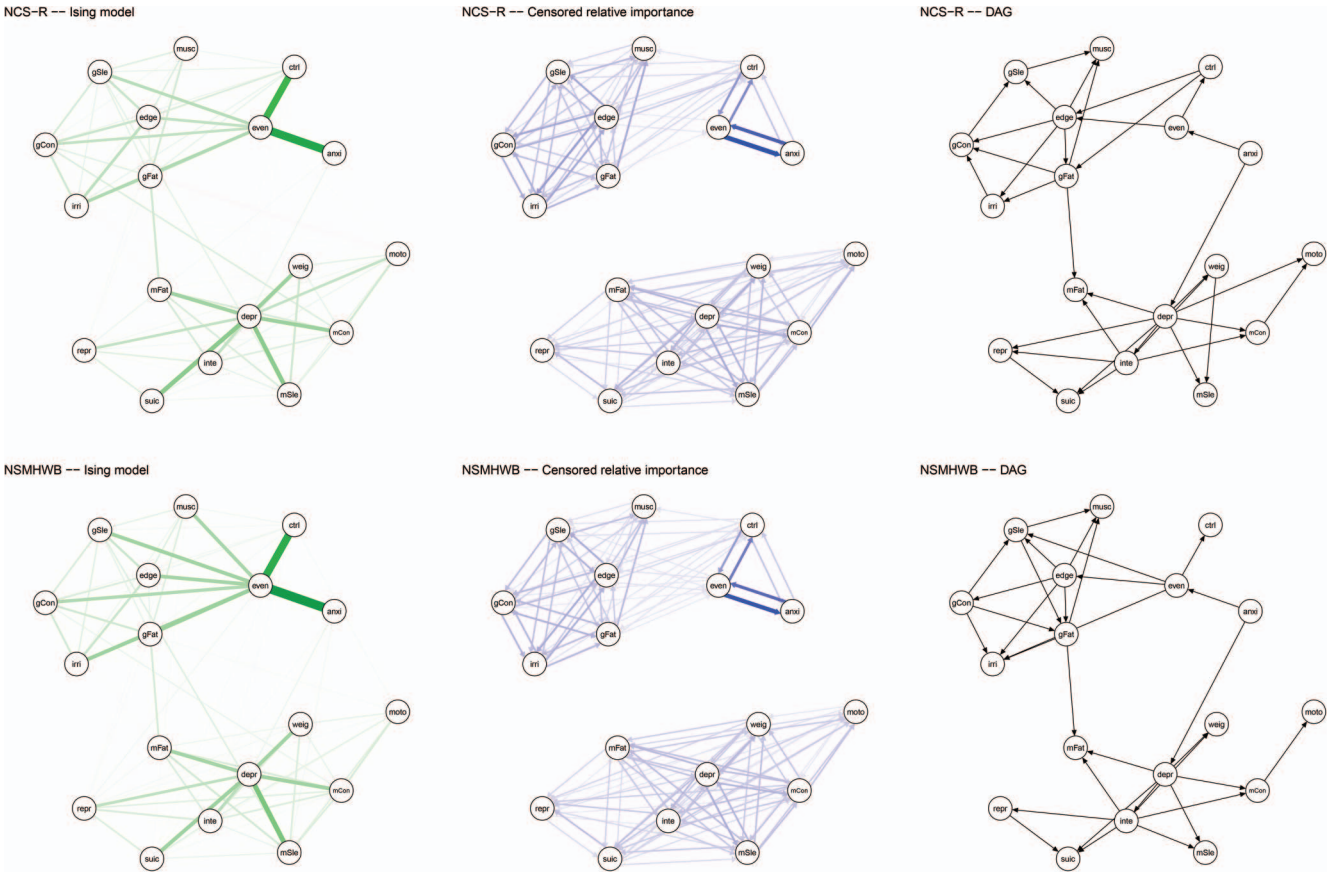

Figure 1. Network structures estimated with the Ising model (left panels), censored relative importance networks (middle panels), and directed acyclic graphs (DAGs; right panels) for the National Comorbidity Survey Replication (NCS-R; top panels) and National Survey of Mental Health and Wellbeing (NSMHWB; bottom panels) data.

two data sets. Of note, two edges (gFat–gCon and gCon–irri) switch direction between the data sets.

Cross-Method Replicability

Forbes et al. (2017) counted how often edges showed up in different network estimation routines. It is clear from the way they interpreted the resulting findings that they assumed that one should expect these different networks to converge to 100%. This, however, is not true. For instance, suppose the data arose from the DAG A → B ← C → D. Then one would not expect to find the Ising model to return the network A–B–C–D, because B is a common effect of A and C, and therefore A and C must be conditionally dependent given B[5] (Pearl, 2009). Instead, one expects the network to also include a direct relation between A and C. In addition, given this network structure, one would never expect any correlations to be nonzero in the association network: Because all variables are connected, one instead expects a fully connected association network. Thus, counting how often individual edges replicate across these different network structures is of limited utility, because it is implausible to expect them to be the same.

In addition, network estimation techniques differ in sensitivity and specificity (van Borkulo et al., 2014), meaning that some techniques more often err on the side of caution and, as such, identify fewer edges, which should be accommodated in assessing replicability. For instance, in relative importance networks all connections are estimated, whereas Ising models estimate only connections that improve the fit of the model (van Borkulo et al., 2014). Similarly, given the stronger causal interpretation of edges in a DAG opposed to Ising models, it is sensible that DAG estimation methods should be more conservative than are Ising model estimation methods, leading DAGs to be sparser. Thus, in addition to principled differences between the edges the methods should detect in the first place, there are also differences in sensitivity and specificity that should be accounted for.

Therefore, rather than counting how many edges are present in different networks, one should investigate a *nesting relationship:* A sparser network should not estimate edges that are absent from

[5] This is because if A and C are independent causes of B, then knowing that B is present means that if A is not present and thus did not produce B, then C must have been the cause of B.

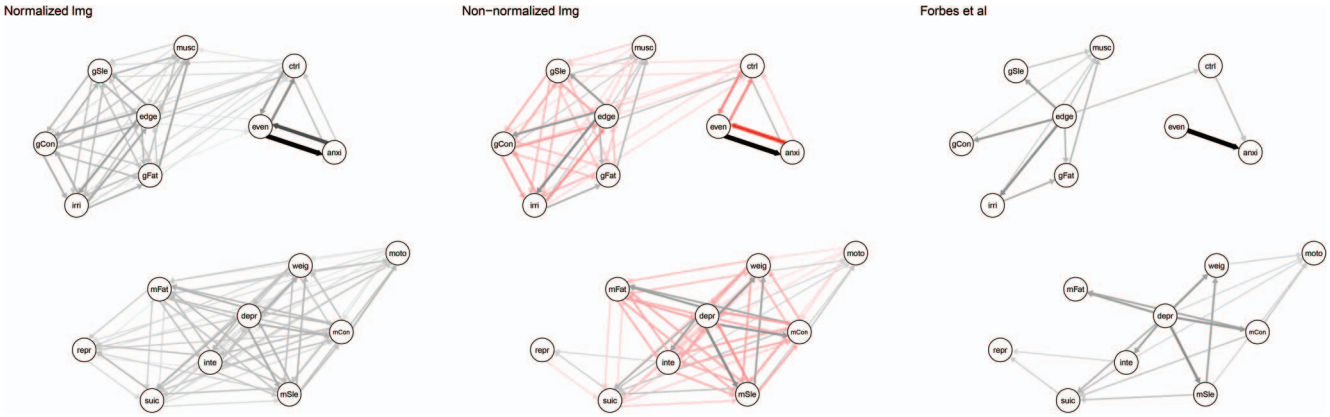

Figure 2. Relative importance networks (estimated on the National Comorbidity Survey Replication data) using normalized *lmg* (left panel, as used by Robinaugh, LeBlanc, Vuletich, & McNally, 2014; 118 edges) and nonnormalized *lmg* (middle and right panels). Red edges in the middle panel (99 edges) indicate edges that are removed by the thresholding rule used by Forbes, Wright, Markon, and Krueger (2017), and the right panel shows the network they reported (31 edges).

the denser network, and a denser network should not leave out edges that are present in the sparser network. When assessing this nesting relation, we found that 100% of the edges in the NCS-R–DAG (the sparser network) were present in the NCS-R–Ising model (the denser network). The same holds for the NSMHWB data. Strikingly, when we compared DAGs and Ising models across data sets, 97% of the NCS-R–DAG edges were included in the NSMHWB–Ising model, and 100% of the NSMHWB–DAG edges were included in the NCS-R–Ising model. In addition, we found that 100% of the edges that were missing in the Ising models were also missing in the DAGs. This is the case both in the split-half analysis and in the replication analysis. Cross-method replication could hardly be better.[6]

Why Did Forbes et al. (2017) Underestimate the Replicability of Psychopathology Networks?

It is remarkable that our results differed so much from those of Forbes et al. (2017), especially given the strong conclusions they drew. After studying their methodology in detail, we argue that the different conclusions originated from two sources. First, Forbes et al.'s analyses contained several statistical inaccuracies. By "statistical inaccuracies" we mean to identify statistical computations that we expect Forbes et al. to acknowledge, upon reflection, as yielding a suboptimal representation of the relations in the data.[7] Unfortunately, these inaccuracies have had strong impact on the results. Second, their results rest on debatable methodologies. By "debatable methodologies" we mean to identify issues that we see as problematic but that can be legitimately disputed depending on one's point of view on what psychopathology networks should deliver or even on one's underlying philosophy of science. We discuss these issues in turn.

Statistical Inaccuracies

When studying Forbes et al.'s (2017) methodology, we found that they did not adopt the standard estimation procedure for relative importance networks introduced by Robinaugh et al. (2014), nor any other published procedure.[8] It is unclear why they deviated from the standard procedure that is used in Robinaugh et al. (2014) and, to the best of our knowledge, in all other articles that have used relative importance networks (Bryant et al., 2017; Heeren & McNally, 2016; Hoorelbeke, Marchetti, De Schryver, & Koster, 2016; McNally, 2016; McNally et al., 2015).

First, Forbes et al. (2017) used nonnormalized instead of normalized estimates for the *lmg* metric to assess relative importance. Although the optimal choice here is debatable, this poses a deviation from standing methodology that should have been acknowledged. Second, Forbes et al. strictly thresholded networks by permanently excluding edges under .05 from the network, whereas Robinaugh et al. (2014) removed these edges for visualization but not in the computation of centrality measures. Third, and most important, Forbes et al. deviated from existing work by introducing a thresholding procedure that has extreme consequences: Whenever an edge between two nodes (e.g., A → B) did not have a weight of at least .005 points higher than the corresponding reciprocal edge for the same two nodes (i.e., A ← B), Forbes et al. removed that edge from the network. This thresholding rule has not been used anywhere else in the literature, and for good reason. For suppose A explains 50% of the variance in B, and B explains 50% of the variance in A: Even though these could be the strongest

[6] We did not investigate the cross-method replicability including relative importance networks, because these do not feature careful edge selection methods.

[7] One of these inaccuracies was already acknowledged: The reader may note that the DAGs in Forbes et al. (2017) are different from the version of their article that was accepted for publication in the *Journal of Abnormal Psychology,* which is the version that the authors had made available online and that we were asked to comment on (we were subsequently sent the revised version). The difference is due to an error in the implementation of DAGs that Forbes et al. caught in time to correct the article between acceptance and publication.

[8] If we deviate in the same way from the literature, we can reproduce their reported results, and hence we are certain this deviation is the source of the differences (see Figure 2 for details and Borsboom et al., 2017, for code replicating both our and Forbes et al.'s analyses).

edges in the network, *they both would be removed* because neither of these edges is >.005 points higher than the other.

The consequences of Forbes et al.'s (2017) procedure are illustrated in Figure 2. When the relative importance network is computed on the NCS-R data as described by Robinaugh et al. (2014), the resulting network retains 118 edges (see the left panel). Using nonnormalized *lmg* with the same threshold results in a network that retains 99 edges (see the middle panel). Finally, applying the deviant thresholding procedure used by Forbes et al. duplicates their analysis, leaving only 31 out of the original 118 edges (see the right panel; the red edges in the middle panel network indicate those removed by Forbes et al.'s thresholding rule). Occasionally this procedure indeed deletes both edges between two nodes; for example, both edges between *even* and *ctrl* are deleted, as one can see by comparing the correctly computed network (see Figure 2, left panel) to the network reported by Forbes et al. (see Figure 2, right panel).

Thus, these analyses did not replicate the standard procedure introduced by Robinaugh et al. (2014), or any other procedure currently in the literature, and introduced a thresholding rule that caused many edges, including some of the strongest, to be deleted. We suggest that, as a result, the conclusions presented by Forbes et al. (2017) that pertain to relative importance networks are not trustworthy and that our results, as presented in Table 1, should be consulted instead. It should be noted that these results should still be interpreted with care, because it is unclear whether relative importance networks, as used by Robinaugh et al. on continuous data, generalize well to the binary data analyzed here in the first place; relative importance networks are computed using linear regressions, which introduces an inappropriate distributional assumption. However, in contrast to the inaccuracies mentioned, we were not able to resolve this in the current work; hence, the reader should keep in mind that both Forbes et al.'s article and our reanalysis are based on an incorrect distributional assumption insofar as relative importance networks are concerned.

A second issue that we consider to qualify as a statistical inaccuracy concerns Forbes et al.'s (2017) use of a distorted tetrachoric correlation matrix, which underlies both their factor analyses and their association networks. To see why this correlation matrix is distorted, first note that the Composite International Diagnostic Interview (Kessler & Ustün, 2004), which yielded the symptom data, involves a skip structure. This means that the full symptomatology of MDE is interrogated only if at least one of the core symptoms of *depressed mood* and *loss of interest* was present; the full symptomatology of GAD is interrogated only if the interviewee reported the presence of *anxiety*, *anxiety about multiple events,* and *loss of control about the worry*. As such, both the NCS-R and the NSMHWB data contain a high percentage of missing values. In both data sets, Forbes et al. imputed 0s for these missing values. This practice assumes Guttman scale properties for the skipped symptoms; that is, if one does not have the symptom of *feeling sad* over a period of 2 weeks (a symptom that acts as a gateway in the skip structure), one cannot have the symptom of *insomnia* (a nongateway item). This practice is acceptable in many contexts, and although the procedure can strongly affect all network models, it does not necessarily invalidate their results. For instance, as can be seen in Figure 1, the GAD skip structure translates to the sequence *anx* → *eve* → *ctrl* in the DAG, with *eve* being the most important gateway item connecting to the other symptoms, whereas the MDE skip structure translates to the sequence *depr* → *inte* in MDE. These sequences accurately reflect the actual order of the symptoms in the interview, and thus the DAGs correctly pick up the skip structure, which reflects a true causal structure in the data (see also Borsboom & Cramer, 2013, Figure 7).

Unfortunately, however, imputing zeros for missing values is not advisable when the goal is to analyze or represent a correlation matrix. This is because it alters the correlations in the data enormously, as is graphically represented in Figure 3. To give an indication of how serious these distortions are, we note that the *average* correlation between depression symptoms in the correlation matrix as used by Forbes et al. equals .94 for the NCS-R data and .96 for the NSMHWB data. This is unrealistically high and nowhere near the average tetrachoric correlation of .33 that characterizes the data if missing values are handled with, for example, pairwise deletion. Also, these values do not resemble correlations typically found for these kinds of symptoms (e.g., see Beard et al., 2016).

In addition, the imputation process introduces deterministic dependencies in the data, which in this case leads the correlation matrices for both the NCS-R and the NSMHWB data to become nonpositive definite (this means that the matrices do not have the characteristics every proper correlation matrix should have and, therefore, should not be used in standard statistical analyses). As a result, these correlation matrices are untrustworthy and unrepresentative of the associations present in the data. Because of this, the results of both association networks and factor analyses reported by Forbes et al. are unreliable. Note that the effects of the imputation strategy are visible in all analyses that Forbes et al. reported and that they affected our analyses in the same way. At present we are unaware of an analytic strategy that could address this issue satisfactorily.

It is important to recognize that, because of the problems we outlined, all statistics reported by Forbes et al. (2017) that pertain to association networks and relative importance networks are either inaccurate or corrupted to an unknown extent by Forbes et al.'s imputation strategy. This has direct consequences for their findings with respect to cross-method replicability. For example, their abstract presents, as a main result, that "only 13%–21% of the edges were consistently estimated across these networks" (p. 1000). These percentages are uninformative, not only because one does not in fact expect different networks to converge upon the same structure, as explained in the previous section, but also because the underlying computations are compromised by statistical inaccuracies, as identified in this section. In fact, the only interpretable results on cross-method replicability that Forbes et al. could have obtained pertain to the comparison between Ising models and DAGs, because these are the only models that they estimated without problems.[9] With respect to this comparison, however, Forbes et al. claimed that 41 edges of the NCS-R–DAG were also present in their NCS-R–Ising model (see their footnote 7). Unfortunately, their NCS-R–DAG contained only 34 edges, which means it is impossible that 41 edges would replicate. We

[9] The reader should take care to interpret this statement as applying to Forbes et al.'s (2017) published article (accompanying this commentary) and not to the version that the authors had made publicly available, which did not implement DAGs correctly and which the authors subsequently corrected.

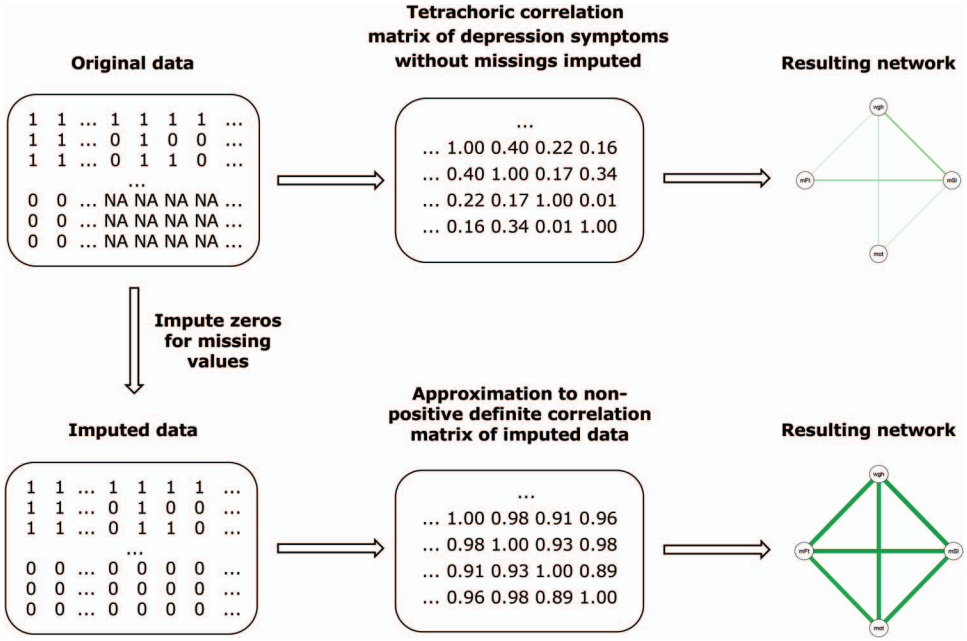

Figure 3. The effect of the imputation strategy used by Forbes, Wright, Markon, and Krueger (2017) on the network structure of the National Comorbidity Survey Replication (NCS-R) data. The figure shows (a) how imputation alters the tetrachoric correlation between depression symptoms and (b) the resulting networks. The correlations shown represent actual NCS-R correlations between four nonskip items (weight problems, sleep problems, psychomotor problems, and fatigue) before and after imputation. See the online article for the color version of this figure.

therefore have no other option than to conclude that none of the statistics on cross-method replicability reported by Forbes et al. are accurate.

Debatable Methodology for Assessing Replicability

After pointing out the statistical inaccuracies in Forbes et al.'s (2017) analyses, this section covers the methodology used to evaluate the replicability of network models. In contrast to the issues mentioned in the previous section, one can have legitimately different points of view on the appropriateness of the measures in question and the importance of the problems they encounter. In our view, the main problem with Forbes et al.'s assessment of replicability is that they do not use any measures that would seem of immediate relevance to any such analysis (e.g., correlations between the edge weights across samples, as reported in Table 1, or statistical tests such as the NCT) and instead rely on several replicability and stability measures that have not been validated and that are problematic for reasons we explain in this section.

First, Forbes et al. (2017) computed the *percentage* of change of the value of a parameter from one data set to the next and then averaged this percentage over all parameters. This percentage is relative to the original size of the edge. This means that small changes in parameters very close to zero can result in huge differences: For instance, when the same parameter is .00001 in one dataset and .00003 in the second, the computations of Forbes et al. convert this into a 300% change, which may be entirely inconsequential for the interpretation of the network structure. Figure 4 (left panel) illustrates, for the Ising model, how it is possible for parameter values to feature an average 30% change across data sets, even though the network parameters are in fact nearly identical. The reason is indeed that large percentage changes are much more likely to occur in small edge weights: Strong edge weights hardly change at all. As a result, the correlation between edge weights remains extremely high (see Figure 4, right panel).

To show that this problem arises in latent variable models as well as networks, we also computed Forbes et al.'s (2017) measure for the parameters of a two-dimensional item response theory model fitted on the NCS-R data; when replicating this model on the NMSHWB data, the percentage parameter change equals 44%, whereas the correlation between the discrimination parameters in the two samples equals .96. Moreover, a small simulation in which we simulated data from a two-factor model and applied Forbes et al.'s measure resulted in an average parameter change of no less than 483%, even though the parameters of the model correlated .99 across samples. Thus, factor models show roughly the same behavior as do network models with this measure.

We conclude that it is inadvisable to attach normative evaluations to the absolute estimates of this metric, as Forbes et al. (2017) did when they interpreted the percentage differences in parameter estimates ("these are all substantial changes in the context of a model that is promoted for its specificity"; p. 1013). The average parameter change metric may be productively used in various methodological investigations (e.g., to compare different models or estimation routines in simulation studies), but it is unfit to serve as an arbiter of replicability.

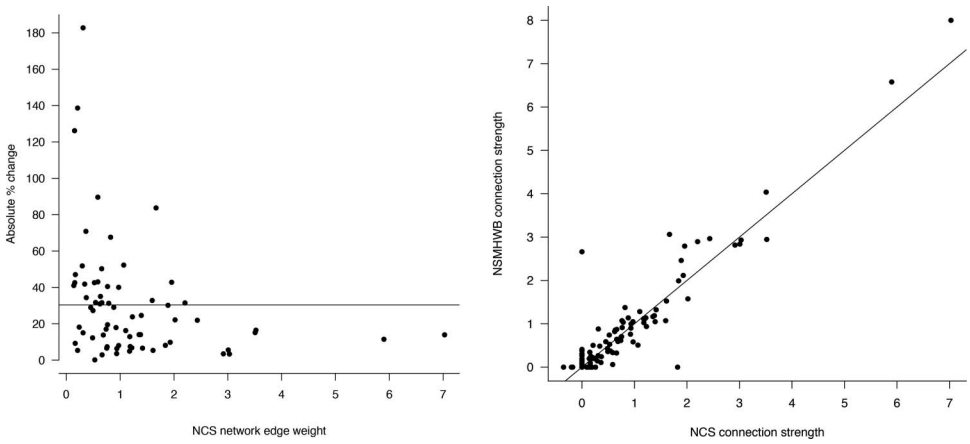

Figure 4. The absolute percentage change in edge weights across data sets relative to the size of the edge weights (left panel). This panel shows that smaller edge weights show larger changes expressed as a percentage of the original weight. The right panel shows that these changes are mostly irrelevant: The strong linear relation ($r = .95$) between edge weights in the National Comorbidity Survey Replication (NCS-R) and National Survey of Mental Health and Wellbeing (NSMHWB) data (right panel) is unaffected by the parameter changes.

Second, Forbes et al. (2017) considered how well the *absolute* position of nodes in the centrality ordering replicates, that is, the question whether a node that ranks 6th in one data set also ranks 6th in the other. Because edge weights and centrality measures are, as are all other statistics, affected by sampling error, nodes can shift positions in the rank ordering due to sampling fluctuations. How strongly sampling fluctuations affect these statistics depends on (a) the sample size and (b) the differences between nodes in terms of centrality at the population level (i.e., the network structure). Epskamp et al. (2017) gave the extreme example where, at a population level, there are no differences in centrality at all (i.e., all nodes are equally central). In this case, one should not expect that order to replicate at all, because any absolute ordering differences in a given sample must be due to sampling error.

Therefore, instead of expecting the orderings to replicate by default, one should inspect both the network structure and the *sampling variability* of centrality measures, which shows how reliably they are estimated and whether differences between them are statistically significant. Fortunately, the R package *bootnet* (Epskamp et al., 2017) can be used for this purpose. Running *bootnet* on the Ising model results obtained by Forbes et al. (2017) showed that most of the edge weights, which are the basis of centrality calculations, were estimated reliably (see Appendix C of the online supplemental materials); however, the edges related to the gateway items used in the skip structure (especially Item 11, which is the symptom *being anxious about multiple events*) are much less reliable, which is likely due to structural zeros in the contingency tables for these items, as induced in Forbes et al.'s treatment of missing values. Inspecting the robustness of the centrality ordering itself reveals that although strength centrality is estimated stably, closeness and betweenness were much less stable. Appendix C of the online supplemental materials explains that this is due to a particularity in the data that likely results from the skip structure; hence, one should hesitate to generalize this result to other data sets or modeling contexts. We advise that, in future research, investigators would do well to interpret centrality results in the context of a robustness analysis using *bootnet*.

In addition, correspondence of the absolute positions of nodes in the centrality ordering across samples, to which Forbes et al. (2017) attached primary significance, is extremely strict as a primary measure of replicability. To see this, suppose one has 26 nodes, corresponding to the letters in the alphabet, for which centrality measures induce the ordering A, B, . . ., Z in one data set. Then one executes the same analysis in another data set, which yields the ordering Z, A, B, C, . . . Y. Because none of the variables occupy the exact same place in the ordering, Forbes et al. would interpret this as evidence that psychopathology networks do not replicate (in fact, there would be no correspondence at all in this case). However, only Z changed position, from least to most central, and although no node occupies the exact same absolute position, one should at the same time conclude that the centrality order does replicate to a large degree, because the *relative* positioning is nearly entirely preserved. This does not invalidate Forbes et al.'s measure of correspondence in absolute position, which can still be useful, but it does mean that this metric should be viewed with caution and, of importance, should always be assessed (a) in the light of stability of the relative positioning of nodes as assessed by the correlation between centrality scores of nodes across samples (e.g., .94 for strength and betweenness and .76 for closeness in the Ising model) and (b) in the light of sampling variability.

Third, Forbes et al. (2017) expressed concern over the fact that different centrality measures identified different nodes as central. However, just as the various network estimation methods get at different aspects of the data and should not be expected to yield the same network solution, centrality measures such as strength, betweenness, and closeness are not interchangeable measures that will converge on "the most influential node," as Forbes et al. suggested (p. 1011). Instead, they are indices that assess different *kinds* of centrality. Thus, if strength centrality is highest for *depressed mood* but *fatigue* shows the highest score on closeness, or when *anxiety about multiple events* has the highest strength in the Ising model but *depressed mood* has the highest strength in the DAG, that signals neither a problem nor a cause for concern.

Instead, these results, if robust across samples and assessment methods, should be viewed as potentially important clues about the structure of a psycho(patho)logical construct under consideration.

So What About Measurement Error?

Because the various network models replicate very well across data sets, the reader may wonder how this fits in with Forbes et al.'s (2017) explanation of the supposed poor replication results in terms of measurement error. That is, Forbes et al. hypothesized that, because edges between two nodes were controlled for other nodes in the network, networks primarily work on residual variances that are largely composed of measurement errors. The results of our reanalysis provide a direct refutation of this theory: If Forbes et al.'s explanation were correct, one should expect bad replicability, but our analyses in fact show replicability to be good. Also, if Forbes et al.'s explanation were correct, one would expect simulation studies and robustness analyses to show that network models produce unreliable results, which is not the case (Epskamp et al., 2017; van Borkulo et al., 2014).

Indeed, despite the suggestive Venn diagrams used in Forbes et al.'s (2017) article, the psychometric intuitions that underlie their argumentation are inaccurate. The following thought experiment may help elucidate why this is the case. Suppose one encountered a situation in which all systematic relations between depression symptoms were due to a latent variable and everything else was pure random measurement error. If Forbes et al. were correct, this would imply that a network model should be expected to return a spurious network without any robust connections: After all, because in their view partial correlations are largely correlations between measurement errors, and measurement errors are not structurally related, there is nothing real for the network to go on. However, this is not what one would find: If a latent variable model gave rise to all correlations between variables, then one would not find an empty network but a fully connected one (Ellis & Junker, 1997; Epskamp et al., in press). Thus, a latent variable model corresponds to a dense network of systematic relations (Marsman, Maris, Bechger, & Glas, 2015) and not to an empty or spurious network, as Forbes et al.'s theory would suggest.

More generally, one can prove that every latent variable structure implies a specific network structure, as Molenaar (2003) already suspected and as Maris and his coworkers have been recently able to formally prove (Epskamp et al., in press; Kruis & Maris, 2016; Marsman et al., 2015). Thus, even though network models do not explicitly represent shared variance in a separate node that renders the other nodes conditionally independent (i.e., a latent variable), they do imply the presence of shared variance in sets of connected nodes. In fact, given that the known mathematical equivalence relations between the models implies that they produce the same joint probability distribution for the items, the models should not be expected to differ in this respect. This has the somewhat ironic consequence that, if network structures replicated badly across two data sets, then this would imply that factor structures (i.e., the configuration of loadings in exploratory factor models) would replicate badly as well. Measurement error has little to do with this, because both latent variable models and network models operate on the same systematic relations in the data.

Despite this, however, we do note that additional methodological research is necessary to systematically study the replication properties of different models under various conditions, because these would likely be influenced by various factors such as the overall fit of the model, the number of parameters (and an important caveat of network models is that they typically do require many parameters to be estimated), and the strength of the associations in the data. Psychometric intuition, however, is an unreliable guide in this respect. Thus, mathematical analyses and simulation studies are required to study these issues, especially when making critical generalized claims about an entire psychometric field based on the analysis of two data sets.

Best Practices for Future Research

Despite the inadequacy of the data and analyses used by Forbes et al. (2017), we stress again that we consider both stability and replicability of networks to be extremely important topics. Therefore, we commend Forbes et al. for taking up these issues. Regarding stability, we agree with Forbes et al. that model stability should be tested in all statistical models, including both network and factor models. Thus, we hope that Forbes et al.'s article—together with the *bootnet* R package and the accompanying tutorial article (Epskamp et al., 2017)—will shift the attention of both applied and technical researchers to this topic. Regarding replicability, we offer a roadmap for network replication studies in this section that may aid future researchers in obtaining more objective and trustworthy results.

The Method: Replication as a Nonempirical Question

First, we address a central issue in the design of Forbes et al. (2017): They confound evidence for replication problems that concern a *particular estimated model* with evidence for problems of *the model in general*. This is a non sequitur. For suppose that one fitted a specific regression model to two different samples and the regression coefficients were different from each other. Nobody would conclude from such a result that "regression analysis has limited replicability". The problem with equating "not the same result in two data sets" to "method does not work" is that one does not know whether the "true" relationship between variables is the same across samples. In the absence of this knowledge, one cannot know for sure whether differences in results are due to differences in sample characteristics or to a flawed method.

One may think that this problem is circumvented in Forbes et al.'s (2017) evaluation of split-half results, which are based on the correspondence of networks within the same sample. However, this only partly addresses the problem. First, because one does not know whether split-half performance with this particular kind of data (here: MDE–GAD symptom data obtained with interviews containing skips) generalizes to other kinds of data, as is necessary for blanket statements like "popular network analysis methods produce unreliable results," as touted in Forbes et al.'s (2017) General Scientific Summary. Second, because even in a given sample one does not know whether any given network model is true, let alone which one, and in the absence of this knowledge it is impossible to assess which part of model instability arises from defects in the methodology and

which part arises from model misfit, population heterogeneity, violations of distributional assumptions, and so forth.

Thus, if the primary aim of research is to assess the general methodological adequacy of a method, the evaluation of two specific empirical data sets is of limited use. Putting a network *method* to the test requires that one know the "true" network structure, and this can be done only by (a) establishing mathematical proof that the method converges on the true structure in the long run (as, e.g., Meinshausen & Bühlmann, 2006, have done for the Gaussian graphical model and Ravikumar, Wainwright, & Lafferty, 2010, for the Ising model) or (b) simulating such "true" network structures and, subsequently, assessing the capability of a method, in a variety of settings, to accurately estimate that "true" network structure (as executed by van Borkulo et al., 2014, for the Ising model). This motivates the rule that *methodological adequacy should be established on methodological grounds.*

Network Structure of a Psychological Construct: Replication as an Empirical Question

Once a particular method is proven to accurately retrieve a "true" network structure using methodological studies, there is another question of replicability that is empirical in nature; namely, what is the particular network structure of a psychological construct such as major depression, or generalized anxiety disorder? Answering this question *does* entail the comparison of network structures across many data sets and many participants. As we have shown earlier, the design used by Forbes et al. (2017) is suboptimal in this respect, and this raises the question what kind of methodological design would be needed to properly assess replicability in network analysis. Although the following list is not meant to be exhaustive (see Anderson & Maxwell, 2016, for additional issues in replication research), we suggest these best practices:

1. *No skip structure.* If one desires a replicability assessment that is not confounded by methodological design, one needs data that do not contain a skip structure. We realize this may be a challenge given that many data sets, such as NCS-R and NSMHWB, do contain a skip structure. We also realize that we are guilty as charged in this respect because we, too, used NCS-R data, albeit it for illustration or hypothesis-generating purposes (Borsboom & Cramer, 2013; Cramer, Waldorp, van der Maas, & Borsboom, 2010). Also, in certain cases there is no other option than to use a design with skip structures (e.g., one cannot ask persons who do not drink whether they got into legal problems because of drinking; Rhemtulla et al., 2016). Future studies in data sets without skip structure will enable one to gauge the replicability of psychopathology networks, and we are glad to see that such studies are already on the way (e.g., network replicability across four large clinical posttraumatic stress disorder data sets; Fried et al., 2017).

2. *Open access data and code.* Reproducibility studies should themselves be reproducible. The NSMHWB data used in this research, however, are not publicly accessible, which means that third parties cannot replicate either our results or those of Forbes et al. (2017) without engaging in a lengthy, cumbersome, and costly procedure to gain access to the data (we were charged US$947 just to be able to check the veracity of Forbes et al.'s analyses). This is highly undesirable. Replication studies are different from other studies in that their consequences may be more far-reaching, because they can discredit or invalidate whole research programs. Therefore, one must be sure that the analyses and reported results are sound. The only way interested third parties can verify this is through free access to the data used. We acknowledge that freely available data sets containing clinical patient data may be challenging, for example due to issues concerning extending informed consent of patients to third parties. However, we feel encouraged that important progress is forthcoming due to a recent article about replicability in clinical science that contains a multitude of valuable recommendations (Tackett et al., in press). Analysis code should naturally always be available, because it is needed to replicate and verify reported analyses—the current report illustrates how important this is—and we commend Forbes et al. for sharing their code.

3. *Preregistration of analyses.* Replication research differs from the exploratory designs in which network analyses are most often used, because researchers have a clear idea of the hypothesis to be tested: replication across samples. In addition, especially in replication research, the selection of measures used to gauge replicability is important: After the data are in, it is always possible to come up with a particular selection of measures that emphasizes evidence for or against replicability. To minimize the influence of subjective choices made after the data are in, we encourage any replicability effort to be preregistered, for example at the Open Science Framework (OSF). Preregistration has an additional advantage, because interested researchers are able to check (a) a priori hypotheses and (b) the analysis plan. The OSF also allows for uploading the code that was used for the analyses, so other researchers can check the veracity of the reported results *before* the article is even submitted for review. This reduces the probability of submitting or even publishing article that later turn out to be ill founded. In the current study, such a procedure would have safeguarded against the statistical inaccuracies manifest in Forbes et al. (2017).

Conclusion

We think that practically all researchers are united by a common goal: the pursuit of scientific knowledge. As such, we stress the importance of expanding the knowledge about psychopathological networks and acknowledge the challenges ahead (Fried & Cramer, 2017). If one day we were to find out that networks are either not replicable or that they cannot be suitable candidate models for explaining psychopathology, then we would consider this a victory for clinical science—despite our investment in these models. Falsification is an essential component of the scientific enterprise, and the burden of doing so should fall on everyone and on all theories and hypotheses.

In our comprehensive reanalysis, however, we have shown that Forbes et al.'s (2017) devastating conclusions are not licensed by their analysis. We conclude that the main conclusion of Forbes et al. that "popular network analysis methods produce unreliable results" is a strongly overstated generalization that is not warranted on the basis of their research design and statistical analyses. The replicability issue, however, is not settled with the publication of either Forbes et al.'s article or our commentary. It is for this reason that we have formulated best practices to investigate the important

replication issue properly, by using adequate data and optimal analysis designs. Our hope is that future work will lead toward the robust and replicable scientific knowledge that everyone should be looking for.

References

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods, 21,* 1–12. http://dx.doi.org/10.1037/met0000051

Beard, C., Millner, A. J., Forgeard, M. J. C., Fried, E. I., Hsu, K. J., Treadway, M. T., . . . Björgvinsson, T. (2016). Network analysis of depression and anxiety symptom relationships in a psychiatric sample. *Psychological Medicine, 46,* 3359–3369. http://dx.doi.org/10.1017/S0033291716002300

Borsboom, D., & Cramer, A. O. J. (2013). Networks: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology, 9,* 91–121.

Borsboom, D., Fried, E. I., Epskamp, S., Waldorp, L., van Borkulo, C., van der Maas, H., & Cramer, A. (2017, July 28). *Codes: False alarm? A comprehensive reanalysis of "Evidence that psychopathology symptom networks have limited replicability" by Forbes, Wright, Markon, and Krueger*. Retrieved from https://osf.io/akywf/

Bryant, R. A, Creamer, M., O'Donnell, M., Forbes, D., McFarlane, A. C., Silove, D., & Hadzi-Pavlovic, D. (2017) Acute and chronic posttraumatic stress symptoms in the emergence of posttraumatic stress disorder: A network analysis. *JAMA Psychiatry, 74,* 135–142. http://dx.doi.org/10.1001/jamapsychiatry.2016.3470

Cramer, A. O. J., Waldorp, L. J., van der Maas, H., & Borsboom, D. (2010). Comorbidity: A network perspective. *Behavioral and Brain Sciences, 33,* 137–193.

Ellis, J. L., & Junker, B. W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika, 62,* 495–523. http://dx.doi.org/10.1007/BF02294640

Epskamp, S. (2017). *Network psychometrics* (Doctoral dissertation). Retrieved from http://sachaepskamp.com/Dissertation

Epskamp, S., Borsboom, D., & Fried, E. I. (2017). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*. Advance online publication. http://dx.doi.org/10.3758/s13428-017-0862-1

Epskamp, S., Maris, G., Waldorp, L. J., & Borsboom, D. (in press). Network psychometrics. In P. Irwing, D. Hughes, & T. Booth (Eds.), *Handbook of psychometrics*. New York, NY: Wiley.

Forbes, M., Wright, A., Markon, K., & Krueger, R. (2017). Evidence that psychopathology symptom networks have limited replicability. *Journal of Abnormal Psychology, 126,* 1011–1016.

Fried, E. I., Eidhof, M. B., Palic, S., Costantini, G., Huisman-van Dijk, H. M., Bockting, C. L. H., . . . Karstoft, K.-I. (2017). *Replicability and generalizability of PTSD networks: A cross-cultural multisite study of PTSD symptoms in four trauma patient samples*. http://dx.doi.org/10.17605/OSF.IO/2T7QP

Fried, E. I., & Cramer, A. O. J. (2017). Moving forward: Challenges and directions for psychopathological network theory and methodology. *Perspectives on Psychological Science*. Advance online publication. http://dx.doi.org/10.17605/OSF.IO/BNEKP

Fried, E. I., van Borkulo, C. D., Cramer, A. O. J., Boschloo, L., Schoevers, R. A., & Borsboom, D. (2017). Mental disorders as networks of problems: A review of recent insights. *Social Psychiatry and Psychiatric Epidemiology, 52,* 1–10. http://dx.doi.org/10.1007/s00127-016-1319-z

Heeren, A., & McNally, R. J. (2016). An integrative network approach to social anxiety disorder: The complex dynamic interplay among attentional bias for threat, attentional control, and symptoms. *Journal of Anxiety Disorders, 42,* 95–104. http://dx.doi.org/10.1016/j.janxdis.2016.06.009

Hoorlebeke, K., Marchetti, I., De Schryver, M., & Koster, E. H. W. (2016). The interplay between cognitive risk and resilience factors in remitted depression: a network analysis. *Journal of Affective Disorders, 195,* 96–104. http://dx.doi.org/10.1016/j.jad.2016.02.001

Kessler, R. C., & Ustün, T. B. (2004). The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *International Journal of Methods in Psychiatric Research, 13,* 93–121.

Kruis, J., & Maris, G. (2016). Three representations of the Ising model. *Scientific Reports, 6,* 34175. http://dx.doi.org/10.1038/srep34175

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2015). Bayesian inference for low-rank Ising networks. *Scientific Reports, 5,* 9050. http://dx.doi.org/10.1038/srep09050

McNally, R. J. (2016). Can network analysis transform psychopathology? *Behaviour Research and Therapy, 86,* 95–104.

McNally, R. J., Mair, P., Mugno, B. L., & Riemann, B. C. (2017). Co-morbid obsessive-compulsive disorder and depression: A Bayesian network approach. *Psychological Medicine, 47,* 1204–1214. https://doi.org/10.1017/S0033291716003287

McNally, R. J., Robinaugh, D. J., Wu, G. W., Wang, L., Deserno, M. K., & Borsboom, D. (2015). Mental disorders as causal systems: A network approach to posttraumatic stress disorder. *Clinical Psychological Science, 3,* 836–849.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics, 34,* 1436–1462. http://dx.doi.org/10.1214/009053606000000281

Molenaar, P. C. M. (2003). *State space techniques in structural equation modeling*. Retrieved from http://bit.ly/2ssau1K

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). http://dx.doi.org/10.1017/CBO9780511803161

Ravikumar, P., Wainwright, M. J., & Lafferty, J. D. (2010). High-dimensional Ising model selection using l1-regularized logistic regression. *Annals of Statistics, 38,* 1287–1319. http://dx.doi.org/10.1214/09-AOS691

Rhemtulla, M., Fried, E. I., Aggen, S. H., Tuerlinckx, F., Kendler, K. S., & Borsboom, D. (2016). Network analysis of substance abuse and dependence symptoms. *Drug and Alcohol Dependence, 161,* 230–237. http://dx.doi.org/10.1016/j.drugalcdep.2016.02.005

Robinaugh, D. J., LeBlanc, N. J., Vuletich, H. A., & McNally, R. J. (2014). Network analysis of persistent complex bereavement disorder in conjugally bereaved adults. *Journal of Abnormal Psychology, 123,* 510–522.

Santos, H. J., Fried, E. I., Asafu-Adjei, J., & Ruiz, J. (2017). Network of perinatal depressive symptoms in Latinas: Relationship to stress-related and reproductive biomarkers. *Research in Nursing & Health, 40,* 218–228. http://doi.org/10.1002/nur.21784

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., . . . Shrout, P. E. (in press). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science*.

van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A., & Waldorp, L. J. (2014). A new method for constructing networks from binary data. *Scientific Reports, 4,* 5918. http://dx.doi.org/10.1038/srep05918

van Borkulo, C. D., Boschloo, L., Kossakowski, J., Tio, P., Schoevers, R. A., Borsboom, D., & Waldorp, L. J. (2016). *Comparing network structures on three aspects: A permutation test*. Manuscript submitted for publication. http://dx.doi.org/10.13140/RG.2.2.29455.38569

Received May 16, 2017
Revision received July 18, 2017
Accepted July 18, 2017 ■