Further evidence that psychopathology networks have limited replicability and utility: Response to Borsboom et al. (2017) and Steinley et al. (2017)

Article in Journal of Abnormal Psychology · October 2017
DOI: 10.1037/abn0000313

4 authors, including:

Miriam K Forbes
University of Minnesota Twin Cities
36 PUBLICATIONS 358 CITATIONS

Aidan G. C. Wright
University of Pittsburgh
156 PUBLICATIONS 3,382 CITATIONS

Some of the authors of this publication are also working on these related projects:

Sexual dysfunctions in the internalizing spectrum View project

Higher-order structure of psychopathology View project
Further evidence that psychopathology networks have limited replicability and utility:

Response to Borsboom et al. and Steinley et al.

Miriam K. Forbes
University of Minnesota

Aidan G. C. Wright
University of Pittsburgh

Kristian E. Markon
University of Iowa

Robert F. Krueger
University of Minnesota

This pre-print is the authors’ final copy accepted for publication at Journal of Abnormal Psychology. This article may not exactly replicate the final published version. It is not the copy of record.

Author note: Please note that the published version of the target article (Forbes, Wright, Markon, & Krueger, this issue) includes corrections made to the results for the directed acyclic graphs in the preprint that was available on the Open Science Framework from April 7th-24th 2017. Miriam Forbes’ work on this commentary response was supported by a NIDA training grant (T320A037183). Aidan Wright’s efforts were supported by NIMH (L30MH101760). Robert Krueger’s efforts were supported by NIH (R01AG053217; U19AG051426) and by the Templeton Foundation.

Word count:
Main text: 3100
Abstract: 166
**Abstract**

In our target article, we tested the replicability of four popular psychopathology network estimation methods that aim to reveal causal relationships among symptoms of mental illness. We started with the focal data set from the two foundational psychopathology network papers (i.e., the National Comorbidity Survey–Replication) and identified the National Survey of Mental Health and Wellbeing as a close methodological match for comparison. We compared the psychopathology networks estimated in each dataset—as well as in ten sets of random split-halves within each dataset—with the goal of quantifying the replicability of the network parameters as they are interpreted in the extant psychopathology network literature. We concluded that current psychopathology network methods have limited replicability both within and between samples, and thus have limited utility. Here we respond to the two commentaries on our target article, concluding that Steinley, Hoffman, Brusco and Sher’s (2017) findings—along with other recent developments in the literature—provide further conclusive evidence that psychopathology networks have poor replicability and utility.

**General Scientific Summary:** This response to the two commentaries on our target article reiterates the conclusion that popular network analysis methods produce unreliable results.
Introduction

The two commentaries on our target article (Forbes, Wright, Markon, & Krueger, 2017), from advocates of psychopathology networks (Borsboom et al., 2017) and scholars versed in multivariate statistical methodology (Steinley, Hoffman, Brusco & Sher, 2017) differ substantially in their assessment of our findings. Borsboom et al. question the accuracy and validity of our work, and present a re-analysis of the data that they suggest provides evidence that “network models replicate very well” (p. 3). We argue this conclusion is not consistent with the data, and is at odds with emerging evidence in the literature as well as with many of their own previously stated conclusions. Steinley et al. describe an innovative method to test whether network parameters and measures of correspondence between networks differ from what would be expected in random data (i.e., by chance). They conclude that “the problem is likely worse” than our results indicated (p. 1). Our conclusions closely align with Steinley et al., and we consider their proposed method to be a key contribution to strengthening the network literature. Given space constraints here, we therefore spend much of this response briefly addressing the key issues raised by Borsboom et al. (see online supplement Table S1 for our response to some of the less central points). We then focus on evidence from others’ work—including Steinley et al.’s commentary—that the replicability of psychopathology networks remains a substantial problem. Indeed, replicability is only one of many problems facing this new field.

Borsboom et al.’s main criticisms of our article

The use of zero-imputation to account for the skip-structure in the structured diagnostic interviews of the two surveys biased our results. It is remarkable that Borsboom et al. chastise us for this, given they used the same process on the same data to argue for the utility of network models (Borsboom & Cramer, 2013; Cramer et al., 2010). In fact, this was their methodological decision, which we emulated in an effort to mirror their methods—“We
interpreted missing values that arose from the skip structure of the questionnaire as absent symptoms and replaced these by zeros, which seems a reasonable course of action given the way the DSM-IV is set up.” (Borsboom & Cramer, 2013, p. 104). We subsequently emulated this approach in the NSMHWB data, as have all other network analyses of psychopathology data based on structured interviews with a skip-structure that we know of (e.g., Boschloo et al., 2015; Rhemtulla et al., 2016). Zero-imputation is thus a potential limitation of extant network approaches, rather than a “statistical inaccuracy” (p. 15) of our work.

**The methods we used to estimate the relative importance networks render these results invalid.** Our goal was to test the replicability of the psychopathology networks as they are interpreted in the extant literature. Studies that include relative importance networks routinely emphasise stronger edges with high relative importance in their interpretation—particularly with respect to purported clinical utility (e.g., Bryant et al., 2017; Heeren & McNally, 2016; Hoorelbeke, Marchetti, De Schryver & Koster, 2016; McNally, 2016; Robinaugh, LeBlanc, Vuletich & McNally, 2014). We therefore chose to examine the replicability of these edges, as well as of uncensored relative importance networks (i.e., with all edges estimated). However, Borsboom et al. point out that our censoring rule for the censored relative importance networks had the unintended effect of removing both edges in pairs with nearly identical edge weights. This point is well taken and we are happy to accept that these particular results may not reflect the expected replicability of the extant relative importance network literature. We also note that in Borsboom et al.’s analyses the relative importance networks had better replicability—and may therefore represent a better avenue for future research—than the putatively “state-of-the-art” Ising models, which are used in the majority of psychopathology network research.

**The assumptions underlying our methods were misguided.** Borsboom et al. stress “the importance of assessing stability and replicability of network structures stands beyond
They also emphasise their openness to evidence that network models are flawed in these regards, as this would represent progress towards researchers’ common goal of robust and replicable scientific knowledge. However, they readily reject the results of our target article, suggesting that the design of our study was not appropriate for assessing replicability. Specifically, they suggest that it is not reasonable to expect that networks should replicate in methodologically matched samples—or even in two random halves of a single dataset (p. 25-26). This position is perplexing to us, given that generalisability and replicability are both fundamental to the utility of research (e.g., why estimate psychopathology networks at all if we do not think they are able to identify meaningful relationships?). In fact, a review paper from their own group specifically recommends “cross-validation across similar samples” to investigate replicability (Fried & Cramer, in press, p. 38).

Borsboom et al. also argue that networks should not be expected to replicate across different network estimation methods (e.g., Ising models, relative importance networks, and directed acyclic graphs [DAGs]), citing the fact that they each “get at different aspects of the data” (p. 22) and differ in sensitivity and specificity. However, Borsboom et al. do propose that concordance between methods can be examined based on a nesting approach that tests only the edges estimated in sparser networks (e.g., DAGs) against the edges estimated in less sparse networks (e.g., Ising models). Their suggested approach highlights the predictable conclusion that the edges from the sparser DAGs were replicated in the less sparse Ising models within each sample, but offers no guidance for how to interpret the other cross-

1 Indeed, this goal was the ultimate motivation for writing our target article.
2 This line of argument highlights how important it is that researchers stop using these methods interchangeably as if they identify the same inferential targets (i.e., dynamic causal relationships among symptoms). Instead, it is vital to take the time to determine which method, if any, can uncover true underlying models (e.g., in simulated data with a known structure). Similarly, Borsboom et al. suggest that different centrality measures (e.g., strength, closeness, betweenness) should not be used either interchangeably or in conjunction with one another to identify the most central or influential node in a network. Importantly, both points contradict current practices as reflected in published studies, and represent important new messages for researchers using psychopathology network methods.
method replicability results in their online materials (i.e., that only 41.3–42.5% of the edges in the Ising models were replicated in the DAGs within and between samples; and that all DAG and Ising model pairs had different nodes with the highest in strength, out strength, and closeness centrality). Their conclusion that “Cross-method replication could hardly be better.” certainly appears to be unwarranted (p. 12).

In short, Borsboom et al. suggest that the replicability of network models cannot be tested based on the comparisons of different samples or different methods. Despite this, based on these same comparisons they conclude that networks models replicate very well across different samples and different methods (when using the “right” metrics for quantifying replicability).

We used the wrong metrics for quantifying replicability. Borsboom et al. state “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” (p. 18). It is hard to determine how they arrived at this conclusion, given our metrics were specifically selected to quantify replicability of the network parameters at the level they are used and interpreted in the literature. For example, the key results from our target article are based on the research questions “is edge A—B estimated in both networks?”, if so “how much did the strength of this edge change from network 1 to 2?”, and “does node A have the highest (or second, third, fourth highest, etc.) centrality in both networks?”. These characteristics represent the features of psychopathology networks that correspond to network theory, and their purported clinical utility (Borsboom, 2017; Fried et al., 2016). In contrast, “the most intuitive and important” of the metrics Borsboom et al. propose for assessing replicability (p. 7) is a Pearson correlation between the edge lists of a pair of networks. This translates to the research question “at a global level, how similar is the list of all edges estimated in network 1 to the list of all edges

---

3 As well as how they determined our indices of replicability to be “debatable” (p. 6), while labelling their own as “intuitive” (p. 7) and “powerful tools” (p. 8).
estimated in network 2”. This metric does not have any clear relevance to the inferences psychopathology networks are promoted to provide.

Borsboom et al.’s proposed methods for quantifying replicability “led to results directly opposed” to ours (p. 3) and this is most clear when comparing the conclusions that our study and the two commentaries arrived at regarding the Ising models in the full samples (see Table 1). The differences are remarkable and highlight that not only do Borsboom et al.’s proposed methods fail to align with the level that network parameters are interpreted, but that they have a striking lack of sensitivity. This lack of sensitivity is highly problematic, as it instils a false sense of confidence in results that evidently lack validity. Even when holding all the characteristics of the networks equal between samples (i.e., identical node selection, similar sample size, the same study methods, and the same analytic method) the key details of the model do not replicate. Combined with the results of Steinley et al., it is clear to us that although the global characteristics appear to be replicable (e.g., the number of edges and average edge strength⁴), the detailed parameters of psychopathology networks as currently formulated are too empirically weak to provide a meaningful basis for understanding psychopathology, much less building meaningful theories.

**Evidence of poor replicability from the psychopathology network literature**

Borsboom et al. claim that “inadequacy of the data and analyses” (p. 25) in our target article invalidates the results.⁵ Not surprisingly, we disagree. However, even if this claim were true, one need only look at the emerging network literature to find an abundance of evidence for problems in replicating the results of network analyses. Take, for example, the eight papers we are aware of that examine PTSD symptom networks (i.e., not including other

---

⁴ We also note that interpreting the global network characteristics requires caution; for example, Terluin et al. (2016) found that the differences in edge strength between symptoms can be artefactual, representing differential range restriction across severity subgroups (i.e., should not be interpreted as differences in impact of one symptom on another).

⁵ We remind the reader that these were the same data and analyses used in the major foundational papers for these methods (Borsboom & Cramer, 2013; Cramer et al., 2010) and the basis for Borsboom et al.’s (this issue) conclusion that “network models replicate very well” (p. 3).
disorders or covariates) in adult samples using putatively “state-of-the-art” network methods: Armour, Fried, Deserno, Tsai and Pietrzak, 2017; Birkeland and Heir, 2017; Epskamp, Borsboom and Fried, 2017; Fried et al., under review; McNally, Heeren and Robinaugh, 2017; Mitchell et al., 2017; Spiller et al., 2017; and Sullivan, Smith, Lewis and Jones, 2016. All eight papers use graphical LASSO regularisation to eliminate weak and unreliable edges in conjunction with the bootnet package (Epskamp et al., 2017) to “safeguard against false positive results, and also help us to identify consistent pathways that are highly reliable across studies” (Fried & Cramer, in press, p. 40). Between these eight papers, fourteen (87.5%) of the sixteen symptoms common to DSM-IV and DSM-5 PTSD diagnostic criteria were reported in-text as having particularly high centrality (see Table 2). Six (42.9%) of these purportedly highly influential symptoms were reported as highly central in only a single paper, and none (0%) in a majority of the papers. Similarly, there were fourteen specific edges emphasised as strong and reliable between the eight papers (see Table 3). Only one (7.1%) of these edges was consistently estimated as a strong relationship (i.e., hypervigilance—startle); others varied substantially in strength, and seven (50%) were absent altogether from the PTSD network in at least one of the papers.

In short, the PTSD network literature shows that neither gLASSO regularisation nor bootnet are sufficient for identifying reasonably replicable results. This finding highlights the central role that Steinley et al.’s proposed method should have in psychopathology network research going forward. As Steinley et al. illustrate, network parameters that are significant and stable but not different from random chance are uninteresting and unlikely to replicate. It also leaves us in the difficult situation of not knowing what, if anything, can be

---

6 Birkeland & Heir (2017), McNally et al. (2017), and Mitchell et al. (2017) listed a larger number of highly central symptoms than was supported by bootnet and made this distinction in-text. If we narrow it down to counting only their “bootnet-central” symptoms, eleven symptoms were reported as highly central across the eight papers; five (45.5%) in only a single paper, and none (0%) in the majority of papers.

7 All but one (97.8%) of the estimated edges identified by Steinley et al. as significantly different from chance in the NCS-R Ising model was replicated in the NSMWHB Ising model.
meaningfully concluded from the extant network literature with respect to the onset, maintenance, or treatment of psychopathology.

**The future of psychopathology networks**

The road ahead for empirical psychopathology network research is challenging. While network *theory* (e.g., Borsboom, 2017) calls attention to interesting and potentially important avenues for future research, the current psychopathology network *methodologies* are plagued with substantial flaws. For example, in the few months since our target article was accepted for publication, a series of papers have emerged on substantial problems facing the methods, in addition to poor replicability (Bos et al., 2017; Epskamp et al., 2017; Fried & Cramer, in press; Guloksuz, Pries & van Os, 2017; Steinley et al., 2017; Wichers, Wigman, Bringmann & de Jonge, 2017). Taken together, this growing literature highlights the remarkable sensitivity of psychopathology network results to a multitude of factors including the study design, variables included, characteristics of the data, and analytic methods (e.g., Bulteel, Tuerlinckx, Brose & Ceulemans, 2016; Terluin, de Boer & de Vet, 2016; Wichers et al., 2017). The generalizability of psychopathology networks is further compromised by the fact that cross-sectional symptom networks change over time (Bryant et al., 2017), do not reflect temporal symptom dynamics derived in the same data (i.e., have different symptom-to-symptom associations and different central symptoms; Bos et al., 2017), and that group-level networks are not expected to generalize to individuals (Borsboom & Cramer, 2013; Fried & Cramer, in press; Fried et al., 2016).

It therefore seems fair to conclude—as we did in our target article—that cross-sectional psychopathology networks evidently have limited utility. We reiterate the warning of Steinley et al. that continuing to apply the current methods “runs the very real risk of creating a series of publications that contain results that are not reproducible and likely no different than what is expected under one of the most basic models of chance…” (p. 23).
Developing appropriate methods to ensure the reliability and validity of the results is thus a crucial step that needs to be addressed before the proliferation of psychopathology networks continues.

Ultimately, there is a fundamental disconnect between network theory—which has a central tenet that “mental disorders arise from the causal interaction between symptoms in a network” (Borsboom, 2017, p. 6)—and the vast majority of contemporary network research. Current cross-sectional network study designs and methods do not and cannot test the research questions or hypotheses of network theory (Guloksuz et al., 2017; Wichers et al., 2017). Fried and Cramer (in press) thus conclude that it is up to network researchers “to be careful not to over-interpret results of network analyses as representing reality” (p. 20), and cross-sectional psychopathology networks are instead often framed as a hypothesis-generating exercise (e.g., Borsboom & Cramer, 2013; Fried & Cramer, in press). However, each hypothesis regarding the clinical importance of an individual symptom, or the existence of dynamic causal relationships between symptoms (e.g., each result in Tables 2 and 3) requires an extensive program of research based on the collection of comprehensive longitudinal and/or experimental data. In a few short years, a multitude of hypotheses have been generated in psychopathology network research. It is now time to start the hard work of testing these hypotheses using appropriate study designs and analytic methods.
References


Table 1. Summary of key differences among the Ising model results in the target article and commentaries

<table>
<thead>
<tr>
<th>Network characteristic</th>
<th>Forbes et al.</th>
<th>Steinley et al.</th>
<th>Borsboom et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated edges</td>
<td>13–14% of the edges failed to replicate between NCS-R and NSMHWB</td>
<td>44% of the estimated edges in the NCS-R Ising model were due to chance</td>
<td>Edge lists correlate $r &gt; .95$</td>
</tr>
<tr>
<td></td>
<td>47% of the bridging edges failed to replicate from NCS-R to NSMHWB</td>
<td>28% of the absent edges were due to chance</td>
<td>The network comparison test (NCT) failed to reject the null hypothesis that the structure is completely identical between NCS-R and NSMHWB</td>
</tr>
<tr>
<td></td>
<td>The replicated edges differed in strength, on average, by 30%</td>
<td></td>
<td>NCT found that there were zero significantly different edges between NCS-R and NSMHWB</td>
</tr>
<tr>
<td>Node strength centrality</td>
<td>83% of the nodes did not have the same strength centrality rank-order in both networks</td>
<td>28% of the strength centrality estimates in NCS-R were due to chance</td>
<td>Strength centrality values correlate $r = .94$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>bootnet suggested that the strength centrality parameters were estimated reliably and that rank-order is interpretable</td>
</tr>
<tr>
<td>Conclusion</td>
<td>“These are all substantial changes in the context of a model that is promoted for its specificity (i.e., its ability to detect and exclude false positives from the model)…” and “the poor replicability of the bridging edges is of particular concern.” (p. 14).</td>
<td>The fact that many of the Ising NCS-R results are indistinguishable from chance “renders these types of depictions of networks almost useless” (p. 14).</td>
<td>The NCS-R and NSMHWB Ising models are “nearly identical” (p. 8).</td>
</tr>
</tbody>
</table>

Note. We focus on node strength centrality here because Borsboom et al. conclude that closeness and betweenness estimates are “much less stable” (p. 21) than strength (although it is noteworthy that Epskamp et al.’s (2017) CS-coefficients would suggest otherwise, as all but one are ≥ .46, which exceeds the lenient cut-off of .25 and approaches the conservative cut-off of .5 for inferring stable and interpretable centrality values). NCS–R = National Comorbidity Survey—Replication; NSMHWB = National Survey of Mental Health and Wellbeing.
Table 2. Whether the symptoms in both the DSM-IV and DSM-5 diagnostic criteria were emphasised as highly central in the eight PTSD studies.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrusive thoughts/memories</td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Distressing dreams</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Flashbacks</td>
<td></td>
<td>x</td>
<td></td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>N/A</td>
</tr>
<tr>
<td>Emotional reaction to cues</td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
<td>b</td>
<td>-</td>
<td>x</td>
<td>-</td>
</tr>
<tr>
<td>Physiological reaction to cues</td>
<td></td>
<td>x x</td>
<td></td>
<td>x</td>
<td>b</td>
<td>x</td>
<td>-</td>
<td>N/A</td>
</tr>
<tr>
<td>Avoidance of thoughts</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Avoidance of external reminders</td>
<td></td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>-</td>
</tr>
<tr>
<td>Inability to remember</td>
<td></td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>N/A</td>
</tr>
<tr>
<td>Diminished interest</td>
<td></td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>N/A</td>
</tr>
<tr>
<td>Detachment</td>
<td></td>
<td>x x</td>
<td></td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>x</td>
</tr>
<tr>
<td>Restricted affect</td>
<td></td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>N/A</td>
</tr>
<tr>
<td>Irritable or angry</td>
<td></td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>x</td>
</tr>
<tr>
<td>Hypervigilance</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>N/A</td>
</tr>
<tr>
<td>Exaggerated startle</td>
<td></td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>N/A</td>
</tr>
<tr>
<td>Problems with concentration</td>
<td></td>
<td>- x</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sleep disturbance</td>
<td></td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

*Note. x = noted as highly central in-text; italics if the authors made the distinction that the symptom was not also reliably more central than others; N/A = symptom not included in analyses.

a These results are based on the cross-sample network.
b These symptoms were concatenated in Fried et al.’s analyses, but separated here to maximise concordance with other studies.*
Table 3. Edges between symptoms in both the DSM-IV and DSM-5 diagnostic criteria that were emphasised as strong and reliable in the eight PTSD studies

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrusive thoughts—distressing dreams</td>
<td>x</td>
<td>xx</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>xx</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Intrusive thoughts—flashbacks</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>xx</td>
<td>xx</td>
<td>x</td>
<td>x</td>
<td>N/A</td>
</tr>
<tr>
<td>Intrusive thoughts—emotional reaction to cues</td>
<td>x</td>
<td>xx</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>xx</td>
<td>x</td>
<td>xx</td>
</tr>
<tr>
<td>Distressing dreams—flashbacks</td>
<td>xx</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>xx</td>
<td>N/A</td>
</tr>
<tr>
<td>Distressing dreams—sleep disturbance</td>
<td>x</td>
<td>x</td>
<td>xx</td>
<td>xx</td>
<td>x</td>
<td>-</td>
<td>xx</td>
<td>N/A</td>
</tr>
<tr>
<td>Flashbacks—sleep disturbance</td>
<td>x c</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>xx d</td>
<td>N/A</td>
<td>xx</td>
</tr>
<tr>
<td>Avoidance of thoughts—avoidance of external reminders</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>xx</td>
<td>x</td>
<td>xx</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Diminished interest—detachment</td>
<td>x</td>
<td>x</td>
<td>xx</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>N/A</td>
</tr>
<tr>
<td>Diminished interest—problems with concentration</td>
<td>x</td>
<td>x</td>
<td>xx</td>
<td>x</td>
<td>xx</td>
<td>x</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Detachment—restricted affect</td>
<td>xx</td>
<td>xx</td>
<td>x</td>
<td>xx</td>
<td>xx</td>
<td>x</td>
<td>x</td>
<td>N/A</td>
</tr>
<tr>
<td>Detachment—irritable or angry</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>xx</td>
<td>x</td>
<td>-</td>
<td>xx</td>
</tr>
<tr>
<td>Irritable or angry—problems with concentration</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td>xx</td>
<td>x</td>
<td>-</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Irritable or angry—sleep disturbance</td>
<td>-</td>
<td>x</td>
<td>-</td>
<td>xx</td>
<td>x</td>
<td>-</td>
<td>xx</td>
<td>N/A</td>
</tr>
<tr>
<td>Hypervigilance—startle</td>
<td>xx</td>
<td>xx</td>
<td>xx</td>
<td>xx</td>
<td>xx</td>
<td>xx</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Note. xx = emphasised as strong and reliable; x = estimated in the network, but not emphasised as strong and reliable; - = not estimated in the network; N/A = at least one of the symptoms in the edge was not included in the network.

a These results are based on the cross-sample network.

b These results are based on Table 2 from Sullivan et al.

c This edge overlaps with others in Figure 1 of Armour et al., so is not visible, but is reported in the supplementary materials.

d Spiller et al. mistakenly reported this edge as intrusive thoughts—sleep disturbance in-text, but examination of the network and supplementary materials made it clear the estimated edge they were referring to was flashbacks—sleep disturbance.
<table>
<thead>
<tr>
<th>Point in commentary in the order they are presented in-text</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>There are a number of small variations in the re-analyses presented in Borsboom et al.’s Table 1 and supplementary materials compared to our results in Tables 2-4 using the same metrics.</td>
<td>These differences are almost invariably due to three minor methodological differences: (1) Borsboom et al. allowed nodes to have multiple simultaneous ranks in computing matches in centrality rank-order while we did not (i.e., we selected a single rank order that had the maximum possible matches); (2) different rounding rules; and (3) three pairs of random split-halves in the NCS-R data were inverted (i.e., baseline versus replication) in our Ising model analyses contrary to the labelling convention on the files MKF sent Borsboom et al. for analysis. In addition, there appear to be two computation or transcription errors on Borsboom et al.’s part and one on our part. However, none of the differences in the tables affect the interpretation of the results, and given the Tables 2-4 of the target article summarise 1,397 estimated parameters, the overwhelming pattern is in the high concordance of the two sets of analyses. The code that Borsboom et al. generated to reproduce our results (<a href="https://osf.io/akywf/">https://osf.io/akywf/</a>) represents an excellent resource for other researchers interested in studying these methods.</td>
</tr>
<tr>
<td>Borsboom et al. say “As Forbes et al. note themselves… estimated Ising networks are nearly identical” (p. 8).</td>
<td>We did not note this, and in fact conclude largely the opposite. This may be based on a misinterpretation of the sentence “The replicability of the edges in the Ising models was remarkably similar in the between and within samples comparisons (see Tables 2–4).” (p. 7)—that is, the results from comparing the Ising models based on the full NCS-R and NSMHWB samples were similar to the results from comparing the split-halves within each sample (rather than any of the results showing high levels of replicability).</td>
</tr>
<tr>
<td>Regarding cross-method replicability, Borsboom et al state “It is clear from the way they interpret the resulting findings that they assume that one should expect these different networks to converge to 100%” (p. 10).</td>
<td>This is not accurate. As noted in footnote 2 of our reply, the different conditional independence networks are purported to identify the same inferential targets (i.e., dynamic causal relationships among symptoms). It is thus an important research question for the field to examine whether these methods converge on the same fundamental results. Our analyses of cross-method replicability in the target article were aimed at testing this research question, not based on an assumption of convergence.</td>
</tr>
<tr>
<td>Borsboom et al. imply that we included association networks in our assessment of cross-method replicability: “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.” (p. 11).</td>
<td>We did not include association networks in our cross-method replicability, as they have rarely been used in the network literature to make causal inferences about symptom-to-symptom relationships.</td>
</tr>
<tr>
<td>Borsboom et al. point out a typo in the number of replicated edges we report between Ising models and DAGs in footnote 7 of the target article, leaving them “no other option than to conclude that none of the statistics on cross-method replicability reported by Forbes et al. are accurate” (p. 18)</td>
<td>This inference is clearly at odds with the results in their online materials that find the same proportion (%) of replicated edges that we report for all of the Ising and DAG model comparisons.</td>
</tr>
</tbody>
</table>
Borsboom et al. suggest that the % change in replicated edges metric is too sensitive, particularly for small parameters, and suggest that this metric “is unfit to serve as an arbiter of replicability.” (p. 20)

Borsboom et al. suggest that our use of matches in node centrality rank-order metric is too sensitive because “nodes can shift positions in the rank ordering due to sampling fluctuations” (p. 20) and give the example of node rank order A, B,…Z in network 1 versus Z, A, B, C,…,Y in network 2 as evidence for the undue sensitivity of the metric (p. 21). They also emphasise the importance of interpreting this metric in the context of the sampling variability of the centrality estimates as well as how reliably the network structure is estimated (p. 20-22).

The “So what about measurement error?” section (p. 22-24) is largely based on the straw man that we hypothesize a sparse network.

Later in the “So what about measurement error?” section, Borsboom et al., go on to say that the mathematical equivalence between specific network and latent variable models means that “if network structures replicated badly across two datasets, then this would imply that factor structures (i.e., the configuration of loadings in exploratory factor models) would replicate badly as well.” (p. 24).

Borsboom et al. suggest “best practices” for future research that include:

1. Using bootnet to ensure the reliability and replicability of networks (p. 25); and
2. The need to preregister replication research but not standard psychopathology network research (p. 29). They also imply that we selected measures post hoc to emphasise evidence against the replicability of networks.

We agree that this is a sensitive metric of change that may be biased by changes in small parameters. (Notably, we also examined change in individual edges between networks using a second less sensitive measure as well—the proportion of edges that replicated.) However, all the proposed alternative metrics in Borsboom et al. evidently err on the side of not being sensitive enough to capture differences between networks (see discussion in-text), and we would suggest that the % change in edges metric may thus be a useful complement to the blunter metrics, such as correlating edge lists.

This metric is simply an extrapolation of testing whether the “most central node” within each centrality index is the same in two networks (i.e., is the second most central node the same; the third… etc.), to reflect the way node centrality rank-order is interpreted in the literature. We would suggest that when the least “influential” node ’Z’ in one network is the most influential in another, this highlights fundamental unreliability in the conclusions regarding node centrality. Importantly, as Borsboom et al. illustrate, this substantial difference between network 1 and network 2 would be missed using their proposed metric of correlating the centrality indices between the two networks. Further, the results for the Ising models highlight that nodes do indeed shift due to random sampling fluctuations, suggesting there is little reliable information in the rank-orders: The bootnet results suggest that the networks are estimated reliably and the strength centrality estimates are highly stable, but only 16.7% of the nodes had the same strength centrality rank-orders in NCS-R and NSMHWB.

We do not hypothesise a sparse network, but rather that much of the evident sensitivity of the conditional independence networks is based on the predominance of measurement error in the estimated edges (i.e., accounting for the changeability in the presence and strength of edges within and between the data sets). This is consistent with the evidence in our target article, Steinley et al.’s (this issue) analyses of the data, as well as the evidence in the literature more broadly (see discussion in-text).

Like the metrics Borsboom et al. use to re-analyse replicability in their commentary, this inference fails to consider the level on which replicability would be meaningfully assessed in the two models: Latent variables capture the (stable) shared variance among the indicators; networks interpret the presence, strength, and direction of individual conditionally independent relationships as meaningful. Even though both are based on the same underlying covariance matrices, replicability takes very different forms in these contexts. As Steinley et al. put it, “the bar is much higher” (p. 7) for replicability in network models.

(1) Steinley et al.’s work unequivocally shows that establishing the stability of parameter estimates alone does little to ensure that network results are meaningful and/or replicable.

(2) The suggestion that replication-focused research specifically should be preregistered so “other researchers” (p. 29)—presumably those already working in the field of interest—can check and approve the hypotheses and analysis plan is problematic for a few reasons. For example, Coyne (2016) highlighted that this sort of requirement for replication studies inadvertently reinforces the problems in the literature that replication aims to address, and holds replication studies to a higher standard than the original research that may well be generating unreplicable results. Ioannidis (2012) also discusses the perils of fostering obliged replication when proponents of a theory can largely select and mold the results and interpretation of studies. We aimed for independent replication, and—as discussed in-text—designed our metrics specifically for assessing replicability based on how psychopathology networks are interpreted in the literature.
References