Quantifying and addressing the impact of measurement error in network models

Ron, J. de; Robinaugh, D.J.; Fried, E.I.; Pedrelli, P.; Jain, F.A.; Mischoulon, D.; Epskamp, S.

Behaviour Research and Therapy 157 (2022) 104163

Contents lists available at ScienceDirect

Behaviour Research and Therapy

journal homepage: www.elsevier.com/locate/brat

Quantifying and addressing the impact of measurement error in network models☆

Jill de Ron [a,*], Donald J. Robinaugh [b,c], Eiko I. Fried [d], Paola Pedrelli [b], Felipe A. Jain [b], David Mischoulon [b], Sacha Epskamp [a,e]

[a] Department of Psychological Methods, University of Amsterdam, the Netherlands
[b] Department of Psychiatry, Massachusetts General Hospital & Harvard Medical School, USA
[c] Department of Applied Psychology, Northeastern University, USA
[d] Department of Clinical Psychology, Leiden University, the Netherlands
[e] Centre for Urban Mental Health, University of Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Measurement error
Replicability
Single-item indicators
Multiple-item indicators
Latent network modeling

ABSTRACT

Network psychometric models are often estimated using a single indicator for each node in the network, thus failing to consider potential measurement error. In this study, we investigate the impact of measurement error on cross-sectional network models. First, we conduct a simulation study to evaluate the performance of models based on single indicators as well as models that utilize information from multiple indicators per node, including average scores, factor scores, and latent variables. Our results demonstrate that measurement error impairs the reliability and performance of network models, especially when using single indicators. The reliability and performance of network models improves substantially with increasing sample size and when using methods that combine information from multiple indicators per node. Second, we use empirical data from the STAR*D trial (n = 3,731) to further evaluate the impact of measurement error. In the STAR*D trial, depression symptoms were assessed via three questionnaires, providing multiple indicators per symptom. Consistent with our simulation results, we find that when using sub-samples of this dataset, the discrepancy between the three single-indicator networks (one network per questionnaire) diminishes with increasing sample size. Together, our simulated and empirical findings provide evidence that measurement error can hinder network estimation when working with smaller samples and offers guidance on methods to mitigate measurement error.

1. Introduction

In recent years, there has been an immense increase in empirical studies utilizing network psychometrics (Robinaugh, Hoekstra, Toner, & Borsboom, 2020). Network psychometrics is a collective term for statistical models that estimate a network structure on psychological data (Epskamp, Maris, Waldorp, & Borsboom, 2017). Despite its popularity, the replicability of network models—that is, the extent to which network properties generalize across samples and measurement scales—is still under scrutiny (Fried, 2017, Forbes, Wright, Markon, & Krueger, 2017; Borsboom et al., 2017, Funkhouser et al., 2020; Jones, Williams, & McNally, 2021). The debate about network replicability prompted the development of methods that assess the stability of networks, such as bootstrapping (Epskamp, Fried, & Borsboom, 2018) and permutation tests (Van Borkulo et al., 2017). These metrics investigate parameter precision via resampling, showing that sampling variability is one main source diminishing network replicability. Additionally, recent work by Herrera-Bennett and Rhemtulla (2021) found that variability in measurement scales is a source of inconsistencies between networks, underscoring the importance of measurement in network replicability. In the current paper, we investigate the impact of measurement error on cross-sectional network models.

Many network studies investigate relationships among variables measured by a single item, typically by including each item of a

☆ National Institute of Mental Health Career Development Award1K23MH113805-01A1the NWO Veni Grant016-195-261National Institutes of HealthK76AG064390

* Corresponding author. Department of Psychology, Psychological Methods, University of Amsterdam, Nieuwe Achtergracht, 129-B, 1018 WT Amsterdam, the Netherlands.
E-mail address: j.deron@uva.nl (J. de Ron).

https://doi.org/10.1016/j.brat.2022.104163
Received 4 March 2021; Received in revised form 30 June 2022; Accepted 12 July 2022
Available online 3 August 2022

questionnaire as its own node in the network. In such single-indicator networks, variables are treated as *observed* and measurement error is not taken into account. However, measurement error is typically assumed to be present in psychological data, given that the variables of interest in psychological research are often not directly observable (Flake & Fried, 2020; Schmidt & Hunter, 1999; Schuurman & Hamaker, 2019). This may seem obvious for complex, multi-dimensional constructs such as depression, but even with relatively simple constructs, such as the symptom depressed mood, the observed behavior (e.g., the response to an item on a questionnaire) is unlikely to be solely determined by the variable of interest. The failure to account for this measurement error may diminish the precision of network estimation.

Measurement error is not a problem unique to psychological networks: It has long been known that ignoring measurement error can bias estimated relationships between variables (e.g., Cole & Preacher, 2014; Jaccard & Wan, 1995). For instance, measurement error attenuates the effects of non-zero partial correlations (i.e., the correlation between two variables after controlling for other variables; Liu, 1988). In contrast, when the true partial correlation is zero, measurement error makes it impossible to perfectly condition on variable(s) that would make the variables of interest conditionally independent, leading to spurious partial correlations (Liu, 1988). As most cross-sectional network studies use undirected, weighted networks, where edge weights represent partial correlations (Robinaugh et al., 2020) measurement error may, thus, be expected to lead to spurious edges while attenuating the edge weights of true edges.

Anticipating these potential problems posed by measurement error, researchers have frequently recommended the use of multiple indicators per node as a means of improving measurement precision (e.g., self- and peer-report; Epskamp, Rhemtulla, & Borsboom, 2017; Bringmann & Eronen, 2018; Fried & Cramer, 2017; Herrera-Bennett & Rhemtulla, 2021). However, there is a lack of clarity around how to include multiple indicators per node: with some researchers calling for the inclusion of latent variables in the network, while others recommend using averages or sum scores. Such choices are not trivial as they can impact someone's score on the variable of interest and, subsequently, the resulting network structure (e.g., Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016; McNeish & Wolf, 2020).

In this project, we respond to the calls for improving network reliability by examining (1) the extent to which measurement error impairs cross-sectional network estimation and the conditions under which such impairment may occur, and (2) the extent to which different multiple-indicator methods can mitigate its potential effects.[1] We do so in two studies. In Study 1, a simulation study, we evaluate the impact of measurement error on reliable estimation by assessing agreement in network structure 1) among three single-indicator networks (Model 1; M1a, M1b, M1c, based on different questionnaires) and 2) among three methods that address measurement error by incorporating information from multiple indicators per node: models based on average scores (M2), models based on factor scores (M3), and latent network models (M4). We then evaluate how well each of these models (M1-M4) recover the true underlying network. For both aims, we repeat the analyses across varying conditions of measurement error and sample size. In Study 2, an empirical study, we evaluate each of these models using data from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial (Fava et al., 2003; Rush et al., 2004). The STAR*D study assessed the nine DSM-5 symptoms of Major Depressive Disorder (MDD) with three different questionnaires in a large and well-characterized sample as part of the study's baseline assessment, thereby providing three indicators per node. Our empirical analyses of this dataset mirrors the analyses performed in our simulation study, investigating the alignment among three single-indicator networks (M1a, M1b, M1c) and among multiple-indicator networks (M2-M4) across varying sample sizes.

[1] Studies on time-series data (e.g., vector autoregressive models, VAR) show that unaccounted measurement error leads to attenuated autoregressive effects (Schuurman, & Hamaker, 2019; Schuurman, Houtveen, & Hamaker, 2015; Staudenmayer & Buonaccorsi, 2005). Although it is possible to estimate latent network models from time-series data (Epskamp, 2020), in the context of time-series studies, researchers are more likely to rely on single-item measurements, given the need for brief questionnaires when assessments are repeatedly administered over relatively brief time intervals.

2. Study 1: simulation

In empirical data, both the underlying data-generating structure and the amount of measurement error are unknown, making it difficult to determine the impact of measurement error on a method's ability to recover the data-generating structure. By contrast, in a simulation study, the true underlying network is known, and the precise amount of measurement error can be specified. Accordingly, the impact of different amounts of measurement error on the ability to recover the true network can be precisely determined. In this simulation study, we investigate the effect of random measurement error in networks using nodes assessed by a single indicator as well as networks based on methods that combine multiple indicators per node.

There are several potential methods by which one could use multiple indicators to overcome measurement error, each with its strengths and weaknesses. One approach is to take the average score of the indicators (M2). Averaging indicators is often used as a proxy for a latent variable (Rhemtulla, van Bork, & Borsboom, 2020; McNeish & Wolf, 2020; e.g., Briganti, Fried, & Linkowski, 2019). This method has the advantage of being computationally simple. However, because measurement error is not explicitly modelled, the variance is inflated as it contains both true construct variance and measurement error variance (Cole & Preacher, 2014; Ledgerwood & Shrout, 2011). A second approach (M3) is to use factor scores. In contrast to averaging, factor scores explicitly account for measurement error by extracting the shared variance among the indicators, which is assumed to reflect the latent construct, and the unique variance of each indicator, which is seen as measurement error. However, using factor scores as observed variables in a subsequent analysis can lead to an inflated covariance matrix due to factor indeterminacy (i.e., the indicators of a factor have infinite ways to satisfy the same factor model; Acito & Anderson, 1986).

Recently, Epskamp and colleagues developed a third approach to address measurement error in the context of network analysis (M4). This method, known as latent network modeling (LNM, Epskamp, 2020; Epskamp, Maris, et al., 2017), avoids factor indeterminacy by directly estimating a network on the implied latent variance-covariance structure of the data. Simulation research on latent network modeling shows good recovery of the parameter estimates, but it is computationally intensive, suggesting it may not be suitable for all substantive areas of interest (e.g., for networks with many nodes; Epskamp, Maris, et al., 2017). In addition, it has never been compared with other methods combining multiple indicators (e.g., M2 and M3). Given the absence of direct comparison of these three methods and each approach's relative strengths and weaknesses, we investigated all three methods in this study.

2.1. Methods

2.1.1. Simulation overview

Each simulation completed for this study consisted of three steps. In Step 1, we generated a true network and simulated data from that network under varying sample size and measurement error conditions. For simplicity and computational power, we used a chain network (i.e., each node connects to its two neighboring nodes) as a true network (see Fig. 1). To mirror the empirical data in our second study as much as possible, the generated number of indicators per node was three. The average edge weight was 0.33. In Step 2, we estimated six network models on the generated data: three *single-item networks* (M1a, M1b,

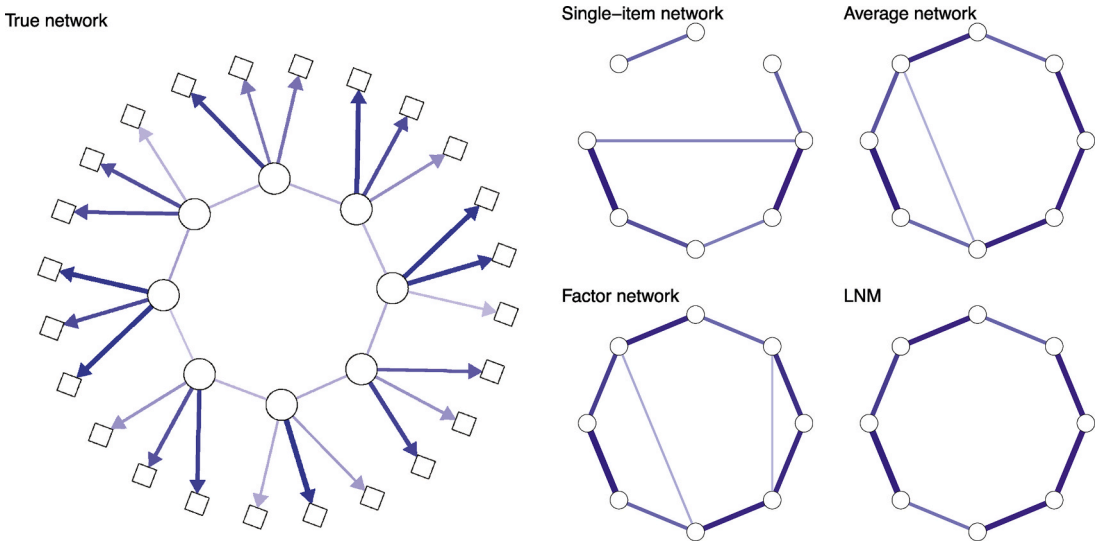

Fig. 1. An example of the true and estimated networks. The left panel depicts one of the true network structures we generated, in which each latent variable (i.e., circular nodes) is assessed by three indicators (i.e., square nodes) plus random measurement error; factor loadings are represented by directed edges (i.e., those with an arrow). The true network is a chain graph, where all latent variables are related to two others, and we set all of these relations variables to be equal. Undirected edges (i.e., those without an arrow) represent partial correlations. Blue edges represent positives relations. Greater saturation of the edge indicates a stronger relationship. The right panel depicts examples of estimated networks on data simulated from the true network with a sample size of 1000, and a medium level of measurement error. Note that the methods differ in the extent to which they have recovered the true network of latent variables. The single-item network exhibits the lowest sensitivity, failing to identify two edges present in the true network, as well as identifying one spurious edge. Conversely, the average and factor network exhibit perfect sensitivity, but also include spurious edges not present in the true network, and therefore have lower precision and specificity. The LNM network performs best, recovering all edges in the true network and not exhibiting any spurious edges, although not all edges are recovered with equal edge weights. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

M1c) estimated based on a single indicator for each node; an *average network* (M2) based on the average of each node's indicators; a *factor score network* (M3) based on extracted factor scores for each node; and a *latent network model* (M4) based on latent variables. In Step 3, we evaluated how much the single- and the multiple-indicators networks differed from each other. Furthermore, we evaluated how well the models (M1a-M4) recovered the true network (see Fig. 1). We completed these simulations under varying conditions of measurement error (with a maximum value of 0, 0.5, 1, 1.5 and 2) and sample size (250, 500, 1,000, 2,500, and 5,000). Based on prior simulation work, we consider samples of n = 250 'small' and samples of n = 5000 'large' (Epskamp, Borsboom, & Fried, 2018). We note, however, that there is no formal definition of what sample size is considered small or large, as the power to detect edge weights reliably depends on many characteristics, including true strength of the edge weights, the variance in the items, and the number of nodes in the network. We performed 100 simulations for every combination of these conditions, resulting in 2,500 (5 x 5 x 100) simulations. We review each step of the simulation process in further detail in the next section.

2.1.2. Network generation, estimation, and evaluation

We conducted all data analyses in the statistical program R (Team & R Development Core Team, 2016). The code can be found on https://osf.io/tfwmh/.

In Step 1, we first generated the latent network via the genGMM function from the *bootnet* R-package (Epskamp, Fried & Borsboom, 2018). This latent network served as the true network that our analyses aimed to recover. We next generated three indicators from every node (i.e., latent variable) in the true network, where indicators were a function of each node plus measurement error. Data generation was the same as described in Epskamp, Rhemtulla, and Borsboom (2017), except that we added independent measurement error. To generate measurement error that varies across items, we sampled from a uniform distribution from 0.25 to 1, either multiplied by 0, 0.5, 1, 1.5 and 2 depending on the measurement error condition. Accordingly, our labels for the different conditions indicate the maximum possible measurement error in that condition.

In Step 2, we estimated the four different network models on the generated data frame. For the single-item networks, we selected one indicator per latent variable and estimated a network structure. For the factor score networks, we used the cfa function from the *lavaan* package (Rosseel, 2012) and extracted the estimated factor scores with the Bartlett predictor of the lavPredict function from the *lavaan* package (Rosseel, 2012). On the resulting factor scores, we estimated a network model. We estimated the networks M1 to M3 using the ggm function and for M4, the LNM, we used the lnm function; both functions are from the *psychonetrics* package (Epskamp, 2020). We used the prune and modelsearch function from the *psychonetrics* package (Epskamp, 2020) on all estimated networks, M1 through M4. Prune removes non-significant edges, and model search performs a stepwise model search by minimizing the Bayesian Information Criterion (BIC).

In Step 3, we first investigated the similarity among different single-indicator networks by computing the mean correlation between the edge weights of the three networks defined by different single indicators. These analyses indicate how stable the findings are across different single-indicator networks under varying measurement error and sample size conditions. We performed this same analysis for the three multiple-indicator networks, indicating how consistent the findings are across different methods of combining information across multiple indicators. Furthermore, we evaluated the performance of each estimated network, M1-M4, in recovering the true network by examining (a) the *correlation* between edge weights of the estimated network and the true network, (b) the *precision* of the estimated network (i.e., the proportion of correctly identified edges relative to all estimated edges), (c) the *sensitivity* or true-positive rate (i.e., the proportion of edges in the estimated network that were also in the true network), and (d) the *specificity* or true-negative rate (i.e., the proportion of missing edges in the estimated

network that were also missing in the true network).[2] Higher scores on all metrics indicate better recovery of the true network, with a 1 indicating complete recovery.

2.2. Results study 1

Fig. 2 shows the results of the correlation between the edge weights of the three single-indicator networks (examining the similarity between networks based on different indicators) and among the three multiple-indicator networks (examining the similarity between networks using the same indicators but different methods for combining indicators). In both sets of analyses, the correlation among edge weights of the networks worsens with increasing measurement error. This decrease is substantially less severe in large samples relative to small samples. For single-indicator networks, the correlations between edge weights are weak when measurement error is high and sample size is small, indicating that the networks differ substantially from each other when a significant amount of measurement error is present and samples are small. For the multiple-indicator networks, the correlation between edge weights is quite high, even in the context of high measurement error and relatively small samples, suggesting consistently strong agreement among the three different methods of combining information across multiple indicators (i.e., average scores, factor scores, and latent variables).

The results of our simulation investigating recovery performance appear in Fig. 3. For single-indicator networks, when measurement error is zero, the estimated networks exhibit near-perfect recovery of the true network, even in small samples. However, as measurement error increases, the performance of the single-indicator networks suffers, especially when sample sizes are small: the correlation between edge weights of the true network and edge weights of the estimated network worsens and there is a large decrease in sensitivity. That is, measurement error reduces the number of accurately detected edges in the estimated network. As measurement error decreases the power to detect edges, specificity and precision are less affected by measurement error than sensitivity. For large sample sizes, due to an increase in power, the impact of measurement error on sensitivity and correlation is significantly lessened. Whereas precision and specificity worsen with sample size, indicating an increase in false positives (i.e., edges absent from the true network but present in the estimated network).

For all multiple-indicator networks, the decrease in performance with growing measurement error was substantially less severe. For small sample sizes with measurement error, multiple-indicator networks significantly outperform single-indicator networks. In situations with high measurement error and large sample sizes, single-indicator networks perform slightly better than the average and factor score networks on specificity and precision. Due to very high unaccounted measurement error, single-indicator networks pick up fewer edges (both true and false ones) than the average and factor score networks. Accordingly, with fewer estimated edges, single-item networks exhibit higher specificity and precision, but at the cost of lower sensitivity. Consistent with the analyses presented in Fig. 2, there was little difference between the three multiple-indicator networks in most conditions, especially when the sample size was small. However, LNM outperformed average scores and factor scores when the sample size was large. The LNM is the only network model that converged to the true model with increasing sample size by modeling measurement error in a way that aligns with the data generating model.

[2] We did not conduct statistical analyses on the outcome metrics because with simulation studies, sample sizes can be increased, andy that tiny deviations from the null-hypothesis could lead to its (false) rejection (Cohen, 1995).

2.3. Conclusion study 1

In our simulation study, we find that measurement error is indeed a problem for single-indicator networks, which differ from one another considerably when significant measurement error is present, especially in small samples. Furthermore, in the context of high measurement error and small samples, single-indicator networks show poor sensitivity and poor agreement between the estimated network and the true network. It is noteworthy that the single-indicator networks do become more consistent with one another and better recover the true data-generating network in larger samples, even in the context of high measurement error. Nonetheless, multiple-indicator networks consistently and substantially outperformed single-indicator networks, exhibiting good recovery performance even in the context of moderate sample sizes and high measurement error. Although all methods of combining multiple indicators performed comparably well across most conditions, LNM outperformed other methods with large samples due to a drop in specificity for all other multiple-indicator methods.

3. Study 2: empirical analysis

Although a simulation study gives us the advantages of working from a known true network and of being able to manipulate factors of interest in isolation, it is also essential to investigate the effect of single- and multiple-indicators in empirical data, which may have characteristics not accounted for in the simulation. Thus, parallel to the simulation study, we investigate the similarity of three different single-indicator networks (M1a, M1b, M1c) and three multiple-indicator networks (M2-M4) in the STAR*D data. For this empirical analysis, we address the challenge of not knowing the true network by repeating the analysis from our simulation study evaluating the agreement among three single-indicator networks and among three different multiple-indicator networks. Based on Study 1, we expect in Study 2 that if measurement error is impairing performance, then the correlation among different single-indicator networks within the same sample should be low when the sample size is small, and should increase as the sample size grows. Our simulation study also suggests that if multiple indicators are improving network reliability, then the multiple-indicator networks should show agreement with each other, even in small samples.

3.1. Methods

3.1.1. Data

The STAR*D trial (Fava et al., 2003; Rush et al., 2004) investigated the effectiveness of different depression treatments in a large sample of participants diagnosed with Major Depressive Disorder (MDD). The data from this trial are uniquely well suited for our purposes here because they include multiple assessments of depression symptoms via three versions of the Inventory of Depressive Symptomatology (IDS; Rush et al., 1986): the full 30-item clinician rated version (IDS-C), a self-rated 16-item version of the scale (QIDS-S) and clinician-rated versions of that same 16-item scale (QIDS-C). For each questionnaire, each item yields a score between 0 and 3, with 0 indicating no symptoms and 3 indicating severe symptoms. All questionnaires were administered to the participants pre-treatment, but QIDS-C was administered at the intake and IDS-C and QIDS-S were administered at the following assessment. Reliability estimates lay between 0.76 and 0.82 in a depressed subpopulation and between 0.92 and 0.94 in a sample with a combination of healthy and depressed participants (Rush, Gullion, Basco, Jarrett, & Trivedi, 1996).

We selected items from each questionnaire assessing the nine symptoms of the DSM: (1) depressed mood, (2) loss of interest or pleasure, (3) decrease or increase in appetite, (4) insomnia or hypersomnia, (5) psychomotor agitation or retardation, (6) diminished energy, (7) feelings of worthlessness or guilt; (8) diminished ability to think or concentrate and (9) suicidal ideation. Symptoms 3, 4 and 5 reflect

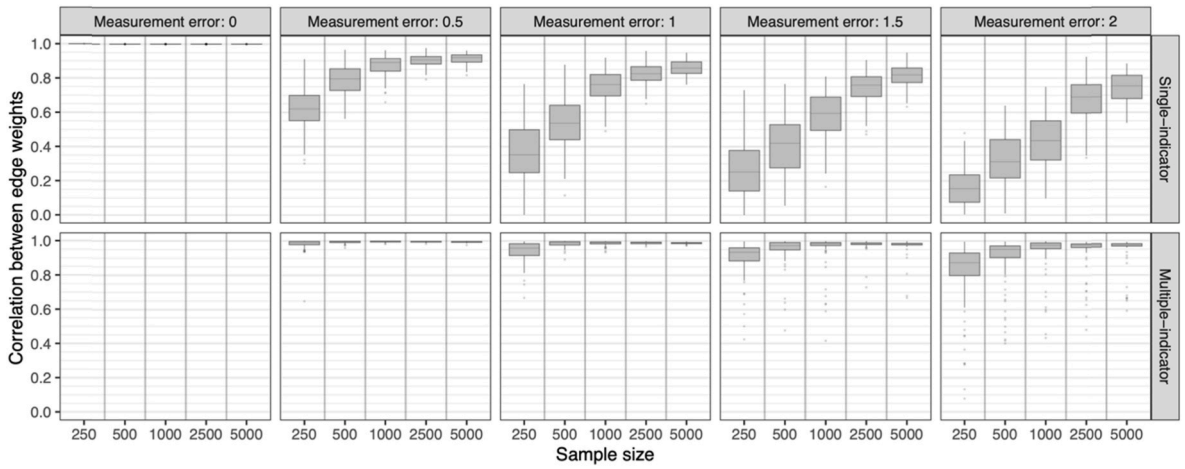

Fig. 2. Simulation results for the similarity between three single-indicator networks and the three multiple-indicator networks under varying measurement error and sample size conditions. The simulation consists of two sets of analyses. The top panel depicts correlations between the edge weights of the three different single-indicator networks (M1a, M1b, M1c), where the indicators are different across networks. The bottom panel depicts the correlation between the three different multiple-indicator networks (M2-M4), where the indicators are the same but the method of combining information from multiple indicators is different. The vertical panels indicate the maximum amount of measurement error. Every condition was simulated 100 times and the boxplots represent the distribution of the mean correlation between the edge weights (i.e., 25th quartile, median, 75th quartile). Importantly, these analyses do not permit direct comparison between the single-indicator networks, which examined the similarity between networks based on different indicators, and multiple-indicator networks, which examine the similarity between networks using the same indicators but different methods for combining those indicators.

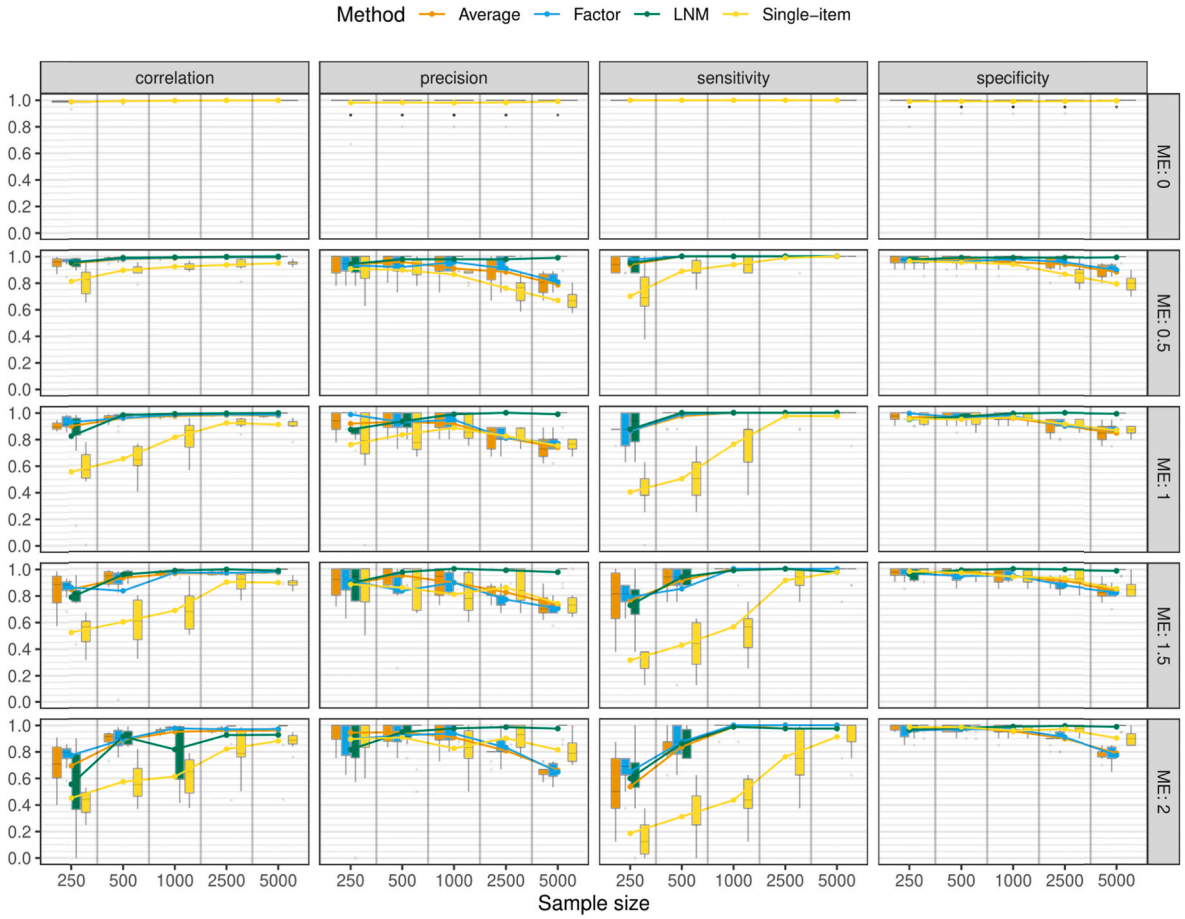

Fig. 3. Simulation results for recovery performance for M1 to M4. The vertical panels indicate the different measures: correlation, precision, sensitivity, and specificity. Horizontal panels indicate the amount of measurement error, abbreviated as ME. Every condition was simulated 100 times and the boxplots represent the distribution of those measures (i.e., 25th quartile, median, 75th quartile); the dots represent the mean of those measures connected by a line to indicate the trend. In conditions with no measurement error, some simulation runs for the factor score networks and the LNM were not completed as the covariance matrix was not positive definitive; this is because some items were a linear combination of other items, leading to eigenvalues of zero.

opposing conditions (e.g., decrease vs. increase in appetite), and all three of the IDS questionnaires (i.e., IDS-C, QIDS-S, QIDS-C) included multiple items to assess the distinct manifestations of these symptoms. For these symptoms, we divided the symptom into two nodes each (e.g., separating the symptom sleep problems into the node *hypersomnia* and the node *insomnia*), resulting in a network of 12 nodes. To investigate if our results are robust against this analytic choice, we also ran the analysis with the 'compound' symptoms that incorporate opposing conditions in a single node in the network, resulting in a network representing the nine DSM symptoms (see Supplementary materials).

We are aware of three studies that use (part of) the data from the STAR*D study to estimate network models. Madhoo and Levine (2016) estimated a network of 14 from the 16 QIDS-S items administered at the beginning of the first treatment stage. Fried, Epskamp, Nesse, Tuerlinckx, and Borsboom (2016) estimated two separate networks based on 28 out of 30 items concerning depression symptoms from the IDS-C questionnaire: They estimated one network on 15 items that are part of the DSM-5 criteria for MDD, and one network on the 13 other items from the IDS-C questionnaire that are non-DSM symptoms. de Ron, Fried, and Epskamp (2021) estimated a network on the 17-item Hamilton Rating Scale for Depression (HRSD) assessed at the beginning of the first treatment stage. In contrast to the current study, all previous studies used each questionnaire item as separate nodes in the network.

3.1.2. Participants

Participants of the original STAR*D study (Fava et al., 2003; Rush et al., 2004) were recruited through mental health and medical care practices and needed to meet to the DSM-IV criteria for single episode or recurrent MDD and have at least 14 points or higher on the Hamilton Rating Scale for Depression (HRSD; Hamilton, 1960). The age of the participants ranged from 18 to 75 years. From the 4041 participants who started the STAR*D study, we examined a subset of participants (n = 3731) from whom data were available on all three questionnaires.

3.1.3. Data analysis

We conducted all analyses in the software program R (Team & R Development Core Team, 2016). We estimated and visualized the same network models used in our simulation study (M1-M4) on the full STAR*D sample. For our single-item networks, we estimated separate symptoms networks for the three (Q)IDS measurement scales (M1a, M1b, M1c). For M2-M4, we combined each of the three items assessing each symptom using average scores (M2), factor scores (M3) or a LNM (M4). Furthermore, we conducted empirical analyses parallel to our simulation study. We estimated all network methods (M1-M4) across varying conditions of sample size, by either using the full sample (N = 3, 731) or randomly selected subsets of the STAR*D data with 250, 500, 1000, 2500 observations. From these networks we computed the mean correlation between edge weights of the different single- and multiple-indicator networks.

3.2. Results study 2

Fig. 4 displays all network models on the complete STAR*D dataset; Table 1 provides an overview of the characteristics of those networks. In line with what we would expect from our simulation study given the high sample size, the overall network structure is similar across different network models: The network density and average edge weights are relatively similar and salient features of the network are consistent across analyses. For example, the association between low energy (ENGY) and lack of interest (INTR) is consistently the strongest or

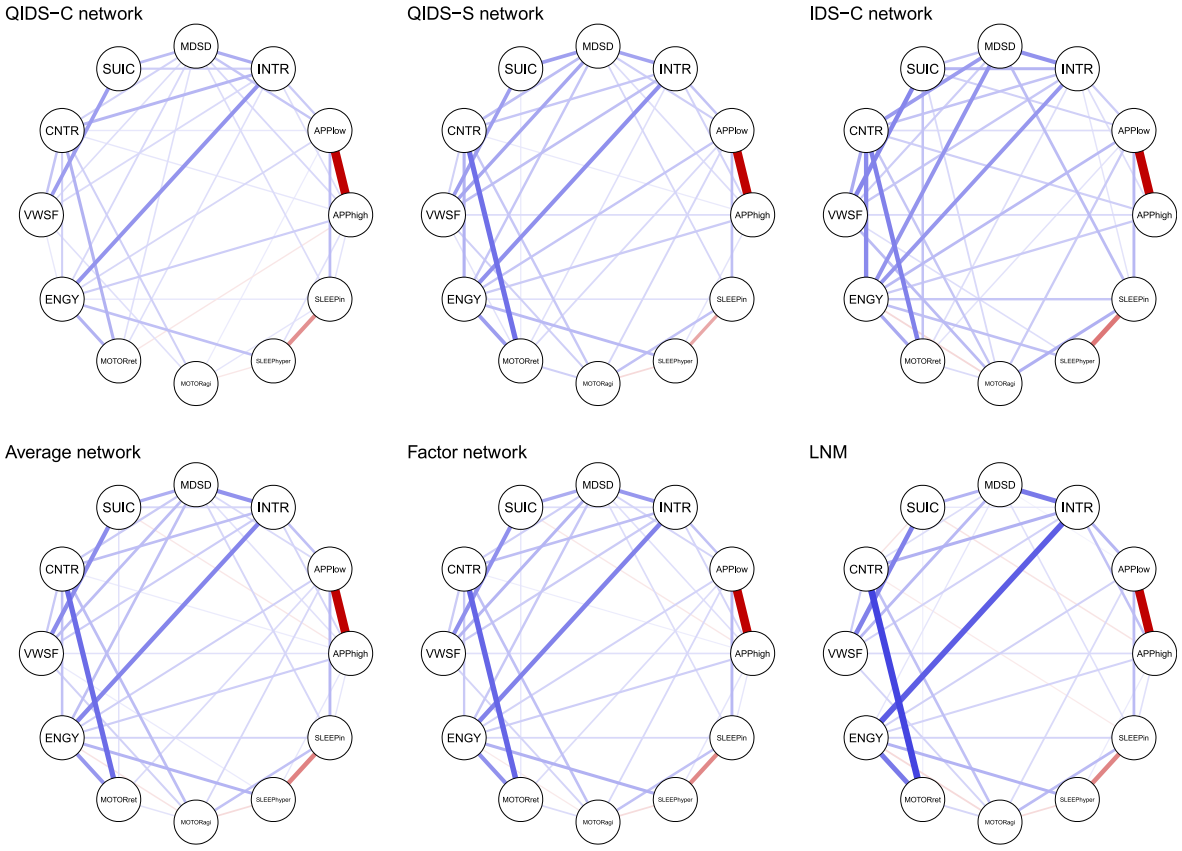

Fig. 4. Six possible network models to infer relations between 12 symptoms of Major Depression on the complete STAR*D data (n = 3,737). DMSD = depressed or sad mood, INTR = loss of interest, APPlow = decrease in appetite, APPhigh = increase in appetite, SLEEPin = insomnia, SLEEPhyper = hypersomnia, MOTORagi = psychomotor agitation, MOTORret = psychomotor retardation, ENGY = diminished energy, VWSF = view of self (feelings of worthlessness or guilt), CNTR = diminished ability to think or concentrate, and SUIC = suicidal ideation.

Table 1
Overview of the network characteristics of the six network models.

	QIDS-C	QIDS-S	IDS-C	Average	Factor	LNM
Number of Edges	39	38	40	41	40	38
Network Density	0.59	0.58	0.61	0.62	0.61	0.58
Edge Weight Mean (SD)	0.06 (0.12)	0.08 (0.12)	0.08 (0.11)	0.08 (0.14)	0.08 (0.14)	0.09 (0.18)

among the strongest associations across each network. Similarly, diminished ability to think or concentrate (CNTR) is consistently closely associated with psychomotor retardation (MOTORret). Furthermore, suicidal thoughts (SUIC) are strongly associated with view of self (feelings of worthlessness or guilt; VWSF). As expected, all the networks display a negative relationship between hyper- and insomnia and between increase and a decrease in appetite.

Fig. 5 shows the correlation between edge weights among the three single-indicator networks (i.e., the QIDS-C, QIDS-S, and IDS-C networks) and among the three multiple-indicator networks (i.e., average, factor score, and latent variable networks) estimated with varying sample sizes of the STAR*D data. The correlation between the edge weights shows the same pattern as the simulation study (see Fig. 2). For single-indicator networks, the correlation among networks is the lowest at small samples, suggesting unstable results. As the sample size increases, the agreement among these networks grows substantially. For multiple-indicator networks, the three methods that combine multiple indicators show a higher correlation between edge weights, suggesting stronger agreement among different methods for combining multiple indicators at low sample sizes.

3.3. Conclusion study 2

In our empirical example, we found that when the sample size is small, different single-indicator networks (i.e., those based on a single questionnaire) produced more distinct results, as would be expected in the context of measurement error. As sample size increases, agreement between single-indicator networks grows substantially. Consistent with the simulation results, each of the methods for incorporating multiple indicators are in reasonably close agreement with each other.

4. Discussion

In this study, we investigated the effects of measurement error on the performance of cross-sectional network models. We began with a simulation study, in which we could investigate the impact of random measurement error in a context where both the underlying network is known, and the amount of measurement error could be manipulated. In these analyses, when measurement error was absent, single-indicator networks were in near perfect agreement and exhibited near perfect recovery of the true network, even when sample sizes were small. However, performance worsened with increasing measurement error, especially in the context of small samples. When sample size was small and measurement error was high, single-indicator networks were in poor agreement with one another and with the true network. In particular, measurement error substantially decreased the ability of single-indicator networks to detect edges present in the true network. As sample size increased, these deleterious effects were substantially diminished: agreement among different indicator networks, agreement with the true network and sensitivity all substantially improved, though there was a small increase in the number of spurious edges. These findings were paralleled in our empirical analyses: we found that as sample size grew, the single-indicator networks came into increasingly close agreement with one another. Together, these empirical and simulation results suggest that measurement error may impair the ability of single-indicator networks to detect the true network but that this impairment is mitigated when the sample is large.

In our simulation study, we found that, when measurement error was present, incorporating multiple indicators per variable considerably improved the performance of network models. Networks based on average scores, factor scores, and latent variables each demonstrated comparably good sensitivity and a strong correlation between edges in the estimated and true networks for samples as small as 500. For sample sizes greater than 1,000, latent network models outperformed both average scores and factor scores, exhibiting both better precision and better specificity. Thus, while large samples may mitigate the effects of measurement error, these simulation results suggest that it is considerably more efficient to address measurement error by improving node reliability (e.g., by gathering multiple indicators per node).

4.1. Implications for empirical research

Our results suggest moderate to large levels of measurement error can indeed be problematic when estimating psychometric networks. This problem can be ameliorated when working with large samples, but there is a clear need for methods that reduce the impact of measurement error. On the basis of these results, we make several suggestions for researchers interested in using network psychometric approaches. First,

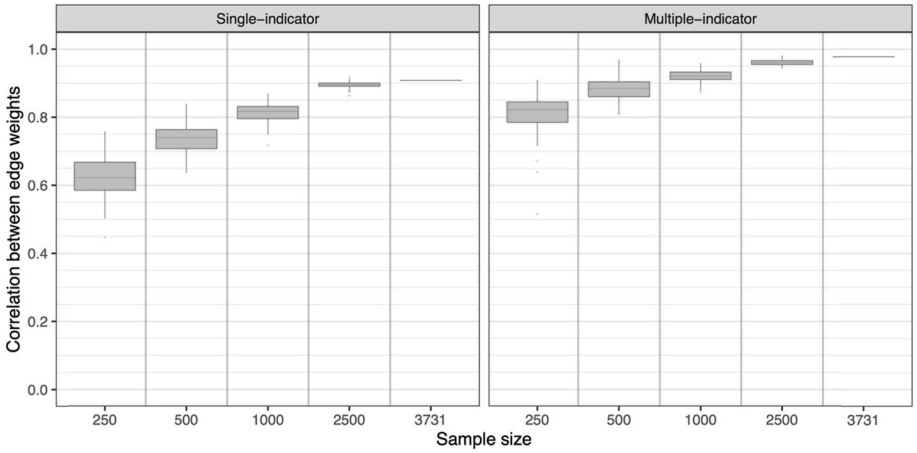

Fig. 5. Simulation results for the similarity between three single-indicator networks and the three multiple-indicator networks estimated with varying sizes of sub-samples of the STAR*D data. The left panel depicts correlations between the edge weights of the three different single-indicator networks (M1a, M1b, M1c), where the indicators are different across networks. The right panel depicts correlations between the three different multiple-indicator networks (M2-M4), where the indicators are the same but the method of combining information from multiple indicators is different. Every condition was sampled 100 times and the boxplots represent the distribution of the mean correlation between the edge weights (i.e., 25th quartile, median, 75th quartile). Importantly, these analyses do not permit direct comparison between the single-indicator networks on the one hand, which examined the similarity between networks based on different indicators, and multiple-indicator networks on the other, which examine the similarity between networks using the same indicators but different methods for combining those indicators when assessing a given node.

and most fundamentally, researchers should take steps to reduce measurement error as they assess their nodes of interest. This is, of course, not novel advice (Bollen & Lennox, 1991; Cronbach & Meehl, 1955), but bears repeating in the context of the network literature where questionnaires were often developed with a broad focus on measuring conceptually complex syndromes (e.g., depression) rather than on precisely measuring the individual symptoms (e.g., feelings of inadequacy) that are of interest in most network psychometric studies of mental health. In the current study, we focused on improving the reliability of node estimation via the use of multiple indicators, as this has been repeatedly recommended in the debate on network replicability. However, any efforts to reduce measurement error, such as developing good single-item indicators, will strengthen our ability to accurately recover the true relationships among symptoms.

A second, related point is that the cross-sectional network psychometric literature will benefit from the development of questionnaires designed with network analysis purposes in mind, namely items targeted at the symptom level instead of the disorder level. As we have demonstrated here, methods that combine information from multiple indicators are an effective way of mitigating the effects of measurement error. However, to fully evaluate and realize the value of this approach, it will be necessary to gather data that includes multiple indicators per node. An example of a questionnaire designed on symptom level is the Inventory of Depression and Anxiety Symptoms scale (IDAS; Watson et al., 2007). Unfortunately, questionnaires that assess every node of interest with multiple items are rare, further underscoring the value of developing questionnaires specifically designed for network analysis. More commonly, questionnaires may include multiple items for some symptoms, but not others. For example, the Center of Epidemiological Studies in Depression Scale (CESDS, Radloff, 1977) includes items assessing "feeling blue", "sad mood", and "depressed mood." Combining these items will potentially not only reduce the impact of measurement error, but also the bias in network estimation that can come from including semantically overlapping items in a single network (Fried & Cramer, 2017; Hallquist, Wright, & Molenaar, 2019).

Importantly, the effort to develop questionnaires with multiple items per symptom will require careful consideration of what experiences constitute the symptom. Symptoms are often 'compound' symptoms, in which a given symptom can be met by endorsing qualitatively different, and even opposing, experiences (e.g., decreased, or increased appetite; even though these are opposing, we note that one can endorse both in a time period of e.g., 2 weeks). It is an open question how to treat these compound symptoms, and there are various options. First, one can aggregate items like 'decreased appetite' and 'increased appetite' so that a low score represents decreased appetite, and a high score represents increased appetite. Unfortunately, it would then be very different to compare such nodes to other nodes, given that their extremes have very different meanings to compared to usual symptom scores. Second, one can treat opposing conditions as separate nodes in the network, as done in the current study. Finally, one can aggregate opposing symptoms into one compound variable, where low values represent absence of problems, and high values represent more severe or frequent problems. We repeated our analysis using this approach, and found results to be very similar (as depicted in Fig. S5, see supplementary materials). However, when opposing symptoms were disaggregated and treated as unique components of the network, we observed a slightly greater agreement among the single-indicator networks (Fig. 5) relative to when these opposing items were aggregated as part of the same symptom (Fig. S2). In summary, the effort to better understand and address measurement error for a given symptom will be inextricably tied to our conceptualization of the nature of that symptom and, thus, is an important topic that warrants further examination (Fried & Cramer, 2017; Rhemtulla et al., 2020; Robinaugh et al., 2019; Wilshire, Ward, & Clack, 2021).

Third, our simulation results suggest that choosing which method to combine items may be of less concern than choosing to combine items in some way. However, we note that we simulated data from one of many possible data-generating mechanisms, namely where the indicators adhere to a latent variable model (i.e., indicators are caused only by the underlying variable of interest plus measurement error). Accordingly, the results of our study can only be assumed to hold under these conditions, and our conclusion are hence limited to such (straight-forward) scenarios. In empirical data, the construct-item relations are generally unknown, and the extent to which this assumption holds is a judgment the researcher must make. Researchers should justify their choice of measurement model that goes beyond statistical fit (Rhemtulla, van Bork, & Borsboom, 2020). Possible methods to investigate the plausibility that indicators indeed measure the same construct are clustering (i.e., groups of strongly and fully connected nodes), topological overlap, and theoretical considerations. A more detailed description of these methods is out of scope for the current paper, but we refer the reader to Fried and Cramer (2017) for a discussion on challenges of selecting networks elements. Averages have the advantage that indicators do not have to imply a latent variable structure and earlier research shows that under alternative construct-item relations, such as formative models, averages can be less biased than latent variable models (Rhemtulla et al., 2020). Furthermore, given that latent network modeling is computationally intensive, average scores may be more feasible and our results suggest that they will provide a reasonable alternative.

4.2. Implications for methodological research

To our knowledge, the current study is the first attempt to quantify the impact of measurement error on the estimation of psychological cross-sectional networks. Although we consider this a valuable first step, there is considerable work on measurement error in network estimation yet to be completed. There are several important directions for future research. First, it is an open question to what extent measurement error is present in the assessment of individual symptoms and, thus, unclear how significant a threat measurement error poses for network estimation based on single items. Reliability measures for measurement scales are often reported in network studies, but it is unclear if and how they translate to the reliability of single nodes. It is therefore essential to quantify the degree of measurement error in the assessment of individual symptoms.

Second, it will be important to evaluate how the methods we have examined here perform if the number of indicators differ across latent variables, including the case where there are multiple indicators for some nodes but not others, a scenario that is likely to occur when using questionnaires not specifically designed for the purposes of network analysis. Furthermore, we could investigate the effect of using multiple indicators that are themselves composite scores, as in our empirical analysis in the supplementary material, three depression symptoms in the STAR*D study were composite scores themselves.

Third, future research could assess if methods such as clustering analysis and topological overlap, could help to inform researcher when to combine multiple indicators as composite scores or latent variables (Fried & Cramer, 2017). We investigated to what extent the assumptions of the latent variable model (and thus the factor score network and LNM) hold up in the STAR*D data. We found that the data exhibits fuzzy boundaries between clusters (each cluster consisting of a depression symptom, such as sad mood, measured three times in different ways), suggesting the clear-cut demarcation between construct assumed by the multiple-indicator networks may be violated. Furthermore, we found correlated measurement error, which violates the assumption of local independence on which the latent variable model rests. In our simulation study, we did not investigate the effect of systematic measurement error (caused by factors such as time and method of administering, i.e., self- and clinical report), which is likely to be present in empirical data. Thus, further research guiding when and how to combine items and when to leave them as distinct nodes within the network will be highly valuable.

5. Conclusion

Conventional wisdom is that psychological data are likely to contain measurement error, or even that there is no such thing as measurement without error, and that measurement error distorts the effects of interest. The goal of the current study was to investigate the consequences of this conventional wisdom in relation to network models. Our simulation results and empirical analysis suggest that, when elevated, measurement error is indeed a problem when estimating psychometric networks and offer guidance on how to mitigate the effects of measurement error. In particular, we found that the reliability of network models improves substantially (1) with sample size and (2) when combining multiple-indicators per variable.

CRediT authorship contribution statement

Jill de Ron: Conceptualization, Methodology, Formal analysis, Writing – original draft. **Donald J. Robinaugh:** Conceptualization, Formal analysis, Methodology, Writing – original draft, Supervision. **Eiko I. Fried:** Conceptualization, Methodology, Writing – original draft, Supervision. **Paola Pedrelli:** Resources, Writing – review & editing. **Felipe A. Jain:** Resources, Writing – review & editing. **David Mischoulon:** Resources, Writing – review & editing. **Sacha Epskamp:** Conceptualization, Software, Methodology, Writing – review & editing, Supervision.

Declaration of competing interest

Dr. Mischoulon has received research support from Nordic Naturals and heckel medizintechnik GmbH. He has received honoraria for speaking from the Massachusetts General Hospital Psychiatry Academy, Harvard Blog, and PeerPoint Medical Education Institute, LLC. He also works with the MGH Clinical Trials Network and Institute (CTNI), which has received research funding from multiple pharmaceutical companies and NIMH.

Dr. Jain has received salary support from the non-profit MGH Clinical Trials Network and Institute (CTNI), which has received research funding from multiple pharmaceutical companies and NIMH.

This manuscript was supported by a National Institute of Mental Health Career Development Award (1K23MH113805-01A1) awarded to Dr. D. Robinaugh, the NWO Veni Grant (016-195-261) awarded to Dr. S. Epskamp and grant support from the National Institutes of Health (#K76AG064390) awarded to Dr. F. Jain.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.brat.2022.104163.

References

Acito, F., & Anderson, R. D. (1986). A simulation study of factor score indeterminacy. *Journal of Marketing Research*. https://doi.org/10.2307/3151658

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305. https://doi.org/10.1037/0033-2909.110.2.305

Borsboom, D., Fried, E. I., Epskamp, S., Waldorp, L. J., van Borkulo, C. D., van der Maas, H. L., & Cramer, A. O. (2017). *False alarm? A comprehensive reanalysis of "evidence that psychopathology symptom networks have limited replicability" by Forbes, Wright, markon, and krueger*. https://doi.org/10.1037/abn0000306, 2017.

Briganti, G., Fried, E. I., & Linkowski, P. (2019). Network analysis of contingencies of self- worth scale in 680 university students. *Psychiatry Research, 272*, 252–257.

Bringmann, L. F., & Eronen, M. I. (2018). Don't Blame the model: Reconsidering the network approach to psychopathology. *Psychological Review* https://doi.org/10.1037/rev0000108.

Cohen, J. (1995). The earth is round ($p < .05$): Rejoinder. In *American psychologist*. https://doi.org/10.1037/0003-066X.50.12.1103

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*. https://doi.org/10.1037/a0033805

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281. https://doi.org/10.1037/h0040957

Epskamp, S. (2020). Psychometric network models from time-series and panel data. *Psychometrika*, 1–26. https://doi.org/10.1007/s11336-020-09697-3

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

Epskamp, S., Maris, G., Waldorp, L. J., & Borsboom, D. (2017). Network psychometrics. In *The wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*. https://doi.org/10.1002/9781118489772.ch30

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika*. https://doi.org/10.1007/s11336-017-9557-x

Fava, M., Rush, A. J., Trivedi, M. H., Nierenberg, A. A., Thase, M. E., Sackeim, H. A., et al. (2003). Background and rationale for the sequenced treatment alternatives to relieve depression (STAR*D) study. In *Psychiatric clinics of north America*. https://doi.org/10.1016/S0193-953X(02)00107-7

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456–465. https://doi.org/10.1177/2515245920952393

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

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

Fried, E. I., Epskamp, S., Nesse, R. M., Tuerlinckx, F., & Borsboom, D. (2016). What are 'good' depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. *Journal of Affective Disorders, 189*, 314–320. https://doi.org/10.1016/j.jad.2015.09.005

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

Hallquist, M. N., Wright, A. G. C., & Molenaar, P. C. M. (2019). *Problems with centrality measures in psychopathology symptom networks: Why network psychometrics cannot escape psychometric theory*. Multivariate Behavioral Research. https://doi.org/10.1080/00273171.2019.1640103

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery & Psychiatry*. https://doi.org/10.1136/jnnp.23.1.56

Herrera-Bennett, A. C., & Rhemtulla, M. (2021). *Network replicability & generalizability: Exploring the effects of sampling variability, scale variability, and node reliability*.

Jaccard, J., & Wan, C. K. (1995). Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation Approaches. *Psychological Bulletin*. https://doi.org/10.1037/0033-2909.117.2.348

Jones, P. J., Williams, D. R., & McNally, R. J. (2021). Sampling variability is not nonreplication: A bayesian reanalysis of Forbes, Wright, markon, and krueger. *Multivariate Behavioral Research, 56*(2), 249–255. https://doi.org/10.1080/00273171.2020.1797460

Ledgerwood, A., & Shrout, P. E. (2011). The trade-off between accuracy and precision in latent variable models of mediation processes. *Journal of Personality and Social Psychology, 101*(6), 1174. https://doi.org/10.1037/a0024776

Liu, K. (1988). Measurement error and its impact on partial correlation and multiple linear regression analyses. *American Journal of Epidemiology, 127*(4), 864–874. https://doi.org/10.1093/oxfordjournals.aje.a114870

Madhoo, M., & Levine, S. Z. (2016). Network analysis of the quick inventory of depressive symptomatology: Reanalysis of the STAR* D clinical trial. *European Neuropsychopharmacology, 26*(11), 1768–1774. https://doi.org/10.1016/j.euroneuro.2016.09.368

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods, 52*(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*. https://doi.org/10.1177/014662167700100306

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*. https://doi.org/10.1037/met0000220

Robinaugh, D., Haslbeck, J., Waldorp, L., Kossakowski, J., Fried, E. I., Millner, A., … Borsboom, D. (2019). *Advancing the network theory of mental disorders: A computational model of panic disorder*.

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

de Ron, J., Fried, E. I., & Epskamp, S. (2021). Psychological networks in clinical populations: Investigating the consequences of Berkson's bias. *Psychological Medicine, 51*(1), 168–176. https://doi.org/10.1017/S0033291719003209

Rosseel, Y. (2012). lavaan: an R package for structural equation modeling and more. *Journal of Statistical Computing*. https://doi.org/10.18637/jss.v048.i02

Rush, A. J., Fava, M., Wisniewski, S. R., Lavori, P. W., Trivedi, M. H., Sackeim, H. A., et al. (2004). Sequenced treatment alternatives to relieve depression (STAR*D): Rationale and design. In *Controlled clinical trials*. https://doi.org/10.1016/S0197-2456(03)00112-0

Rush, J. A., Giles, D. E., Schlesser, M. A., Fulton, C. L., Weissenburger, J., & Burns, C. (1986). The inventory for depressive symptomatology (IDS): Preliminary findings. *Psychiatry Research*. https://doi.org/10.1016/0165-1781(86)90060-0

Rush, A. J., Gullion, C. M., Basco, M. R., Jarrett, R. B., & Trivedi, M. H. (1996). The inventory of depressive symptomatology (IDS): Psychometric properties. *Psychological Medicine, 26*(3), 477–486. https://doi.org/10.1017/S0033291700035558

Schmidt, F. L., & Hunter, J. E. (1999). *Theory testing and measurement error*. https://doi.org/10.1016/S0160-2896(99)00024-0

Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods, 24*(1), 70. https://doi.org/10.1037/met0000188

Schuurman, N. K., Houtveen, J. H., & Hamaker, E. L. (2015). Incorporating measurement error in n= 1 psychological autoregressive modeling. *Frontiers in Psychology, 6*, 1038. https://doi.org/10.3389/fpsyg.2015.01038

Staudenmayer, J., & Buonaccorsi, J. P. (2005). Measurement error in linear autoregressive models. *Journal of the American Statistical Association, 100*(471), 841–852. https://doi.org/10.1198/016214504000001871

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Team, R. D. C., & R Development Core Team, R. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://doi.org/10.1007/978-3-540-74686-7

Van Borkulo, C. D., Boschloo, L., Kossakowski, J., Tio, P., Schoevers, R. A., Borsboom, D., et al. (2017). *Comparing network structures on three aspects: A permutation test* (p. 10). Manuscript submitted for publication.

Watson, D., O'Hara, M. W., Simms, L. J., Kotov, R., Chmielewski, M., McDade-Montez, E. A., et al. (2007). Development and validation of the inventory of depression and anxiety symptoms (IDAS). *Psychological Assessment, 19*(3), 253–268. https://doi.org/10.1037/1040-3590.19.3.253

Wilshire, C. E., Ward, T., & Clack, S. (2021). Symptom descriptions in psychopathology: How well Are they working for us? *Clinical Psychological Science, 9*(3), 323–339. https://doi.org/10.1177/2167702620969215