Cognitive Therapy and Research
https://doi.org/10.1007/s10608-026-10719-0

ORIGINAL ARTICLE

Comparing Multiple-Indicator Approaches to Account for Measurement Error in Dynamic Networks

Reeta Kankaanpää[1] · Jill de Ron[2] · Ria H. A. Hoekstra[2,3] · Riet van Bork[2]

Received: 5 May 2025 / Accepted: 31 January 2026

Abstract

Background To better understand the development of mental disorders, dynamic networks have gained more attention in recent years. Most of these network models use a single indicator per node despite the fact that measurement error may bias parameter estimates. This can lead to incorrect conclusions about the presence or absence of edges, as well as the relative strength of edges in the network. In this study, we compared single-indicator dynamic networks to approaches using information on multiple indicators per node to account for measurement error.

Data and Methods We conducted two simulation studies, using time series (Study 1, N = 1) and panel (Study 2, N > 1) data, to compare the estimation of network parameters in the presence of measurement error in models with single indicators versus models with multiple indicators, namely as latent variables, plausible values, factor scores, and average scores. Across conditions, we varied the variance of the measurement error and the number of observations (in time series: number of timepoints; and in panel data: number of persons and waves). We evaluated the performance of each model by examining the correlation between the estimated and true network edge weights, as well as the sensitivity, specificity, and precision.

Results In both studies, measurement error decreased correlations between the true and estimated network as well as sensitivity among all approaches, while specificity and precision were mostly unaffected. The single-indicator approach was the most sensitive to measurement error and the number of observations compared to other approaches. In Study 1, the factor and average score approaches performed best for temporal networks, and the latent variable approach for contemporaneous networks. In Study 2, generally the best-performing approach was the plausible value score.

Discussion Measurement error may substantially bias estimates in dynamic networks, and multiple-indicator approaches can mitigate this bias. Multiple-indicator approaches generally outperformed the single-indicator approach, but the choice between different multiple-indicator approaches depends on several factors that must be carefully considered before deciding the best method for each study.

Keywords Measurement error · Dynamic network model · Time series data · Panel data · Simulation study · Multiple-indicators · Vector autoregressive models

Background

The psychological network approach conceptualizes psychological constructs as networks of interacting behaviors, feelings, and cognitions (Cramer et al., 2010). In the context of psychopathology, mental disorders can be conceptualized as networks of symptoms that mutually reinforce and sustain each other (Borsboom, 2017). Network psychometrics is the collective term for statistical models that estimate network structures from psychological data (Epskamp et al., 2017), and its use has grown rapidly in recent years (Robinaugh et al., 2020). Today, the network framework offers researchers

✉ Reeta Kankaanpää
reeta.kankaanpaa@tuni.fi

Jill de Ron
j.deron@uva.nl

Ria H. A. Hoekstra
h.a.hoekstra@uva.nl

Riet van Bork
r.vanbork@uva.nl

[1] University of Turku, Turku, Finland

[2] University of Amsterdam, Amsterdam, The Netherlands

[3] Utrecht University, Utrecht, The Netherlands

Published online: 16 March 2026

a wide variety of statistical tools to analyze cross-sectional and longitudinal data (Briganti et al., 2024).

Within this framework, dynamic network models offer a promising method for uncovering how variables interact and evolve over time. In clinical psychology, dynamic network models have been proposed as a way to identify symptom pathways, i.e., symptoms that activate or deactivate other symptoms (Castro et al., 2019). Dynamic networks can be based on time series (N = 1) or panel data (N > 1) (Hamaker & Wichers, 2017). Time series data, especially intensive longitudinal data from a single person, allow researchers to study intraindividual symptom pathways and potentially develop personalized psychotherapies (Bringmann, 2021; Schumacher et al., 2024). Panel data, which involve data of many individuals measured on only a few and more widely spaced time points, are useful for investigating dynamics of more stable traits and for separating intraindividual and between-subject dynamics (Epskamp et al., 2018b).

Because longitudinal studies are more demanding for respondents than cross-sectional studies, researchers often opt for single-indicator measures (Dejonckheere et al., 2022; Eisele et al., 2022). Unfortunately, single-indicator measures have been shown to be prone to measurement error, which can bias parameter estimates. In bivariate associations, measurement error attenuates correlations (Spearman, 1904). In multivariate cases, an attenuated correlation between two variables may lead to inflated correlations with a third variable. As the number of variables increases, the inflating or attenuating effects of measurement error become less predictable (Cole & Preacher, 2014). Longitudinal studies add complexity in predicting the effects of measurement error on parameter estimation as data may include effects within a timepoint as well as over time.

Multiple indicators could be used to account for measurement error by extracting the shared variance across multiple indicators of the same construct. However, multiple indicators add an extra burden to participants, especially in longitudinal studies, and it is therefore important to investigate their added benefit. The current study aims to do exactly this. Using a simulation study, we aim to evaluate how well dynamic networks based on single-indicators and on different multiple-indicator approaches perform when the data contain measurement error.

This paper is structured as follows. First, we discuss two dynamic networks models: for time series (N = 1) and panel (N > 1) data. Second, we describe previous research on the effects of measurement error on (network) models. Third, we introduce the different measurement approaches for multiple indicators, namely latent variable modeling, plausible values scores, factor scores, and average scores. Fourth, in the method section, we describe the set-up of our simulation study. To evaluate the performance of these different approaches, we will examine the correlation between the estimated and the data-generating network, sensitivity, specificity, and precision. Fifth, in the results section we show that measurement error decreased correlations and sensitivity among all approaches, while specificity and precision were mostly unaffected. Generally, the multiple-indicator approaches outperformed the single-indicator approach. Lastly, in the discussion, we provide insights into which multiple indicator approach works best under which conditions, guiding researchers in selecting the most suitable approach when estimating dynamic networks.

Dynamic Networks

Idiographic methods have become increasingly popular for personalizing assessment and therapy (Frumkin et al., 2021; Hamaker & Wichers, 2017; Piccirillo & Rodebaugh, 2019; Wright & Woods, 2020). The rise of intensive longitudinal data collection has enabled the estimation of person-specific models (Wright & Woods, 2020). In time series studies, individuals are repeatedly measured over an extended period. One of the most widely used models within the network approach for such data is the graphical vector-autoregressive (GVAR) model, which can account for the temporal dependencies in the data (Epskamp et al., 2018b; Jordan et al., 2020). Additionally, the residuals of the VAR model are modeled to capture contemporaneous associations.

Consequently, dynamic network models typically consist of two components: a *temporal* network capturing autoregressive and cross-lagged relations between variables, and a *contemporaneous* network capturing associations between variables within the same time point. The temporal network forms a directed network where variables at time t are predicted by variables at time t – 1. Residual variances and covariances unexplained by the temporal dynamics are modeled in the contemporaneous network, resulting in an undirected network structure. In studies with multiple individuals but with a limited number of assessments per individual (e.g., pre-intervention, post-intervention, and follow-up), a GVAR panel network model can be estimated. Besides capturing within-subjects temporal and contemporaneous associations, panel networks also provide between-subjects networks based on the stable trait-like means (Epskamp et al., 2018a; Epskamp, 2020a). However, between-subject networks are not restricted to panel networks; if time-series data are collected from multiple individuals, a between-subject network can also be estimated using a multilevel network model (Epskamp et al., 2018a).

GVAR models include ‘innovations’ (or ‘dynamic errors’), which refer to unmeasured influences on observed variables that persist across time points. However, they do not explicitly account for measurement error. Measurement

error differs from innovations in that it is typically conceptualized as white noise, i.e., random, uncorrelated error over time (Lütkepohl, 2005; Schuurman et al., 2015). In the following section, we discuss previous research on the effect of measurement error on (network) models.

The Effect of Measurement Error

Measurement error has been shown to impair reliability and attenuate edge weights in cross-sectional network analysis (De Ron et al., 2022; Herrera-Bennett & Rhemtulla, 2021). De Ron et al. (2022) demonstrated that measurement error substantially reduced the ability of single-indicator networks to detect true edges. In a similar vein, Herrera-Bennett and Rhemtulla (2021) found that using multiple indicators per node (i.e., sum scores) improved network sensitivity, enhanced estimates of global strength, and increased the consistency of network properties such as edge weights and centrality scores. Moreover, enhancing node reliability (using two indicators instead of one) improved correlation estimates between nodes equivalent to increasing the sample size by 2.5 times. However, these findings are limited to cross-sectional data.

In longitudinal settings, the effect of measurement error is similarly problematic. Staudenmayer and Buonaccorsi (2005) have demonstrated that, in the context of an autoregressive lag 1 (AR(1)) model, the autoregressive coefficient is biased toward zero in the presence of measurement error. Using an empirical study of the AR(1) model, Schuurman et al. (2015) found that over a third of the total variance in indicators could be attributed to measurement error, and neglecting this error resulted in a significant underestimation of autoregressive effects. Extending this to a bivariate AR model, Schuurman and Hamaker (2019) showed that ignoring measurement error can distort both the autoregressive effects as well as the cross-lagged relations between variables.

To address these issues, various methods have been evaluated in light of measurement error in longitudinal data. For single indicator models, Schuurman et al. (2015) compared an autoregressive model with a white noise term (AR + WN) to an autoregressive moving average (ARMA) model concluding that the AR + WN model performed better. Building on this, Schuurman and Hamaker (2019) presented a method that accounts for measurement error in longitudinal single-indicator models.

In dynamic structural equation modeling (DSEM), several studies have demonstrated the biasing effects of measurement error on temporal relations in time series by comparing dynamic factor analysis (DFA) model using multiple indicators to single-indicator (average score) AR models that either ignore or account for measurement error (O'Laughlin et al., 2021; Oh & Jahng, 2023; Oh et al., 2025). All three studies concluded that the AR model ignoring measurement error showed most bias, whereas DFA and the AR model accounting for measurement error performed comparably. These results show that for other types of longitudinal models approaches that deal with measurement error have been developed and lead to improved results.

As measurement error decreases the reliability of the indicators, researchers have developed several methods to estimate reliability in longitudinal studies. For instance, Dejonckheere et al. (2022) proposed two test–retest reliability coefficients. For evaluating intraindividual reliability in multiple-indicator models, person-specific reliability can be estimated using p-factor analysis (Hu et al., 2016), a two-level random dynamic model (Xiao et al., 2023), or the mixed-effects trait-state-occasion model (Castro-Alvarez et al., 2022). Castro-Alvarez et al. (2024a, 2024b) provide a concise overview of the methods to estimate between-person and within-person reliability in longitudinal studies.

Studies by Schuurman et al. (2015), Schuurman and Hamaker (2019), O'Laughlin et al. (2021), Oh and Jahng (2023), and Oh et al. (2025) focused only on uni- or bivariate relations. We think that some of these methods could be modified to apply also to multivariate dynamic network models, however since we did not find such modified versions in the literature, we did not include these methods in the simulation study and we consider the development of such methods beyond the scope of this article. Our study adds to this body of knowledge by comparing the effectiveness of single- and multiple-indicator measurement approaches in recovering the true network structure in the presence of measurement error. In the following section, we discuss various measurement approaches for including multiple indicators per node.

Measurement Approaches for Multiple Indicators

When multiple indicators for each node are available, several different methods can be employed to address measurement error in network estimation. One approach is to account for measurement error using latent variables, where each node in the network is represented by a latent variable measured by multiple indicators (Epskamp et al., 2017). In this approach, the variance shared across the set of indicators for a node reflects the latent variable, while the unique variance in the indicators is treated as measurement error. Recently, the Gaussian Graphical Model (GGM) with latent variables has been extended to include temporal relationships (Epskamp, 2020a). With GVAR models, it is now possible to estimate dynamic networks with latent variables using time series and panel data, a functionality available in the psychonetrics R package (Epskamp, 2020b). However, a

notable caveat of latent variable network models is that they require a large number of parameters to be estimated, necessitating large datasets with multiple observations, which are often unavailable in psychology research (Van Agteren et al., 2021; Wrzus & Neubauer, 2023).

Alternatively, measurement errors can be addressed by using plausible values, factor scores, and average scores. The Bayesian approach of plausible value scoring has shown promise in accurately estimating scores for latent variables using imputation techniques and multiple draws (Von Davier et al., 2009). In this technique, instead of directly estimating a person's value on the latent variable, a posterior distribution for this person's value on the latent variable is estimated that captures the uncertainty and is used to randomly draw plausible values from. Wu (2005) describes plausible values as representing the range of abilities that a subject might reasonably have, given the subject's item responses. Multiple studies have been published demonstrating that plausible values are an efficient way of estimating parameters, in some cases even outperforming maximum likelihood estimation (Laukaityte & Wiberg, 2017; Von Davier et al., 2009; Wu, 2005). The marginal distribution of plausible values has been demonstrated to be a consistent estimator of the true ability distribution, even when the population model is misspecified (Marsman et al., 2016). This suggests that even though plausible value estimation includes imputation, it can be effective even in the absence of covariates in the model. However, similar to latent variable models, estimating plausible values requires a larger number of observations compared to approaches like factor scores or average scores (Marsman et al., 2016).

Factor scores assign unique weights to each indicator based on their contribution to the factor. These scores are typically estimated using regression analysis (Thomson, 1934; Thurstone, 1935) or the Bartlett method (Bartlett, 1937; Thomson, 1938). However, factor scores are subject to factor indeterminacy (i.e., the factor solution is not uniquely determined), and they cannot fully eliminate measurement error (Grice, 2001; Mulaik, 1972; Rigdon et al., 2019; Steiger, 1979). Several methods, such as Croon's correction (Croon, 2002), have been proposed to improve factor scores, but these corrections are not widely accessible and require manual computation (Devlieger et al., 2019). In the average score approach, each node is represented by the average of the values on its indicators. While average scores improve node reliability by averaging out random errors, they treat each indicator equally and do not fully eliminate measurement errors.

The use of approximate scores—such as plausible values, factor scores, or average scores—follows a two-stage process. In these approaches, scores are first estimated and then used as proxies for the latent variable in further analysis. In contrast, when using latent variables, the measurement model and the structural (network) model are estimated simultaneously, necessitating a larger sample size. In this study, we compare these various multiple-indicator approaches for accounting for measurement error to a single-indicator approach that ignores measurement error entirely. We evaluate how well each approach retrieves the true network parameters and investigate the extent to which ignoring measurement error biases network estimates. This comparison will shed light on the effectiveness of different strategies for handling measurement error in dynamic network models and their impact on parameter accuracy.

Methods

This simulation study was preregistered on OSF (https://doi.org/10.17605/osf.io/khtgc) using the preregistration template provided by Siepe et al. (2023). The OSF page also contains the fully reproducible simulation R script and a table with the simulation results. We used R version 4.4.0 (R Core Team, 2021). The empirical datasets we used can be freely accessed online (time series dataset for Study 1 by Kossakowski et al., 2017, and panel dataset for Study 2 by McBride et al., 2021).

The primary aim of the simulation study was to compare the performance between the different measurement approaches (latent variable, plausible value score, factor score, average score, and single-indicator). A secondary target was to explore how the performance of the networks depends on the amount of measurement error. We wanted to investigate at what level of the measurement error variance each network, with the different measurement approaches, would start to perform worse, exhibiting inflated or deflated edge weights, and lowered specificity and sensitivity. Our third aim was to explore which networks (i.e., contemporaneous, temporal, or between-subjects) are most affected by measurement error. As it is known that more complex models (latent variables, plausible value scores, and factor scores) require more observations to produce stable estimates (De Ron et al., 2022; Epskamp, 2020a; Mansueto et al., 2023), we also investigated the minimum required sample size for using more complex models. This was investigated by computing the number of errors as failed estimations (e.g., due to non-convergence) per measurement approach, see Supplementary Materials S5.

To do so, our simulation study consists of four steps. In Step 1, we constructed the data-generating network based on empirical data. In Step 2, we used the data-generating network to simulate three indicators per node plus random measurement error. In Step 3, we used the simulated data to estimate one single-indicator and four multiple-indicator

network models, namely where nodes are modeled as latent variables, plausible values, factor scores, and average scores. In Step 4, we assessed the performance of these methods by comparing the estimated networks with the data-generating network.

We completed these simulations under varying conditions: we varied two factors in the simulation design of Study 1 and three factors in Study 2. First, we manipulated the variance of the measurement error using values 0, 0.5, 1, 1.5, and 2, based on prior research assessing the reliability of construct scales in time series and panel data (Castro-Alvarez et al., 2024a, 2024b; Dejonckheere et al., 2022; Freichel et al., 2023; Schuurman et al., 2015; Schuurman & Hamaker, 2019). Second, in Study 1 (N = 1 time series), we varied the number of observations across time points: $t_{timepoints}$ = 100, 250, 500, 750, 1000. In Study 2 (panel data), we varied both the number of persons: $n_{persons}$ = 100, 250, 500, 1000, 2000, and the number of waves: t_{waves} = 3, 6. Note that the preregistration also included a condition of $n_{persons}$ = 5000, but we have decided to drop this condition as the simulation design was getting too large and computationally expensive. These ranges were based on prior studies reporting typical to ideal dataset sizes (Mansueto et al., 2023; Martín-Gómez et al., 2022; McBride et al., 2021; Rigabert et al., 2020; Wrzus & Neubauer, 2023). The conditions were varied in a fully factorial design. This resulted in 5 (measurement approaches) × 5 (measurement error sizes) × 5 (timepoints) = 125 conditions for Study 1 (N = 1 time series). In Study 2 (panel data), the design comprised 5 (measurement models) × 5 (measurement error sizes) × 5 (sample size) × 2 (waves) = 250 conditions. Each condition was simulated 100 times. We review each step of the simulation study in further detail below.

Step 1: Constructing the Data-Generating Network Based on Empirical Data

The data were generated using dynamic network models estimated on empirical datasets to ensure that the parameter values were realistic. The data-generating network models consisted of six nodes, chosen to align with the limited number of variables typically included in dynamic latent network models. Below, we describe how we constructed the data-generating network for the N = 1 time series data (Study 1) and the panel data (Study 2).

Study 1: N = 1 Time Series Data

The data-generating network for the N = 1 time series study was based on an open-access dataset containing momentary affective states from a patient diagnosed with major depressive disorder (Kossakowski et al., 2017). We estimated a GVAR model using six out of the nine mood-related variables: ("I feel…") "down", "irritated", "lonely", "suspicious", "indecisive", or "strong". These items were asked ten times a day for 84 days[1] using a scale from 1 (not) to 7 (very). We selected these six variables because they resulted in a network that was moderately dense (i.e., it contained several edges but was not fully connected), which works well for assessing both sensitivity and specificity. Prior to estimating the network model, the variables were detrended to avoid violating the assumption of stationarity.

We estimated the data-generating temporal and contemporaneous networks using the function 'tsdlvm1' from the psychonetrics package (Epskamp, 2020b), setting the loadings matrix (lambda) to be an identity matrix. This means that each latent node was associated with only one observed indicator variable (i.e., each variable in the dataset represented a separate node in the network). We chose tsdlvm1 over other available functions to maintain consistency with the methods used in the simulation study, as this function also supports multiple indicators per node. Full Information Maximum Likelihood (FIML) estimation was employed due to the presence of missing values in the dataset.[2] The model was pruned at an alpha level of 0.05, removing edges that were not significantly different from zero.

Study 2: Panel Data

The network model for the panel study was based on the open-access COVID-19 psychological research consortium (C19PRC) dataset (McBride et al., 2021), which includes mental health variables measured over six waves. From this empirical dataset, we estimated a panelGVAR model using six of the nine depression-related variables: "Little interest or pleasure in doing things", "Trouble falling or staying asleep, or sleeping too much", "Feeling tired or having little energy", "Feeling bad about yourself—or that you are a failure or have let yourself or your family down", "Trouble concentrating on things, such as reading the newspaper or watching television", and "Moving or speaking so slowly that other people have noticed? Or the opposite—being so fidgety or restless that you have been moving around more than usual". Participants were asked how often, over the last two weeks, they had been bothered by each of the depressive symptoms, and the response options were "not at all", "several days", "more than half the days", and "nearly every day", scored as 0, 1, 2 and 3, respectively. Just like in the N = 1 study, we selected six variables that resulted in a moderately dense model to enhance the assessment of

[1] Because of the assumed stationarity in the analyses, we selected only the post-assessment phase, following the example from Epskamp (2020a).

[2] Mean frequency of responses per day was 5.8.

specificity and sensitivity. Before estimating the model, the variables were detrended to avoid violating the stationarity assumption.

We estimated the network using the 'dlvm1' function from the psychonetrics package (Epskamp, 2020b), setting the loadings matrix (lambda) to be an identity matrix. After running the model, we pruned it with an alpha level set to 0.05. We estimated within-subjects temporal, contemporaneous, and between-subjects networks from the empirical data, using the estimated edge weights as the true edge weights in our true data-generating model. As with the N = 1 time series data, this resulted in a temporal network. However, unlike the N = 1 time series data, the residual covariance within timepoints, after accounting for temporal dependencies, was used to estimate both contemporaneous relations at the within-subjects level (included in the matrix $\mathbf{\Omega}(\zeta_{\text{within}})$) and at the between-subjects level (included in the matrix $\mathbf{\Omega}(\zeta_{\text{between}})$).

Step 2: Simulate Data with Measurement Error

For Study 1, we used the function 'graphicalVARsim' with the estimated temporal and contemporaneous network as input to generate time series data. For Study 2, we used the function 'generate_paneldata' accessible in our shared code, with the estimated within-subjects temporal, contemporaneous, and between-subjects network as input to generate panel data. To introduce measurement error, each of the six nodes in the data-generating network was treated as a latent variable, and we simulated three indicators for each node. These indicators were modeled as linear functions of the latent node plus random measurement error. Consequently, the simulated dataset contained a total of 18 observed variables (3 indicators × 6 latent nodes). That is, for each latent node, we constructed three normally distributed indicator variables that loaded on the latent node N_j as follows:

$$X_{1j} = \lambda_{1j} N_j + E_{1j}$$

$$X_{2j} = \lambda_{2j} N_j + E_{2j}$$

$$X_{3j} = \lambda_{3j} N_j + E_{3j}$$

where λ_{1j} represents the loadings of the first indicator (X_{1j}) on node j (N_j), and E_{1j} is the measurement error variable of the first indicator of node j. Based on a helpful reviewer's comment, we decided to deviate from what we stated in the preregistration and, instead of giving all indicators the same loading of 1, draw loadings from a uniform distribution ranging from 0.75 to 1 for some variation in loadings. The measurement error variables were generated as multivariate random normal variables with mean 0 and measurement error variances that were equal across indicators, but that we varied over conditions (0, 0.5, 1, 1.5 and 2). The covariance matrix for the measurement error variables, Θ_j, was set diagonal (uncorrelated) and the diagonal elements took the values of the vector of measurement error variances. The covariance matrix of the three indicators X_{1j} to X_{3j}, Σ_j, is a function of the loadings on N_j, the variance of N_j and Θ, $\Sigma_j = \Lambda_j \psi_j \Lambda_j' + \Theta_j$, where Λ_j is the vector with loadings sampled from a uniform distribution ranging from 0.75 to 1, and ψ_j is the variance of node j.

Step 3: Estimate Measurement Approaches Based on Simulated Data

We use five measurement approaches to estimate networks from the simulated data. The first approach involves modeling each node as a latent variable. This approach is implemented in the 'lvgvar' framework (Epskamp, 2020b), which can be applied to both time series and panel data and is available in the open-source *psychonetrics* package. In the second approach, nodes are measured by plausible value scores. We specified a confirmatory factor analysis model for each node, using the three indicators per node, estimated the model using the *lavaan* package (Rosseel, 2012) and obtained the plausible value scores using the *semTools* package (Jorgensen et al., 2022). We estimated a network on the resulting plausible value scores. The third approach measured nodes using factor scores. For each node, we specified a confirmatory factor analysis model with the three indicators based on the "Bartlett" method in the *lavaan* package (Rosseel, 2012). We estimated a network based on the resulting factor scores. The fourth approach involves measuring nodes using average scores. Here, we computed an average score of the three indicators per latent node and used this average score as the observed variable in the model. The fifth approach uses single indicators. In this case, we randomly selected one of the three indicators per latent node to serve as the observed variable (and measure of the node) in the model.

For Study 1 (N = 1 time series data), we used the 'gvar' function from the *psychonetrics* package (Epskamp, 2020b) for single indicators, average scores, factor scores, and plausible value scores. For the latent network model, we used the 'tsdlvm1' function in the time series setting, with identification set to "loadings" and we specified the lambda matrix such that each set of three indicators loaded onto their specified latent node. In Study 2 (panel data), we used the 'panelgvar' function for the single indicators, average scores, factor scores, and plausible value scores. For the latent network model, we used the 'dlvm1' function, setting the lambda matrix in the same way as in Study 1. All these functions are accessible through the 'psychonetrics'

package (Epskamp, 2020b). We performed pruning at an alpha of 0.05 for all networks and ran the models with the approximate_SEs argument set to TRUE due to a high number of convergence issues in the latent variable approach encountered during the simulation pilot. The approximate_SEs argument allows the use of approximate matrix inverse of the Fischer information to obtain standard errors.

Step 4: Compare Estimated Networks with Data-Generating Network

In the last step, we compared the estimated networks of the different measurement approaches with the data-generating network. Our primary performance measures were (1) the correlation between the edge weights of the estimated and the true network, (2) sensitivity, (3) specificity, and (4) precision. The performance measures were computed for temporal and contemporaneous networks in Study 1, and additionally for the between-subjects networks in Study 2. Sensitivity (also known as true positive rate) was computed as the ratio between the true positives and the sum of the true positives and the false negatives. Specificity (also known as true negative rate) was computed as the ratio between the true negatives and the sum of the true negatives and the false positives. Precision was computed as the ratio between the true positives and the sum of the true positives and the false positives. As secondary performance measures, we computed the average bias and absolute bias in order to gain insight into the average deviation of the estimated edge weights from their true values and their relative deviation. The average bias was computed as the difference between the estimated edge weights and the true edge weights. The average absolute bias was computed as the absolute difference between the estimated edge weights and the true edge weights.

Simulation Study 1: Time Series Data (N = 1)

Supplementary Materials S1 show the data-generating networks estimated from the time series data. Figures 19 and 20 in Supplementary Materials S4 show the bias results. Out of the five evaluated measurement approaches, the latent variable approach showed the most bias, which was primarily evident for the contemporaneous network. However, bias was still minimal. For a more detailed discussion on these results, see Supplementary Materials S4.

The results on the main performance measures (correlation, precision, sensitivity and specificity) of the different measurement approaches are presented in Fig. 1 for the temporal network and Fig. 2 for the contemporaneous network. For the condition without measurement error, we only plotted the performance measures of the single-indicator approach, because in case of no measurement error, all three indicators are exactly equal, making it impossible to estimate latent variables, plausible values, or factor scores, and making the average score identical to the single-indicator. We included the condition without measurement error to compare the performance in the absence of measurement error to the performance of different methods in the presence of measurement error. Below, we first discuss the results for the temporal network and then the contemporaneous network. Per network, we discuss the performance per measurement approach (latent variable, plausible value, factor score, average and single indicator) in such a way that for each approach we will consider all four performance measures (correlation, precision, sensitivity and specificity) and make comparisons between approaches. We finish with a summary of the results for this simulation study before moving to Study 2.

Temporal Networks

Figure 1 presents the performance of the different measurement approaches in the temporal networks as boxplots (in addition, we present the results as means with simulation-based confidence intervals in the Supplementary Materials S2). Overall, the performance across different measurement approaches did not differ considerably. Specificity and precision showed good performance overall, whereas sensitivity was generally low but improved to moderate levels as the number of timepoints increased. In the next sections, we evaluate the performance of each of the measurement approaches, but it is important to note that low performance of an approach in a certain condition does not necessarily mean that it is because of that particular way of dealing with measurement error that there is poor performance, but that it could also be that in such conditions network estimation performs poorly in general. For example, Fig. 1 shows that many of the approaches perform poorly on sensitivity with small numbers of timepoints, but this poor performance in sensitivity is also present in the zero-measurement error condition and thus does not reflect a problem with a specific "measurement approach" but rather that network estimation in general will have low sensitivity in conditions with small sample sizes.

Latent Variable Approach

The latent variable approach showed very low correlations in conditions with high measurement error (error variance = 1.5 and 2) and the smallest number of timepoints ($t_{timepoints} = 100$). However, with reduced measurement error and increased timepoints, it achieved high correlations

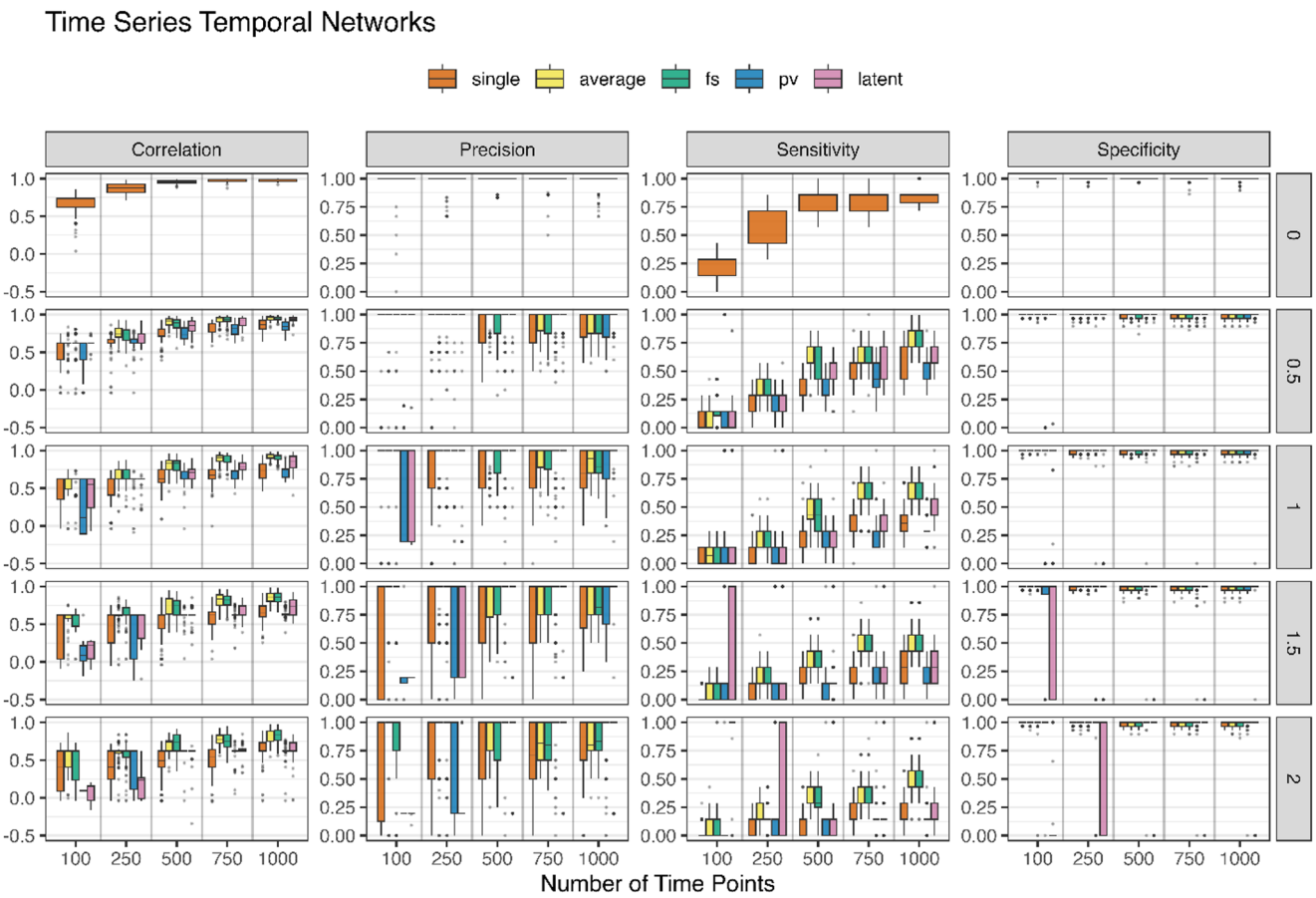

Fig. 1 Simulation results for the performance of the single- and multiple-indicator approaches for temporal N = 1 network models. The vertical panels indicate the different performance measures: correlation, precision, sensitivity, and specificity. Horizontal panels indicate the amount of measurement error. The dots indicate outliers. Every condition was simulated 100 times, and the boxplots represent the distribution of those measures (i.e., 25th quartile, median, 75th quartile). Please view this figure in color for optimal interpretation

with the true network. The sensitivity of the latent variable approach was strongly influenced by both the number of timepoints and measurement error. Unexpectedly, in conditions with high measurement error (error variance = 1.5 and 2), the latent variable model shows higher sensitivity in conditions with low numbers of timepoints compared to high numbers of timepoints. For all other measurement approaches, and for the conditions with small measurement error, this is not the case: sensitivity increases with increasing numbers of timepoints. The pattern in precision and specificity was similar for the latent variable approach: it was low in conditions with high measurement error and small numbers of timepoints and increased as the number of timepoints increased. However, inconditions where precision and specificity are low (high measurement error and small number of timepoints), the latent variable approach performs worse than other approaches.

Plausible Value Score Approach

The plausible value score approach also showed very low correlations in the smallest number of observations with high measurement error, but like the latent variable approach, performance improved with increasing numbers of observations and lower measurement error. Sensitivity remained low across conditions and was consistently below the level of other approaches. However, specificity was optimal except for the smallest number of observations ($t_{timepoints} = 100$). Precision was initially very low but increased to optimal in larger numbers of observations.

Factor Score Approach

The factor score approach yielded correlations with the true network ranging from moderate to high. Along with the average score approach, it showed the highest correlations among all approaches. Sensitivity remained moderate even in the largest sample size and highest measurement error,

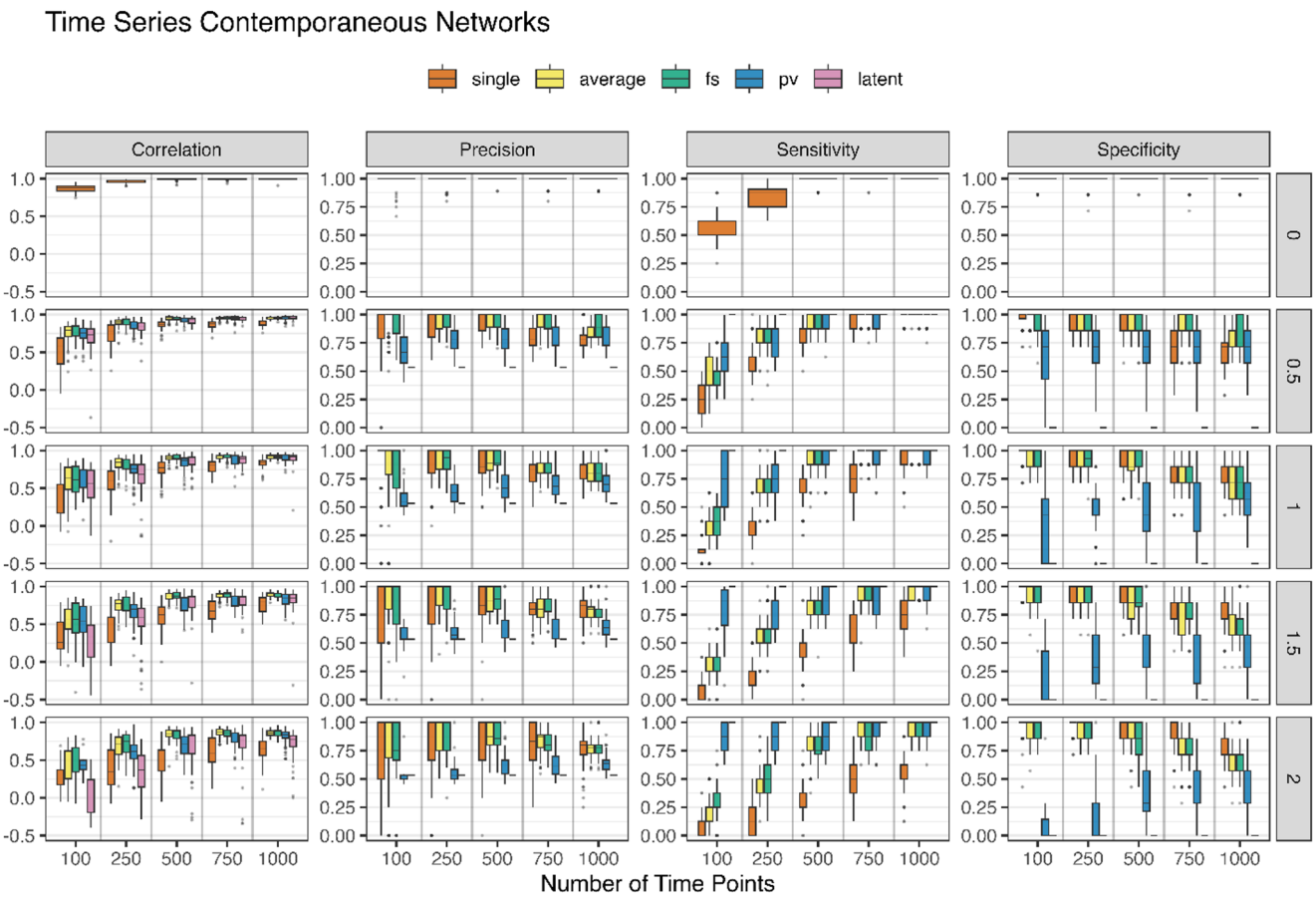

Fig. 2 Simulation results for the performance of the single- and multiple-indicator approaches for contemporaneous N = 1 network models. The vertical panels indicate the different performance measures: correlation, precision, sensitivity, and specificity. Horizontal panels indicate the amount of measurement error. The dots indicate outliers. Every condition was simulated 100 times, and the boxplots represent the distribution of those measures (i.e., 25th quartile, median, 75th quartile). Please view this figure in color for optimal interpretation

improving from very low levels as the number of observations grew. Still, the factor score approach together with the average score approach outperforms other approaches in terms of sensitivity as the number of timepoints increases. Specificity was consistently optimal regardless of the number of observations or variance of the measurement errors. Consequently, precision ranged from moderate to high across numbers of observations.

Average Score Approach

The average score approach demonstrated a nearly identical pattern to the factor score approach. Minor differences were observed in correlations and sensitivity, likely reflecting random variation rather than systematic effects. Together with the factor score approach, the average score approach outperforms other approaches in most conditions.

Single Indicator Approach

In the conditions without measurement error, the single-indicator approach correlated highly with the true structure and had high specificity and precision. Sensitivity is low in conditions with small numbers of timepoints but increases as the number of timepoints increases. For conditions with measurement error, it showed slightly lower correlations overall but outperformed the latent variable and the plausible value score approaches at the lowest number of observations, where those showed particularly poor correlations. Sensitivity declined considerably in the presence of measurement error, indicating difficulty in edge detection. Under low measurement error, sensitivity reached approximately 0.6 but deteriorated as error increased. Specificity remained optimal throughout. Precision displayed a distinctive double-peaked pattern, with values clustering at either 0 or 1 at the smallest number of observations, while precision remained relatively high for larger numbers of observations.

Contemporaneous Networks

Figure 2 shows the performance of the different measurement approaches in contemporaneous networks. Overall, correlations, sensitivity, and precision were higher in contemporaneous networks than in temporal networks. Correlations and sensitivity ranged from low to optimal and were influenced by both the number of observations and measurement error variance.

Latent Variable Approach

Correlations between the estimated latent variable networks and the true network ranged from low to optimal, depending on number of observations and measurement error variance. At the smallest number of observations, it yielded the lowest correlations among all approaches but reached comparable, optimal levels as the number of observations increased. With lower measurement error variance, its correlation levels were similar to those of other multiple-indicator approaches; however, under higher measurement error, achieving parity took longer. Sensitivity was consistently optimal across all conditions, while specificity remained zero, suggesting that all networks were estimated as fully connected. Consequently, precision was only moderate.

Plausible Value Score Approach

The plausible value score approach exhibited high correlations with the true network, though these declined with increasing measurement error. Sensitivity ranged from moderate to high and was less affected by measurement error than the factor score, average score, and single-indicator approaches—but more so than the latent variable approach. Specificity was moderate, reflecting some ability to identify absent edges, though it declined under higher measurement error. Precision was higher than in the latent variable approach but remained clearly below that of the other approaches.

Factor Score Approach

The factor score approach showed moderate to high correlations with the true network, and sensitivity ranged from low to optimal, depending on sample size. However, all performance measures were negatively impacted by measurement errors. Notably, specificity and precision decreased with increasing sample size under high measurement error—an unexpected pattern suggesting over-identification of edges in larger samples.

Average Score Approach

As in the temporal networks, the average score approach performed almost identically to the factor score approach. No systematic differences were observed between the two. Like the factor score approach, it struggled to detect edges in the conditions with smaller numbers of observations, as indicated by lower sensitivity.

Single Indicator Approach

The single-indicator approach yielded somewhat lower correlations with the true network compared to the other approaches, except for the latent variable approach in conditions with small number of observations. Sensitivity was low at smaller numbers of observations, indicating difficulty in edge estimation. Even in the conditions with the largest number of observations, it failed to identify all edges, and sensitivity remained below that of the other approaches. However, specificity was high for smaller numbers of observations, and, although it declined in the conditions with the largest number of observations, deterioration was less pronounced than in the other approaches.

Summary of Simulation Study 1

The Effect of Measurement Error

In temporal networks, measurement error impacted the correlation, sensitivity, and precision, while specificity remained largely unaffected. Under conditions without measurement error, overall performance was high with the exception of low sensitivity in smaller samples. As measurement error increased, sensitivity was most affected—all approaches struggled to detect edges. Correlations declined with increasing measurement error, particularly for the plausible value score and single-indicator approaches, whereas the factor score and average score approaches were less affected.

In contemporaneous networks, measurement error influenced correlations across all approaches, with sensitivity being especially reduced in the single-indicator approach and specificity for the plausible value score approach. However, precision was relatively stable. When no measurement error was present, the single-indicator approach demonstrated near-optimal performance. As measurement errors increased, overall performance declined. The clearest effects of measurement error were seen in reduced correlation and sensitivity, indicating a diminished capacity for edge detection and a tendency to produce networks sparser than the true structure. Interestingly, specificity in

the plausible value score approach declined markedly with increased measurement error.

The Effect of Number of Observations

The number of observations influenced correlations and sensitivity in both network types, while having little effect on precision and no observable impact on specificity. In temporal networks, the latent variable, plausible value score and single-indicator approaches showed clear improvements in performance with increasing numbers of observations, whereas the factor score and average score approaches were more robust and less affected by the number of observations. In contemporaneous networks, the correlation improved with larger numbers of observations, particularly in the latent variable and single-indicator approaches. In terms of sensitivity, the single-indicator approach was again the most clearly affected by sample size.

Simulation Study 2: Panel data (N > 1)

The Supplementary Materials S1 present the data-generating networks estimated from the panel dataset. Figures 21 and 22 in Supplementary Materials S4 show the results for the bias. For three waves of assessment, out of the five evaluated measurement approaches, the latent variable approach showed the most bias, which was primarily evident in the contemporaneous and the between-subjects network. However, in all cases, bias was minimal. For a more detailed discussion on the bias, see Supplementary Materials S4.

Figures 3, 4, and 5 present the performance measures of the different measurement approaches in within-subjects temporal, contemporaneous, and between-subject networks respectively. Similar to Study 1, we only include the performance measures for the single-indicator approach in the condition of no measurement error. We here only discuss the results for the networks based on three waves of assessment. The results for networks based on six waves of assessment are provided in Supplementary Materials S3.

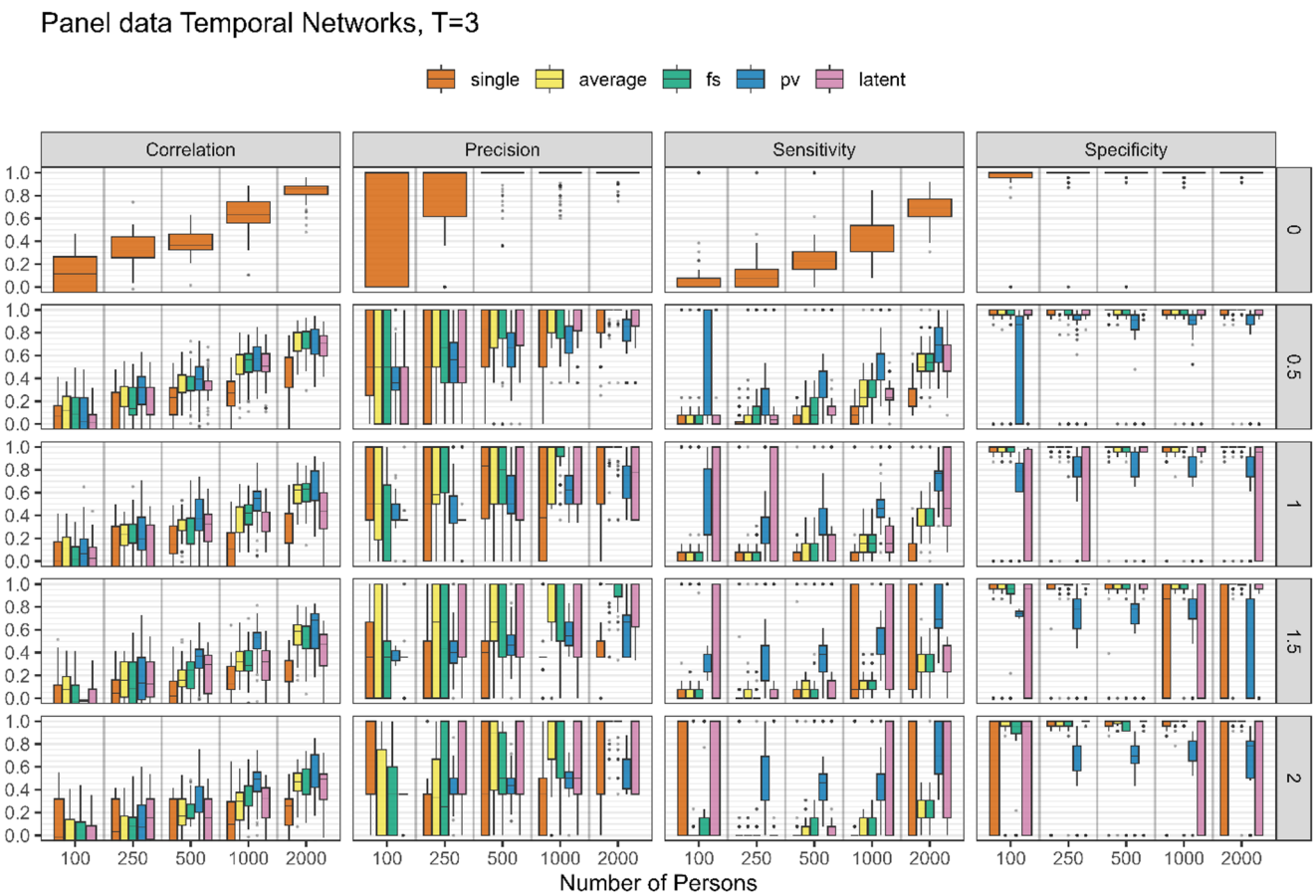

Fig. 3 Simulation results for the performance of the single- and multiple-indicator approaches for temporal network model based on panel data of three waves. The vertical panels indicate the different performance measures: correlation, precision, sensitivity, and specificity. Horizontal panels indicate the amount of measurement error. The dots indicate outliers. Every condition was simulated 100 times, and the boxplots represent the distribution of those measures (i.e., 25th quartile, median, 75th quartile). Please view this figure in color for optimal interpretation

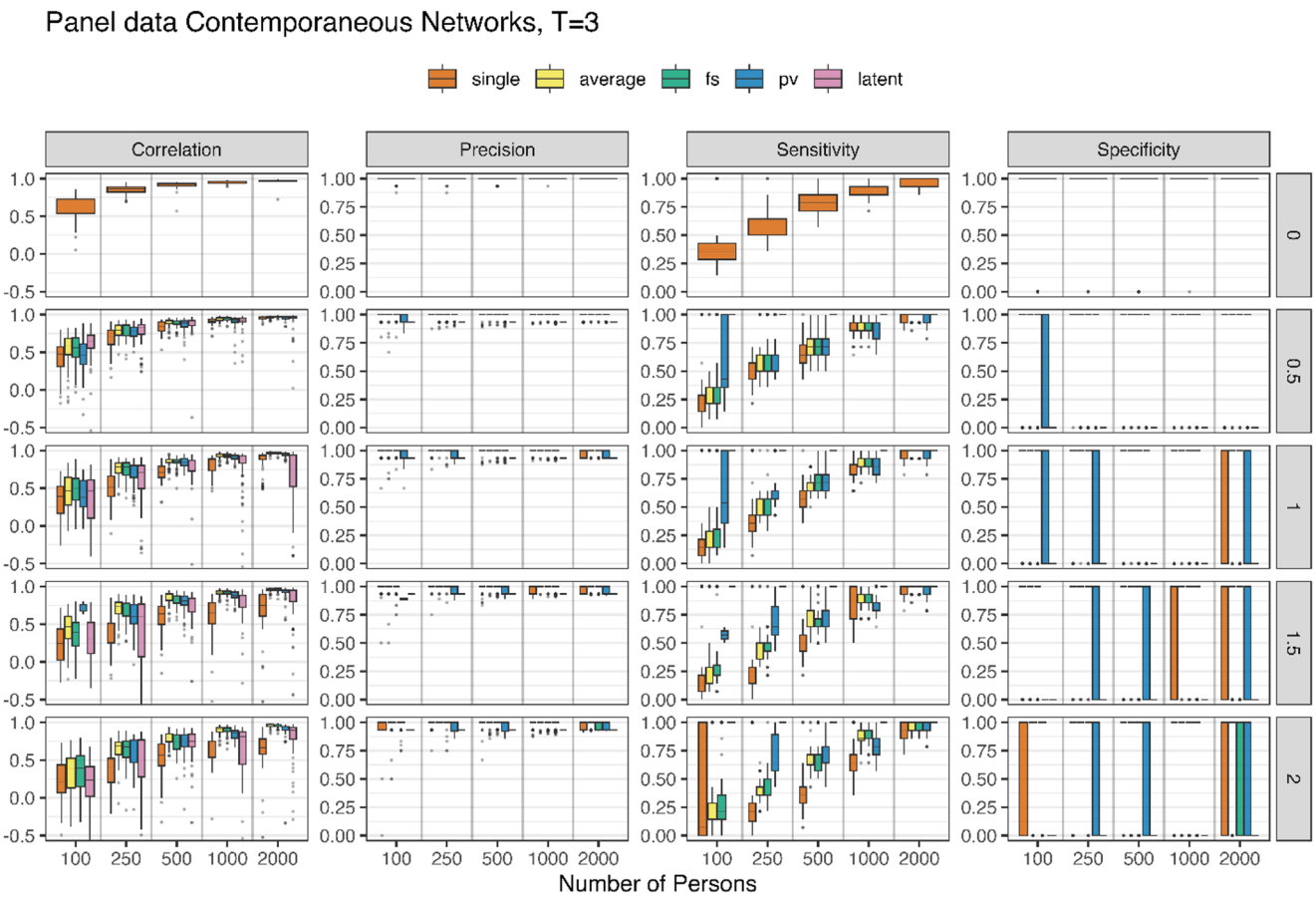

Fig. 4 Simulation results for the performance of the single- and multiple-indicator approaches for contemporaneous network model based on panel data of three waves. The vertical panels indicate the different performance measures: correlation, precision, sensitivity, and specificity. Horizontal panels indicate the amount of measurement error. The dots indicate outliers. Every condition was simulated 100 times, and the boxplots represent the distribution of those measures (i.e., 25th quartile, median, 75th quartile). Please view this figure in color for optimal interpretation

Temporal Networks

Figure 3 shows how different measurement approaches performed across conditions of different sample sizes and degrees of measurement error in temporal networks. In general, the different approaches exhibited comparable performance patterns. Most performance measures improved with increasing sample size and declined with increasing measurement error. Correlations were generally similar across the multiple-indicator approaches, whereas the single indicator approach consistently showed slightly lower correlations. Similar to Study 1, the conditions without measurement error gives some comparison to disentangle to what extent poor performance is due to measurement error and how particular measurement methods deal with that, and to what extent poor performance is also just because network estimation more generally performs poorly in those conditions, ignoring measurement error. Figure 3 shows that for the temporal network, correlations between the estimated and true network edges are low in conditions with small sample sizes, also in the case of absence of measurement error.

Latent Variable Approach

The latent variable approach produced very low correlations at the smallest sample size, with performance improving steadily as the sample size increased. Conversely, greater measurement error led to declining correlations. The highest mean correlation occurred under the lowest measurement error condition. Sensitivity was minimal in the smallest sample size but improved with larger samples. Under high measurement error, the sensitivity distribution followed a bimodal pattern. A similar bimodal pattern emerged in specificity, with high specificity only observed under the lowest measurement error. Precision was high when sample size was large and measurement error was low. However, as measurement error increased, the distribution of precision values widened. In small samples with high measurement error, the latent variable approach frequently yielded

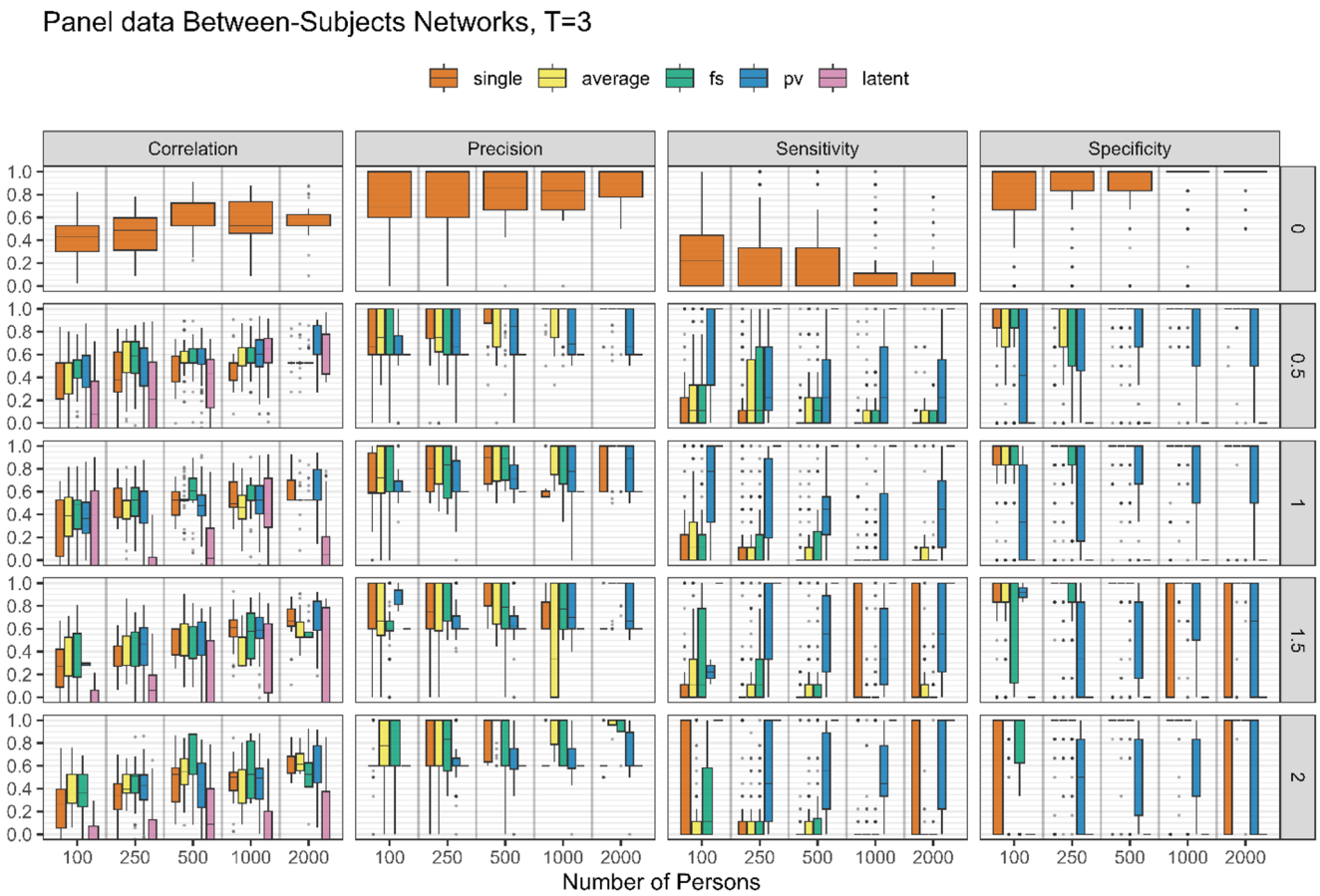

Fig. 5 Simulation results for the performance of the single- and multiple-indicator approaches for the between-subjects network model based on panel data of three waves. The vertical panels indicate the different performance measures: correlation, precision, sensitivity, and specificity. Horizontal panels indicate the amount of measurement error. The dots indicate outliers. Every condition was simulated 100 times, and the boxplots represent the distribution of those measures (i.e., 25th quartile, median, 75th quartile). Please view this figure in color for optimal interpretation

fully connected networks, resulting in a narrow precision centered at 0.36.

Plausible Value Score Approach

Among all approaches, the plausible value score consistently demonstrated the highest correlations, which increased with sample size and declined moderately with measurement error. Under the smallest measurement error and largest sample size, the mean correlation approached 0.8 and remained relatively strong (>0.6) even under substantial measurement error. Sensitivity was also highest for this approach, outperforming others in many conditions. However, convergence issues occurred in the most extreme case—smallest sample size combined with highest measurement error—resulting in missing values. Notably, sensitivity remained largely unaffected by increasing measurement error, a pattern not observed in other approaches. Specificity was generally high and optimal in most conditions with low measurement error. In more error-prone conditions, specificity remained relatively stable, showing more concentrated boxplots than the latent variable approach. Precision was high in the lowest measurement error condition and hovered around 0.5 under higher error levels.

Factor Score Approach

The factor score approach yielded low to moderate correlations with the true network and exhibited very low sensitivity across conditions, with modest improvement at the highest sample sizes. Specificity, however, was consistently optimal, suggesting that the approach tended to estimate sparse networks. Precision showed signs of bimodal distribution at smaller sample sizes but improved as sample size increased.

Average Score Approach

The average score approach closely mirrored the factor score approach in all performance measures. It demonstrated low

sensitivity, high specificity, and moderate to high precision across conditions.

Single Indicator Approach

Performance of the single indicator approach lagged behind the multiple-indicator approaches. Even in the absence of measurement error, its correlations and sensitivity were highly dependent on the sample size, showing very low performance for smaller samples. Specificity was high across sample sizes, but precision was somewhat reduced at the smallest sample size. In other measurement error conditions, sensitivity remained very low throughout. Sensitivity, specificity, and precision displayed a double-peaked distribution in both the smallest and largest sample sizes. Precision was highly sensitive to measurement error: while it was relatively high under minimal error, it declined sharply as measurement error increased.

Contemporaneous Networks

Figure 4 shows the performance of the different measurement approaches in contemporaneous networks. In general, all measurement approaches performed better for contemporaneous networks than for temporal ones. As with temporal networks, performance depended both on sample size and measurement error: larger samples enhanced performance, while increased error impaired it. Correlations were particularly sensitive to measurement error.

Latent Variable Approach

The latent variable approach exhibited strong overall performance, though correlations suffered in smaller sample sizes and under greater measurement error. Compared to other multiple-indicator approaches, correlations were somewhat lower yet still surpassing those of the single-indicator approach. Sensitivity was consistently optimal, while specificity remained at zero—indicating that all networks were estimated as fully connected. Nevertheless, given that the true network contained only a single absent edge, the approach still achieved high precision despite consistently misclassifying that edge.

Plausible Value Score Approach

The plausible value score approach again encountered convergence issues in the most challenging condition—smallest sample size combined with highest measurement error—evidenced by missing values. In all other conditions, performance improved with larger sample sizes: both correlation and sensitivity increased, while specificity and precision remained optimal. These results indicate that the plausible value score approach was effective in correctly identifying the sole absent edge in the network.

Factor Score Approach

The factor score approach yielded similar correlations to both the latent variable and plausible value score approaches, but showed reduced sensitivity, particularly in smaller sample sizes—resulting in networks that were sparser than the true structure. Despite this, specificity and precision remained consistently optimal across conditions.

Average Score Approach

The average score approach showed nearly identical performance patterns to the factor score approach. Sensitivity was low in the smallest sample sizes, while specificity and precision were optimal. Its performance clearly decreased with smaller sample sizes, and was less susceptible to measurement error, suggesting that it estimated sparser-than-true networks in conditions with limited data.

Single Indicator Approach

Under the no measurement error condition, the single-indicator approach had high performance, except for low sensitivity in smaller samples. When measurement error was present, it exhibited lower correlations and sensitivity compared to the multiple-indicator approaches but maintained high specificity and precision. In specificity, it displayed a characteristic bimodal distribution in both the smallest and largest sample sizes, and in sensitivity in the smallest sample.

Between-Subjects Networks

Figure 5 presents the performance of the different measurement approaches in between-subjects networks. Performance in between-subjects networks exhibited greater variability than in contemporaneous networks. Correlations ranged from low to moderate across most conditions, with the exception of the no measurement error scenario, where the single-indicator approach achieved moderate to high correlations. Sensitivity was generally low to moderate, whereas specificity and precision remained relatively high across approaches.

Latent Variable Approach

Unexpectedly, the latent variable approach produced the lowest correlations of all approaches—including the

single-indicator approach—in particular under conditions of small sample size and high measurement error. The latent variable approach displayed optimal sensitivity and zero specificity, indicating that it consistently estimated fully connected networks. Given the substantial number of both present and absent edges in the true network, this led to only moderate levels of precision.

Plausible Value Score Approach

The plausible value score approach again failed to converge under the most challenging condition: smallest sample size and highest measurement error. Across other conditions, it achieved moderate correlations. Somewhat counterintuitively, its sensitivity declined while specificity slightly improved as sample size increased. Although it maintained reasonable overall accuracy, its precision was somewhat lower than that of the factor score, average score, and single-indicator approaches.

Factor Score Approach

The factor score approach mirrored the counterintuitive pattern observed in the plausible value score: increasing sample size led to decreased sensitivity and improved specificity. In the largest sample size, sensitivity dropped to zero while specificity reached optimal levels—suggesting that estimated networks became increasingly sparse. However, precision improved with sample size and reached its peak in the largest condition, indicating that although fewer edges were detected, the edges that were identified tended to be correct.

Average Score Approach

Once again, the average score approach paralleled the behavior of the factor score approach across all performance metrics. It produced low to moderate correlations, exhibited minimal sensitivity—often approaching zero—and maintained high to optimal specificity and precision. These results suggest a conservative estimation pattern, with few false positives but a tendency to overlook true edges.

Single Indicator Approach

In the no measurement error condition, the single-indicator approach had moderate to high correlations and high specificity and precision, but its sensitivity was low in all sample sizes. When measurement error was present, it outperformed the latent variable approach in all measures except sensitivity. Relative to the other multiple-indicator approaches, it lagged slightly. Correlations slightly increased with sample size. Specificity exhibited optimal values in low measurement error and in moderate sample sizes even in high amount of error, but sensitivity and specificity showed a bimodal distribution when the sample was the smallest or largest, and error was highest. Precision increased from moderate to high, with increasing sample sizes and decreasing measurement error variance.

Summary of Simulation Study 2

The Effect of Measurement Error

Measurement error had a relatively modest impact on the performance of measurement approaches in temporal networks. Although increasing measurement error led to some decline in performance, the differences were generally limited. Among all methods, the single-indicator approach was most sensitive to measurement error. The effect of measurement error varied by measure: in larger sample sizes, sensitivity declined especially in the factor score and average score approaches, whereas the plausible value score approach exhibited the clearest drop in precision.

In contemporaneous networks, measurement error had minimal influence. Across all levels, performance remained relatively stable, with the exception of specificity, which displayed a bimodal distribution in the single-indicator and plausible value score approaches as error increased—suggesting greater variability in detecting absent edges. Similarly, between-subjects networks were largely unaffected by measurement error. Only a slight decline in sensitivity was observed for the factor score and average score approaches at higher error levels.

The Effect of Sample Size

Sample size impacted all network types. Smaller samples were associated with weaker performance and reduced edge detection capacity. In contrast, specificity remained high across all sample sizes for the factor score and average score approaches, suggesting that these approaches tended to estimate sparse networks regardless of sample size. In contemporaneous and between-subjects networks, sample size had minimal effect on specificity and precision. This indicates that while smaller samples hampered the ability to detect present edges, they did not meaningfully increase false positives or reduce the accuracy of detected edges in these network types.

Differences Between the Wave 3 and the Wave 6

The results for the wave 6 data are included in Supplementary Materials S3. We do not discuss these results in depth here but present a quick comparison with the wave 3 data. In many conditions the results for wave 3 and wave 6 data are similar, but there are conditions in which the wave 3 data show better performance and there are conditions in which the wave 6 data show better performance. However, bias increased compared to the three waves, see Supplementary Materials S4. Contrary to the three waves, in the six waves condition, the single-indicator approach showed relatively high performance compared to the multiple-indicator approaches. Among the multiple-indicator approaches, the average score approach fared equally to the factor score, except for showing better performance in the smallest sample size, where the factor score failed to estimate.

Discussion

This study compared the performance of different measurement approaches—latent variable, plausible value score, factor score, average score, and single-indicator—in recovering the true network structure when node scores include measurement error, in the context of dynamic network models. In particular, we focused on the GVAR model and the panelGVAR model. To evaluate the performance of each of the measurement approaches, we considered the correlation between the edge weights in the estimated network with those in the true network, sensitivity, specificity, and precision, in conditions that varied the numbers of observations and degrees of measurement error.

Our primary aim was to assess how well each measurement approach performed under these varying conditions. The best-performing measurement approach was different for the time series data (N = 1, Study 1) than for the panel data (N > 1, Study 2). In Study 1 (N = 1), the factor score and average score approaches performed best for recovering the temporal networks, while the latent variable approach performed best in recovering contemporaneous networks, particularly in terms of sensitivity. The observation that multiple-indicator approaches outperform the single-indicator approach in the presence of measurement error, specifically in their sensitivity, is consistent with previous cross-sectional studies that have shown multiple-indicator approaches to improve network sensitivity (De Ron et al., 2022; Herrera-Bennett & Rhemtulla, 2021).

In our study, the factor score and average score approaches performed similarly even though we varied the factor loadings. We expected the factor score to outperform the average when factor loadings are not equal because the factor score is a weighted average that accounts for differences in factor loading. However, the finding that the two approaches perform similarly is consistent with previous studies (O'Laughlin et al., 2021; Oh et al., 2025). Still, future research could consider more conditions under which these two approaches would differ in performance. For example, we randomly sampled loadings for the indicators from a uniform distribution between 0.75 and 1, which was already more variability in factor loadings than was used by O'Lauhglin et al. (2021) and Oh et al. (2025). But it is possible that when loadings across indicators of the same node differ even more from each other, the two approaches will start to diverge in performance because the factor score approach—unlike the average score—could weight indicators differently, potentially leading to different results.

The latent variable approach tended to produce false positives in contemporaneous networks. Stricter pruning could possibly mitigate this issue—albeit at the cost of general sensitivity, particularly in temporal networks. Future research could consider how different pruning thresholds (e.g., stricter) could improve the performance of some of the approaches.

In Study 2 (N > 1), plausible value scores and the latent variable approach performed better than other approaches for many conditions. However, in the between-subjects network, the latent variable approach showed very low correlations. In temporal networks, plausible value scores performed more consistently across conditions—except for the smallest sample size (N = 100)—compared to the latent variable method, which frequently exhibited bimodal performance—estimating either fully connected or empty networks. In addition, the plausible value scores demonstrated higher correlations across conditions than the latent variable approach. Because plausible value scores are less known than factor scores (Marsman et al., 2016; Thomson, 1938; Thurstone, 1935; Von Davier et al., 2009), these results will hopefully encourage researchers to consider plausible value scoring as an alternative to other multiple-indicator approaches. The bimodality observed in both plausible value and single-indicator approaches in contemporaneous networks was likely an artifact of the true structure of the contemporaneous network, which was missing only a single edge. In this specific case, the bimodality in specificity reflects whether the approach accurately detected the absence of that single edge. In future research, it would therefore be good to assess how results generalize to less dense networks.

However, the plausible value approach also exhibited relatively high variability in performance. As one of our anonymous reviewers kindly pointed out, one might prefer an approach that performs slightly worse if it is more consistent. The variability in the performance of more complex

approaches has also been found in previous studies, known as the variability-accuracy trade-off, or the bias-variance trade-off (Ledgerwood & Shrout, 2011; O'Laughlin et al., 2021; Oh et al., 2025).

Our second aim was to explore how performance varied with different degrees of measurement error. All approaches showed decreased correlations and sensitivity with increasing measurement error, while specificity and precision remained largely unaffected, except for a decline in specificity in the plausible value score approach in Study 1(N = 1) in the contemporaneous networks. In Study 2 (N > 1), the single-indicator approach was particularly sensitive to measurement error, with performance deteriorating rapidly. Unexpectedly, the latent variable approach also showed some sensitivity to measurement error in both studies. The negative effect of measurement error on sensitivity has been observed in cross-sectional network studies as well (De Ron et al., 2022; Herrera-Bennett & Rhemtulla, 2021). However, unlike Herrera-Bennett and Rhemtulla (2021), we found that the number of observations had a greater impact than the degree of measurement error.

In the context of longitudinal studies, our findings are in line with Schuurman and Hamaker (2019), who also found that measurement error had a negative impact on network estimation in models that do not account for measurement error. They also proposed a single indicator method to account for measurement error in AR models. It would therefore be an important direction for future research to see whether their method could be applied to the designs we considered in this study and compare their approach to the multiple-indicator approaches we have examined. Given that multiple-indicator approaches place a significantly greater burden on participants, a single-indicator approach that could account for measurement error as effectively (or better) than a multiple-indicator approach would be a welcome addition to the network modeling toolbox.

For our third aim—comparing network types (temporal, contemporaneous, and between-subject) —we conclude that temporal networks were most affected by both measurement error and the number of observations. The finding of temporal networks being the most demanding in terms of the number of observations to estimate is typical for dynamic network estimation and consistent with previous studies (Epskamp, 2020a; Hoekstra et al., 2024; Mansueto et al., 2023). When measurement error was low and the number of observations high, all approaches performed well in estimating temporal networks. However, under high measurement error in Study 1, only the latent variable and single-indicator approaches retained high correlations in conditions with the largest number of timepoints.

As expected, the number of observations had a substantial effect on network estimation across approaches and network types. In both Study 1 and Study 2, larger numbers of observations led to improved correlations and sensitivity. In the time series data (N = 1), the factor score and average score approaches were less affected by smaller numbers of observations compared to the other approaches. In contrast, in the single indicator approach, sensitivity was most negatively affected by smaller numbers of observations, aligning with prior findings that small samples impair sensitivity (De Ron et al., 2022; Epskamp, 2020a; Hoekstra et al., 2023; Mansueto et al., 2023).

This sensitivity to the number of observations is concerning given the common use of small samples in clinical psychology. In clinical research aiming at personalized treatment and using idiographic time series data, sample sizes often include around 100 time points per participant due to practical constraints, such as participant burden (McLean et al., 2017; Ono et al., 2019; Rintala et al., 2019; Wen et al., 2017), the construct's natural timescale (Wilhelm & Schoebi, 2007), and the risk of nonstationarity with longer data collection periods (Epskamp et al., 2018a).

In contrast, in panel designs, the focus shifts to the number of participants rather than the number of time points. Because individual burden is lower, participant numbers tend to be higher, typically ranging from a hundred to several thousand (Martín-Gómez et al., 2022; Rigabert et al., 2020). Still, median sample sizes reported in some of the reviews were only N = 106, and N = 222, so the impact of sample size may be significant also in studies using panel data. Some exceptions with higher numbers of observations exist in both time series and panel designs (McBride et al., 2021; Wichers & Groot, 2016).

The minimum required sample size to use multiple-indicator approaches differed across approaches. For the factor score and average score approaches, the minimum number of observations appeared to be the lowest in both Study 1 and 2, reflecting their relative efficiency even with small samples, as well as their overall superior performance compared to the single-indicator approach. For the latent variable and plausible value approaches, performance deteriorated at smaller sample sizes, indicating that these methods are less suitable when the number of observations is low. Their performance improved substantially when the number of observations was 500 or higher, suggesting a minimal number of observations around 500. In Study 2, the plausible value approach failed to converge in the smallest sample size condition, which underscores the need for caution when applying this approach with smaller samples. But, overall, given the complexity of network model estimation and the known challenges in their accuracy and replicability (Borsboom et al., 2017; Epskamp, 2020a; Hoekstra et al., 2023, 2024; Mansueto et al., 2023), such complex dynamic

network models with multiple indicators as nodes appear to perform relatively well.

Recommendations for Applied Researchers

Our results suggest that multiple-indicator approaches outperform the single-indicator approach in both Study 1 and Study 2, across most network types. However, the optimal multiple-indicator approach varies depending on the study and the specific network. For time series (N = 1) analysis focusing on temporal relationships, factor scores or average scores may be preferred. Factor scores offer the added benefit that the underlying factor model allows researchers to evaluate model fit and thus strengthen the adequacy of the measurement structure before network estimation. In contrast, in panel designs with only three assessment points, plausible value scoring may offer the most consistent results.

When the focus shifts to contemporaneous associations—such as in time series data where the goal is to study instantaneous relations among variables, or in panel data, where assessment intervals are too wide to capture rapid variation—the latent variable approach and plausible value approach perform well. But, since the latent variable approach gives very low (in some cases even negative) correlations for the between-subjects network, we would recommend choosing the plausible value approach over the latent variable approach. In situations where reliability is crucial, particularly in longitudinal studies, investing in additional indicators (rather than relying on a single one) could be advantageous to improve both reliability and validity.

For research prioritizing simplicity and shorter questionnaires, the single-indicator approach may be appropriate, especially when the focus is on contemporaneous networks, or when specificity and precision are more crucial than sensitivity. The single-indicator approach may also be suitable when the number of observations is very low (e.g., T = 100 or N = 100), as performance across all approaches tends to decline in such cases. However, it should also be noted that network estimation is not recommended with very low numbers of observations (Hoekstra et al., 2023; Mansueto et al., 2023). Given vulnerability of the single-indicator approach to measurement error, it is most competitive when measurement error is known to be minimal.

To our knowledge, this is the first study to evaluate the performance of multiple measurement approaches in estimating temporal, contemporaneous, and between-subjects networks across both time series (N = 1) and panel data (N > 1). Thus, this study addresses a critical gap in the literature, contributing to the understanding of how measurement error impacts dynamic networks and offering guidance on the most effective measurement approaches for mitigating such error. However, like all simulation studies, this work is limited by the necessity of selecting a finite set of conditions. We aimed to choose the most relevant conditions to help applied researchers identify the best approaches for their specific studies. Also, we were only able to use 100 replications, although the number of replications is recommended to be higher (Mundform et al., 2011). Future studies could consider increasing the number of replications if possible.

Conclusion

This study compared the performance of different measurement approaches—latent variable, plausible value score, factor score, average score, and single-indicator—for estimating dynamic network models in the presence of measurement error. Overall, multiple-indicator approaches generally outperformed the single-indicator approach, particularly in the presence of higher measurement error or smaller numbers of observations. Among them, the plausible value score approach demonstrated the most robust and stable performance across metrics, particularly in larger samples with low measurement error. In contrast, the latent variable approach exhibited extreme sensitivity and specificity, often resulting in fully connected or empty networks. Factor score and average score approaches often yielded sparse networks with high specificity but low sensitivity and showed highest performance for estimating time series temporal networks. To conclude, our study suggests that multiple-indicator approaches may substantially improve the estimation of dynamic networks, while emphasizing that the optimal multiple-indicator strategy depends on the research question.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10608-026-10719-0.

Author Contributions R.K. played a lead role and R.v.B. played a supporting role in conceptualization of the study. R.K. wrote the main manuscript text, analyzed the results, and prepared figures. R.v.B. played a leading role in supervision. R.K., J.d.R., and R.H.A.H. wrote the simulation code. R.H.A.H. ran the code in an external server using R. All authors have extensively taken part to reviewing and editing the manuscript.

Funding Open Access funding provided by University of Turku (including Turku University Central Hospital). This research was supported by the Research Council of Finland, decision number 345546, Jenny and Antti Wihuri Foundation, The Alfred Kordelin Foundation, and the Mannerheim League for Child Welfare Foundation granted for R.K., R.v.B., J.d.R., or R.H.A.H. did not receive funding for this study.

Data Availability The empirical datasets we used can be freely accessed online (time series dataset for Study 1 by Kossakowski et al.,

2017, and panel dataset for Study 2 by McBride et al., 2021).

Declarations

Competing interests The authors declare no competing interests.

References

Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of Psychology. General Section, 28*(1), 97–104. https://doi.org/10.1111/j.2044-8295.1937.tb00863.x

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry, 16*(1), 5–13. https://doi.org/10.1002/wps.20375

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

Briganti, G., Scutari, M., Epskamp, S., Borsboom, D., Hoekstra, R. H. A., Golino, H. F., Christensen, A. P., Morvan, Y., Ebrahimi, O. V., Costantini, G., Heeren, A., Ron, J. D., Bringmann, L. F., Huth, K., Haslbeck, J. M. B., Isvoranu, A., Marsman, M., Blanken, T., Gilbert, A., … McNally, R. J. (2024). Network analysis: An overview for mental health research. *International Journal of Methods in Psychiatric Research, 33*(4), Article e2034. https://doi.org/10.1002/mpr.2034

Bringmann, L. F. (2021). Person-specific networks in psychopathology: Past, present, and future. *Current Opinion in Psychology, 41*, 59–64. https://doi.org/10.1016/j.copsyc.2021.03.004

Castro, D., Ferreira, F., De Castro, I., Rodrigues, A. R., Correia, M., Ribeiro, J., & Ferreira, T. B. (2019). The differential role of central and bridge symptoms in deactivating psychopathological networks. *Frontiers in Psychology, 10*, Article 2448. https://doi.org/10.3389/fpsyg.2019.02448

Castro-Alvarez, S., Tendeiro, J. N., De Jonge, P., Meijer, R. R., & Bringmann, L. F. (2022). Mixed-effects trait-state-occasion model: Studying the psychometric properties and the person–situation interactions of psychological dynamics. *Structural Equation Modeling: A Multidisciplinary Journal, 29*(3), 438–451. https://doi.org/10.1080/10705511.2021.1961587

Castro-Alvarez, S., Bringmann, L. F., Back, J., & Liu, S. (2024a). *The Many Reliabilities of Psychological Dynamics: An Overview of Statistical Approaches to Estimate the Internal Consistency Reliability of Intensive Longitudinal Data* [Preprint]. https://doi.org/10.31234/osf.io/qyk2r

Castro-Alvarez, S., Zhou, D., Bringmann, L., Tutunji, R., Proppert, R. K. K., Rieble, C., Fried, E. I., & Liu, S. (2024b). *Assessing the Internal Consistency Reliability of Ecological Momentary Assessment Measures: Insights from the WARN-D Study* [Preprint]. https://doi.org/10.31234/osf.io/nrzsc

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods, 19*(2), 300–315.

Cramer, A. O. J., Waldorp, L. J., Van Der Maas, H. L. J., & Borsboom, D. (2010). Comorbidity: A network perspective. *Behavioral and Brain Sciences, 33*(2–3), 137–150. https://doi.org/10.1017/S0140525X09991567

Croon, M. (2002). Using predicted latent scores in general latent structure models. In *Marcoulides, G., & Moustaki, I. (Eds.), Latent variable and latent structure modeling* (pp. 195–223). Lawrence Erlbaum.

De Ron, J., Robinaugh, D. J., Fried, E. I., Pedrelli, P., Jain, F. A., Mischoulon, D., & Epskamp, S. (2022). Quantifying and addressing the impact of measurement error in network models. *Behaviour Research and Therapy, 157*, Article 104163. https://doi.org/10.1016/j.brat.2022.104163

Dejonckheere, E., Demeyer, F., Geusens, B., Piot, M., Tuerlinckx, F., Verdonck, S., & Mestdagh, M. (2022). Assessing the reliability of single-item momentary affective measurements in experience sampling. *Psychological Assessment, 34*(12), 1138–1154. https://doi.org/10.1037/pas0001178

Devlieger, I., Talloen, W., & Rosseel, Y. (2019). New developments in factor score regression: Fit indices and a model comparison test. *Educational and Psychological Measurement, 79*(6), 1017–1037. https://doi.org/10.1177/0013164419844552

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment, 29*(2), 136–151. https://doi.org/10.1177/1073191120957102

Epskamp, S. (2020a). Psychometric network models from time series and panel data. *Psychometrika, 85*(1), 206–231. https://doi.org/10.1007/s11336-020-09697-3

Epskamp, S. (2020b). *Psychonetrics: Structural Equation Modeling and Confirmatory Network Analysis.* (Version 0.7.1.) [R]. https://cran.r-project.org/web/packages/psychonetrics/index.html

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent variable models. *Psychometrika, 82*(4), 904–927. https://doi.org/10.1007/s11336-017-9557-x

Epskamp, S., Van Borkulo, C. D., Van Der Veen, D. C., Servaas, M. N., Isvoranu, A.-M., Riese, H., & Cramer, A. O. J. (2018a). Personalized network modeling in psychopathology: The importance of contemporaneous and temporal connections. *Clinical Psychological Science, 6*(3), 416–427. https://doi.org/10.1177/2167702617744325

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018b). The Gaussian graphical model in cross-sectional and time series data. *Multivariate Behavioral Research, 53*(4), 453–480. https://doi.org/10.1080/00273171.2018.1454823

Freichel, R., Pfirrmann, J., Cousijn, J., De Jong, P., Franken, I., Banaschewski, T., Bokde, A. L. W., Desrivières, S., Flor, H., Grigis, A., Garavan, H., Heinz, A., Martinot, J., Martinot, M. P., Artiges, E., Nees, F., Orfanos, D. P., Poustka, L., Hohmann, S., … IMAGEN Consortium. (2023). Drinking motives, personality traits and life stressors—Identifying pathways to harmful alcohol use in adolescence using a panel network approach. *Addiction, 118*(10), 1908–1919. https://doi.org/10.1111/add.16231

Frumkin, M. R., Piccirillo, M. L., Beck, E. D., Grossman, J. T., & Rodebaugh, T. L. (2021). Feasibility and utility of idiographic models in the clinic: A pilot study. *Psychotherapy Research, 31*(4), 520–534. https://doi.org/10.1080/10503307.2020.1805133

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6*(4), 430–450.

Hamaker, E. L., & Wichers, M. (2017). No time like the present: discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science, 26*(1), 10–15. https://doi.org/10.1177/0963721416666518

Herrera-Bennett, A. C., & Rhemtulla, M. (2021). *Exploring the Effects of Sampling Variability, Scale Variability, and Node Aggregation on the Consistency of Estimated Networks*. https://doi.org/10.31234/osf.io/7vkm8

Hoekstra, R. H. A., Epskamp, S., & Borsboom, D. (2023). Heterogeneity in individual network analysis: Reality or illusion? *Multivariate Behavioral Research, 58*(4), 762–786. https://doi.org/10.1080/00273171.2022.2128020

Hoekstra, R. H. A., Epskamp, S., Nierenberg, A. A., Borsboom, D., & McNally, R. J. (2024). Testing similarity in longitudinal networks: The individual network invariance test. *Psychological Methods*. https://doi.org/10.1037/met0000638

Hu, Y., Nesselroade, J. R., Erbacher, M. K., Boker, S. M., Burt, S. A., Keel, P. K., Neale, M. C., Sisk, C. L., & Klump, K. (2016). Test reliability at the individual level. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(4), 532–543. https://doi.org/10.1080/10705511.2016.1148605

Jordan, D. G., Winer, E. S., & Salem, T. (2020). The current status of temporal network analysis for clinical science: Considerations as the paradigm shifts? *Journal of Clinical Psychology, 76*(9), 1591–1612. https://doi.org/10.1002/jclp.22957

Jorgensen, T. D. ;, Pornprasertmanit, S. ;, Schoemann, A. M., & Rosseel, Y. (2022). *semTools: Useful tools for structural equation modeling*. https://cran.r-project.org/package=semTools

Kossakowski, J. J., Groot, P. C., Haslbeck, J. M. B., Borsboom, D., & Wichers, M. (2017). Data from 'critical slowing down as a personalized early warning signal for depression.' *Journal of Open Psychology Data, 5*(1), 1. https://doi.org/10.5334/jopd.29

Laukaityte, I., & Wiberg, M. (2017). Using plausible values in secondary analysis in large-scale assessments. *Communications in Statistics - Theory and Methods, 46*(22), 11341–11357. https://doi.org/10.1080/03610926.2016.1267764

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), 1174–1188. https://doi.org/10.1037/a0024776

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer.

Mansueto, A. C., Wiers, R. W., Van Weert, J. C. M., Schouten, B. C., & Epskamp, S. (2023). Investigating the feasibility of idiographic network models. *Psychological Methods, 28*(5), 1052–1068. https://doi.org/10.1037/met0000466

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2016). What can we learn from plausible values? *Psychometrika, 81*(2), 274–289. https://doi.org/10.1007/s11336-016-9497-x

Martín-Gómez, C., Moreno-Peral, P., Bellón, J. A., Conejo-Cerón, S., Campos-Paino, H., Gómez-Gómez, I., Rigabert, A., Benítez, I., & Motrico, E. (2022). Effectiveness of psychological interventions in preventing postpartum depression in non-depressed women: A systematic review and meta-analysis of randomized controlled trials. *Psychological Medicine, 52*(6), 1001–1013. https://doi.org/10.1017/S0033291722000071

McBride, O., Murphy, J., Shevlin, M., Gibson-Miller, J., Hartman, T. K., Hyland, P., Levita, L., Mason, L., Martinez, A. P., McKay, R., Stocks, T. V., Bennett, K. M., Vallières, F., Karatzias, T., Valiente, C., Vazquez, C., & Bentall, R. P. (2021). Monitoring the psychological, social, and economic impact of the COVID-19 pandemic in the population: Context, design and conduct of the longitudinal COVID-19 psychological research consortium (C19PRC) study. *International Journal of Methods in Psychiatric Research, 30*(1), Article e1861. https://doi.org/10.1002/mpr.1861

McLean, D. C., Nakamura, J., & Csikszentmihalyi, M. (2017). Explaining system missing: Missing data and experience sampling method. *Social Psychological and Personality Science, 8*(4), 434–441. https://doi.org/10.1177/1948550617708015

Mulaik, S. (1972). *The foundations of factor analysis*. McGraw-Hill.

Mundform, D. J., Schaffer, J., Kim, M.-J., Shaw, D., Thongteeraparp, A., & Supawan, P. (2011). Number of replications required in Monte Carlo simulation studies: A synthesis of four studies. *Journal of Modern Applied Statistical Methods, 10*(1), 19–28. https://doi.org/10.22237/jmasm/1304222580

Oh, H., Hunter, M. D., & Chow, S.-M. (2025). Measurement model misspecification in dynamic structural equation models: Power, reliability, and other considerations. *Structural Equation Modeling: A Multidisciplinary Journal, 32*(3), 511–528. https://doi.org/10.1080/10705511.2025.2452884

Oh, H., & Jahng, S. (2023). Incorporating measurement error in the dynamic structural equation modeling using a single indicator or multiple indicators. *Structural Equation Modeling: A Multidisciplinary Journal, 30*(3), 501–514. https://doi.org/10.1080/10705511.2022.2103703

O'Laughlin, K. D., Liu, S., & Ferrer, E. (2021). Use of composites in analysis of individual time series: Implications for person-specific dynamic parameters. *Multivariate Behavioral Research, 56*(3), 408–425. https://doi.org/10.1080/00273171.2020.1716673

Ono, M., Schneider, S., Junghaenel, D. U., & Stone, A. A. (2019). What affects the completion of ecological momentary assessments in chronic pain research? An individual patient data meta-analysis. *Journal of Medical Internet Research, 21*(2), Article e11398. https://doi.org/10.2196/11398

Piccirillo, M. L., & Rodebaugh, T. L. (2019). Foundations of idiographic methods in psychology and applications for psychotherapy. *Clinical Psychology Review, 71*, 90–100. https://doi.org/10.1016/j.cpr.2019.01.002

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Rigabert, A., Motrico, E., Moreno-Peral, P., Resurrección, D. M., Conejo-Cerón, S., Cuijpers, P., Martín-Gómez, C., López-Del-Hoyo, Y., & Bellón, J. Á. (2020). Effectiveness of online psychological and psychoeducational interventions to prevent depression: Systematic review and meta-analysis of randomized controlled trials. *Clinical Psychology Review, 82*, Article 101931. https://doi.org/10.1016/j.cpr.2020.101931

Rigdon, E. E., Becker, J.-M., & Sarstedt, M. (2019). Factor indeterminacy as metrological uncertainty: Implications for advancing psychological measurement. *Multivariate Behavioral Research, 54*(3), 429–443. https://doi.org/10.1080/00273171.2018.1535420

Rintala, A., Wampers, M., Myin-Germeys, I., & Viechtbauer, W. (2019). Response compliance and predictors thereof in studies using the experience sampling method. *Psychological Assessment, 31*(2), 226–235. https://doi.org/10.1037/pas0000662

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

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Schumacher, L., Klein, J. P., Hautzinger, M., Härter, M., Schramm, E., & Kriston, L. (2024). Predicting the outcome of psychotherapy for chronic depression by person-specific symptom networks. *World Psychiatry, 23*(3), 411–420. https://doi.org/10.1002/wps.21241

Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods, 24*(1), 70–91. 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*. https://doi.org/10.3389/fpsyg.2015.01038

Siepe, B. S., Bartoš, F., Morris, T. P., Boulesteix, A.-L., Heck, D., & Pawel, S. (2023). *Simulation Studies for Methodological Research in Psychology: A Standardized Template for Planning, Preregistration, and Reporting.* https://doi.org/10.31234/osf.io/ufgy6

Spearman, C. (1904). "General intelligence", objectively determined and measured. *The American Journal of Psychology, 15*(2), 201–292.

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

Steiger, J. H. (1979). Factor indeterminacy in the 1930's and the 1970's some interesting parallels. *Psychometrika, 44*(2), 157–167. https://doi.org/10.1007/BF02293967

Thomson, G. H. (1934). The meaning of '*i*' in the estimate of '*g*.'. *British Journal of Psychology. General Section, 25*(1), 92–99. https://doi.org/10.1111/j.2044-8295.1934.tb00728.x

Thomson, G. H. (1938). Methods of estimating mental factors. *Nature, 141*(3562), 246–246. https://doi.org/10.1038/141246a0

Thurstone, L. L. (1935). The vectors of mind: Multiple-factor analysis for the isolation of primary traits. *University of Chicago Press*. https://doi.org/10.1037/10018-000

Van Agteren, J., Iasiello, M., Lo, L., Bartholomaeus, J., Kopsaftis, Z., Carey, M., & Kyrios, M. (2021). A systematic review and meta-analysis of psychological interventions to improve mental wellbeing. *Nature Human Behaviour, 5*(5), 631–652. https://doi.org/10.1038/s41562-021-01093-w

Von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. *IERI Monograph Series, 2*, 9–36.

Wen, C. K. F., Schneider, S., Stone, A. A., & Spruijt-Metz, D. (2017). Compliance with mobile ecological momentary assessment protocols in children and adolescents: A systematic review and meta-analysis. *Journal of Medical Internet Research, 19*(4), Article e132. https://doi.org/10.2196/jmir.6641

Wichers, M., Groot, P. C., Psychosystems, ESM Group, EWS Group. (2016). Critical slowing down as a personalized early warning signal for depression. *Psychotherapy and Psychosomatics, 85*(2), 114–116. https://doi.org/10.1159/000441458

Wilhelm, P., & Schoebi, D. (2007). Assessing mood in daily life. *European Journal of Psychological Assessment, 23*(4), 258–267. https://doi.org/10.1027/1015-5759.23.4.258

Wright, A. G. C., & Woods, W. C. (2020). Personalized models of psychopathology. *Annual Review of Clinical Psychology, 16*(1), 49–74. https://doi.org/10.1146/annurev-clinpsy-102419-125032

Wrzus, C., & Neubauer, A. B. (2023). Ecological momentary assessment: A meta-analysis on designs, samples, and compliance across research fields. *Assessment, 30*(3), 825–846. https://doi.org/10.1177/10731911211067538

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2–3), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005

Xiao, Y., Wang, P., & Liu, H. (2023). Assessing intra- and inter-individual reliabilities in intensive longitudinal studies: A two-level random dynamic model-based approach. *Psychological Methods*. https://doi.org/10.1037/met0000608