METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

Multiverse Analysis for Dynamic Network Models: Investigating the Influence of Plausible Alternative Modeling Choices

Björn S. Siepe [1], Daniel W. Heck [1]

[1] *Psychological Methods Lab, Department of Psychology, University of Marburg, Marburg, Germany.*

Methodology, 2025, Vol. 21(2), 123–143, https://doi.org/10.5964/meth.15665

Received: 2024-09-23 • **Accepted:** 2025-03-09 • **Published (VoR):** 2025-06-30

Handling Editor: Eduardo Estrada, Autonomous University of Madrid, Madrid, Spain

Corresponding Author: Björn S. Siepe, Psychological Methods Lab, Department of Psychology, University of Marburg, Gutenbergstraße 18, 35032 Marburg, Germany. E-mail: bjoern.siepe@uni-marburg.de

Supplementary Materials: Code, Data, Materials, Preregistration [see Index of Supplementary Materials]

Abstract

Specifying complex time series models typically allows for a wide range of plausible analysis strategies. However, researchers typically perform and report only a single, preferred analysis while ignoring alternatives that could yield different conclusions. As a remedy, we propose multiverse analysis to investigate the robustness of dynamic network analysis to arbitrary modeling choices. We focus on group iterative multiple model estimation (GIMME), a highly data-driven approach, and re-analyze two datasets (combined n = 199). We vary seven modeling parameters, resulting in 3,888 fitted models. Group-level and to a lesser extent subgroup-level results were mostly stable. Individual-level estimates were more heterogeneous, with some decisions strongly influencing results and conclusions. The robustness of GIMME to alternative modeling choices depends on the level of analysis. For some individuals, results may differ strongly even when changing the algorithm only slightly. Multiverse analysis is a valuable tool for checking the robustness of results from time series models.

Keywords

time series, network analysis, multiverse, heterogeneity, GIMME

There has been a shift towards the use of intensive longitudinal data in the social sciences (Wright & Zimmermann, 2019). This includes data collection through repeated

daily experience sampling questionnaires or passive sensors such as heart-rate monitors or smartphones (Wright & Zimmermann, 2019). Such data may contribute to a better understanding of the dynamics of intra-individual psychological processes and inter-individual differences therein. There exists a wide variety of different time series models for such types of data (for an overview, see Jordan et al., 2020). Such time series models are often visualized and interpreted as dynamic networks of mutually influencing variables (Bringmann et al., 2022). It is well known that model choice and specification substantially influence the results of such analyses. However, the statistical literature on the comparative performance of different methods is inconclusive, and hence, there is often no consensus on which modeling approach is preferable in a given application. As a result, there is considerable heterogeneity in analysis techniques used to investigate similar questions.

Amongst the various alternatives, group iterative multiple model estimation (GIMME, Gates & Molenaar, 2012) has emerged as a popular approach. It combines the benefits of personalized models with the identification of shared relationships at the (sub-)group level (Wright & Woods, 2020). GIMME performs a data-driven search for directed temporal (across time points) and contemporaneous associations (at the same time point). The approach estimates separate person-specific models while searching for associations between variables shared by a majority of the full sample or subgroups thereof (Beltz & Gates, 2017). The algorithm has several strengths, including its good performance in simulation studies and its use of group information to improve individual model fit (e.g., Hoekstra et al., 2022, Lane et al., 2019, Nestler & Humberg, 2021). In the following, we interpret fitted GIMME models as dynamic networks, where *nodes* represent variables and *edges*/*paths* the statistical association between variables. An *edge weight* refers to the strength of this association. A more extensive introduction to dynamic network models is available in Borsboom et al. (2021) and Bringmann et al. (2022).

However, even when researchers have settled on a specific modeling approach such as GIMME, they still have ample flexibility along the analysis path, ranging from variable choice, preprocessing, and modeling to the interpretation of results. This is especially true for highly data-driven, iterative algorithms such as GIMME. The conventional workflow involves making a certain number of decisions, based on which a single, supposedly most appropriate model is identified and reported. However, focusing on a single analysis disregards other reasonable modeling choices that could have led to meaningfully different conclusions (Steegen et al., 2016).

The idea of repeatedly performing a certain analysis iterating through a range of plausible alternative specifications is the premise of multiverse analysis (Steegen et al., 2016). The goal of a multiverse analysis is to systematically analyze the uncertainty associated with defensible choices in the modeling process. This aspect would not be explored when performing and reporting only a single analysis (Del Giudice & Gangestad, 2021). In experience sampling research, multiverse analyses have been used to investigate

outcomes such as the influence of the choice of central tendency measures (Weermeijer et al., 2022) or measurement flexibility (Dejonckheere et al., 2018). These studies show that investigating the effects of alternative, often arbitrary analysis decisions is important for assessing the robustness of results.

Previous multiverse analyses have usually focused on a single or a few parameters of interest. The stability of a few parameters across different specifications can easily be examined graphically and with inferential statistics (e.g., Hall et al., 2022, Simonsohn et al., 2020). Dynamic network analyses do not necessarily provide a single parameter of interest. Instead, a large number of estimates and network summaries have to be interpreted across (sub-)group and individual network structures (Wright et al., 2019). Therefore, the application of multiverse analyses in this context requires new ways of analyzing results, which we will explain in this manuscript.

Time series modeling using data-driven algorithms is a complex process involving multiple arbitrary decisions. By arbitrary, we mean that these choices are neither theoretically motivated, nor based on extensive statistical evidence. For GIMME, this includes thresholds for (sub-)group effects (defined in terms of a certain percentage of individuals) or cutoffs for good model fit, such as the Root Mean Square Error of Approximation (RMSEA), which are explained below. The cutoffs for good model fit are currently set to fixed, mostly arbitrary default values within the algorithm. While these defaults are motivated by cutoffs commonly used in structural equation modeling, the same cutoffs do not necessarily apply to GIMME models. By now, there is abundant literature criticizing the use of generic fit index cutoffs in the psychometrics literature (e.g., McNeish & Wolf, 2023). This criticism is even more relevant when transferring cutoffs to very different model classes, such as GIMME.

A previous simulation study has already indicated that the default choices, which cannot be changed in the standard version of GIMME, may not always be optimal (Nestler & Humberg, 2021). Nestler and Humberg (2021) also showed that characteristics of group-level results can have an unexpected effect on individual-level results. Thus, changing cutoffs could have downstream implications for individual-level models that are hard to anticipate due to the iterative nature of the algorithm. Relying on a single, estimated GIMME model as the only basis for drawing substantive conclusions ignores the inherent uncertainty when making such arbitrary modeling choices.

Aims

We propose to use multiverse analysis to study the robustness of psychological time series analysis. We explored a range of different fitting criteria and compared the resulting models to models estimated in previously published empirical studies. We aim to provide answers to questions such as: Had the models been specified slightly differently, would results and the corresponding conclusions substantially change? How can we best quantify and visualize the robustness of a complex model? Among several equally

justifiable specifications, which aspects of a specific fitted GIMME model are most robust on a given dataset? In doing so, this manuscript serves as a blueprint and guidance for future multiverse analyses with GIMME.

Method

We used R version 4.3.1. (R Core Team, 2023) and preregistered all analyses. The preregistration (and deviations from it), code, information on the computational environment, all R packages used, and all supplementary materials are available at Siepe and Heck (2025a).

Data

Personality Dataset

The first dataset ('personality dataset') contains daily diary measures of $n = 94$ participants (Wright et al., 2019). Participants had at least one personality disorder diagnosis. Daily diary measures were completed in the evening, including questions about a daily summary of affect, interpersonal behavior, stress, and functioning. Except for functioning (assessed with a single 5-point item), all constructs were measured by multiple items aggregated into sum scores. Participants were only included in the analysis when they provided at least 60 observations. The average time series length of included participants was 91.48 ($SD = 9.00$). More information can be found in Wright et al. (2019).

Emotion Dataset

The second dataset ('emotion dataset') contains experience sampling data of $n = 105$ individuals (Kullar et al., 2024). Participants either had a major depressive or a bipolar disorder, were in remission of a major depressive disorder, or had no history of any mood or anxiety disorder. Smartphone surveys were sent out five times daily for 14 days. In these, participants rated nine momentary emotions on 7-point Likert scales. Further, they stated how long their current emotional state lasted and whether they thought about something else than their current activity. The model by Kullar et al. (2024) included *time-of-day* as an exogenous predictor to handle time trends. Individuals were excluded when they completed less than 50% of the surveys. The average time series length of included individuals was 62.31 ($SD = 8.11$). More information can be found in Kullar et al. (2024).

GIMME

To avoid confusion across the different levels of analyses, we use the following terms: *GIMME model* as the overarching modeling approach, a *specification* as a specific iteration of the multiverse, a *fitted model* as the overall GIMME result within a specific

iteration, and an *individual fitted model* as the result of GIMME for a specific individual within an iteration. GIMME is a form of unified structural equation modeling (uSEM) in which each variable is predicted by itself and all other variables at the previous time point (temporal) as well as all other variables at the same time point (contemporaneous). We provide a full description of the GIMME algorithm in our supplementary materials (Siepe & Heck, 2025a).

Briefly, the algorithm starts by estimating a null model without any paths for each individual.[1] Then, group-level associations are searched in an iterative manner using modification indices for each individual. Paths are added at the group level if they significantly improve model fit for a certain proportion of individuals. Subgroups are then searched based on similarities in the individual fitted models. In the resulting subgroups, subgroup paths are added (i.e., freely estimated per individual) and pruned in the same way as for group effects. Finally, for each individual fitted model, paths are added that significantly improve model fit until an 'excellent' model fit is achieved at the individual level. Per default, this is defined as two out of four fit indices satisfying certain thresholds: Root Mean Squared Error of Approximation (RMSEA < .05), Standardized Root Mean Square Residual (SRMR < .05), Comparative Fit Index (CFI > .95), Non Normed Fit Index (NNFI > .95; see Hu & Bentler, 1999, for more information).

Here, we focus on 'classic' exploratory GIMME as described in Gates et al. (2017) to mimic the analyses in the empirical papers from which we obtained our datasets. In the original GIMME package, users can neither modify the threshold values used for each fit index nor the number of fit indices that need to satisfy these thresholds. The modified package that enables these changes is available at Siepe and Heck (2025d).

Multiverse Analysis

We vary seven parameters of the GIMME approach. This includes two parameters adjustable in the original GIMME package (group and subgroup threshold) and five parameters that have so far been defined as fixed (goodness-of-fit criteria). The first two parameters specify the required proportion of individuals at different levels for which a path must significantly improve model fit (default values designated with a superscripted 'a', i.e., [a]):

1. Group threshold ∈ {50%, 60%, 75%[a], 80%}
2. Subgroup threshold ∈ {50%, 60%, 75%[a], 80%}

Five parameters refer to the fit indices used for model selection:

3. RMSEA cutoff ∈ { .03, .05[a], .08}
4. SRMR cutoff ∈ { .03, .05[a], .08}
5. NNFI cutoff ∈ { .90, .95[a], .97}

1) Except for autoregressive paths in default GIMME.

6. CFI cutoff ∈ { .90, .95[a], .97}
7. Fit measures satisfying the cutoffs ∈ {1, 2[a], 3}

Creating a grid of all possible combinations results in $4^2 \times 3^5 = 3,888$ possible model specifications for each data set. One of these combinations provides the *reference specification* using the GIMME default settings, except for subgroup thresholds. For the emotion dataset, we used a subgroup threshold of 51% as indicated in Kullar et al. (2024). For the personality data set, we used the default GIMME settings (highlighted in bold) as reference specifications, except for the subgroup threshold of 50% used in the original study. In the following, we will refer to the cutoffs for the (sub-)group thresholds ranging from 50% to 80% as 'liberal', 'medium-liberal', 'medium-strict', and 'strict'. The labels for the fit index cutoffs 'strict', 'medium', and 'liberal' refer to increasing maximum cutoffs of.03,.05, and.08, respectively, and decreasing minimum cutoffs of.97,.95, and.90, respectively. For the number of fit measures satisfying the cutoffs, larger values are stricter.

Multiverse analyses have been criticized for combining analysis choices that are not equally justified (Del Giudice & Gangestad, 2021). The different modeling choices we assess can be considered arbitrary since they are neither theoretically motivated nor justified by a wealth of statistical literature. Rather, the two (sub-)group thresholds reflect assumptions of the researcher about how many individuals must have a non-zero path for it to be included at the group level. The different fit index cutoffs serve as stopping criteria and thus adjust the conservativeness of the algorithm. Overall, all seven parameters can be expected to affect the sensitivity-specificity tradeoff.

Changes to the stopping rules of GIMME have not yet been evaluated in simulation studies. Hence, we performed a small simulation study to assess the performance of GIMME when using the modified cutoffs. Results are provided in the supplement (see Siepe & Heck, 2025a). The results showed that the different fit index cutoffs did not substantially affect the overall performance across simulation conditions. We conclude that there is insufficient evidence to generally favor a specific fit-index specification over any other.

Statistics of Interest

We compute various statistics at all levels of the GIMME model structure to compare the 3,888 model specifications. We summarise all statistics in Table 1. At the *group* level, we compare the overall number and the specific group-level edges against the reference model. We ignore autoregressive edges, as they are freely estimated by default. Counting these would artificially inflate homogeneity.[2] Homogeneity is investigated by dividing the number of group edges by the total number of nonzero edges across individuals,

2) We do not ignore autoregressive effects when computing edge weight differences.

as proposed by Hoekstra et al. (2022). High values imply that most non-zero edges are shared across individuals.

Table 1

Summary Statistics for a Fitted GIMME Model

Level	Summary
Group	• Number of group-level edges • Specific group-level edges present/absent • Homogeneity: number of group edges divided by number of total edges
Subgroup	• Number of subgroups • Size of subgroups • Huber-Arabie Adjusted Rand Index (ARI) • Variation of Information (VI)
Individual	• Difference in adjacency matrix • Difference in edge weights • Difference in density • Difference in fit measures • Difference in out-strength • Most central node

Note. Individual results are first calculated per individual and then aggregated across all individuals in a specification.

At the *subgroup* level, we calculate the number and size of subgroups. Furthermore, we evaluate the adequacy of the estimated subgroup solutions by computing the Hubert-Arabie Adjusted Rand Index (ARI) and the Variation of Information (VI). The ARI is a measure for cluster recovery and quantifies the contingency between the true subgroup structure and the estimated clustering while correcting for chance similarity (Hubert & Arabie, 1985). An ARI of at least 0.80 has been used as a threshold for similarity (Gates et al., 2019). The VI is based on information theory and assesses the distance between two clusterings, defined as the information that is lost when moving from one clustering to another (Meilă, 2007).

At the *individual* level, we compute differences between all specifications and the reference specification for the edge weight estimates and the adjacency matrix of an individual. The adjacency matrix encodes the occurrence of certain paths in a binary fashion. As an overall measure of the network structures, we compute network density (the average of absolute edge weights) for temporal and contemporaneous relationships. We used all edge weights, including zero weights, as excluding zero weights could lead to an inflated perception of network density. We also compare the fit statistics for an individual model. Finally, we calculate out-strength centrality defined as the sum of all

edge weights going out of a specific node. This measure has been used as a proxy for the importance of a variable (Bringmann et al., 2019). We compare centrality estimates against the reference model and check whether the most central node is identical, as the most central node is sometimes considered a potential target for treatments (Bringmann et al., 2022).

Visualizations

Visualizations are an important tool for multiverse analysis (Hall et al., 2022). To make our results more accessible, we provide an interactive Shiny app to explore the multiverse. The app is hosted at Siepe & Heck, 2025b or can be downloaded to render locally from Siepe & Heck, 2025c. The app includes functionality to calculate grouped summary statistics per specification, visualize results in different ways, and compute networks for each specification.

Results

Personality Dataset

We were unable to exactly replicate the results of Wright et al. (2019), likely due to changes to the GIMME algorithm over the years. We will use the results obtained with the current version of GIMME since they closely resemble the original results. Only one effect surpassed the group-level threshold. Most individual temporal networks were sparse, with the lagged effect from *Negative Affect* to *Stress* being the most commonly estimated edge (9 individuals) besides autoregressive effects and the (sub-)group effects. Individual contemporaneous networks were more densely connected. Three subgroups were identified (59, 24, and 11 members).[3] A subgroup path emerged only in the first subgroup. At the individual level, *Negative Affect* and *Stress* had the highest temporal and contemporaneous centrality for the largest number of individuals, respectively.

Multiverse: Group Level

All models converged. Around 92.4% of the specifications resulted in a single group-level path from *Stress* to *Negative Affect*, matching the reference model. The remaining 7.6% specifications led to two group-level paths. The homogeneity value varied across specifications and went up to twice the value of the reference fit. As a hypothetical example, if there was one group-level path (for all 94 individuals) and 47 additional, non-group-level paths, homogeneity would be $94/(94 + 47) = 2/3$. Across the multiverse, it ranged from 12.6% to 29.1%, compared to the reference homogeneity of 14.8%.

3) In the original publication, one individual was in the largest subgroup instead of the second largest one.

Multiverse: Subgroup Level

The same three subgroups were obtained in all specifications. Consequently, we did not compute any subgroup comparison metrics. Due to potential differences in the individual-level models, some characteristics of the subgroups, such as the number of paths between certain nodes, may still vary across specifications.

Multiverse: Individual Level

We calculated the absolute difference between the individual fitted and the reference adjacency matrices. Then, we summed these differences across all individuals. The average difference of the adjacency matrix for an individual was 1.745 (SD = 0.718), which means that, on average, the presence (or absence) of slightly less than two paths differed from the reference fit. However, seven individuals had an average difference larger than four.

Figure 1 shows a specification curve analysis (Simonsohn et al., 2020) which visualizes the value of a statistic of interest across all specifications. The first panel shows the statistic of interest at the y-axis (here, the mean of the adjacency matrix differences). For example, a value of 1 indicates that on average, across all individual models in a certain specification, one path differed in its presence to the reference model. On the x-axis, all 3,888 model specifications are ordered using this statistic. The second panel shows the strictness of the different specifications.

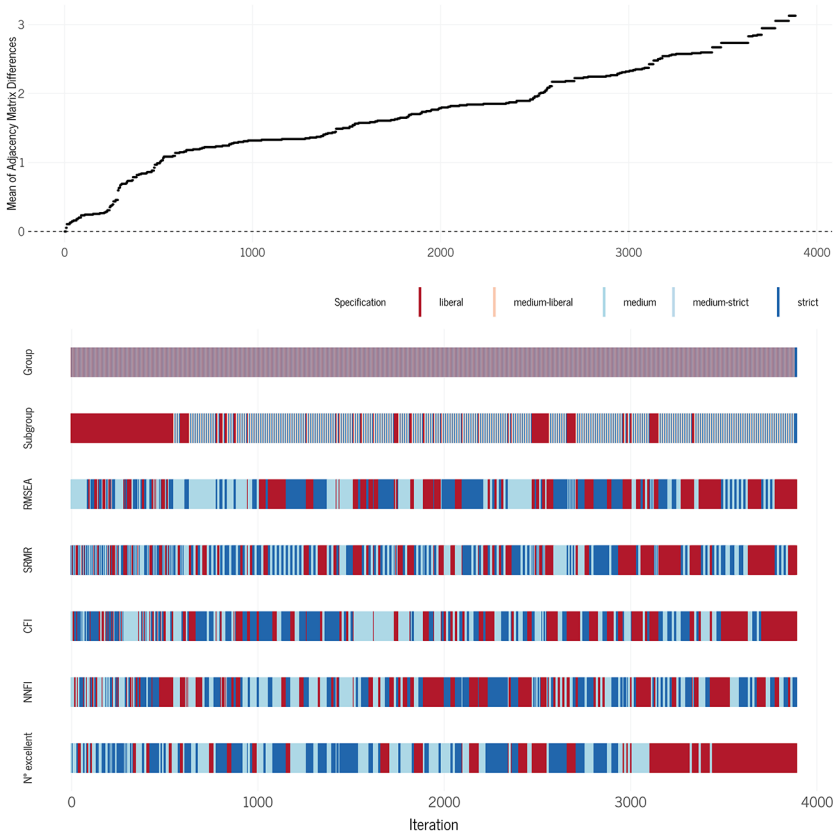
Figure 1 shows a clear trend: When fewer fit indices have to reach a certain criterion for a model to converge (*n.excellent*), the difference to the reference model becomes larger. For the four fit index cutoffs, there appears to be a similar, but weaker, trend. When the cutoffs are more liberal, the difference to the reference model increases.

Next, we investigate the frequency of certain paths being included across all specifications of the multiverse. We present a network multiverse plot in Figure 2 to visualize this. The upper row shows the percentages of individual fitted models in the reference fit containing a specific edge. In the lower row, the plot shows the proportion of all individual models across all specifications that included a specific edge. For example, roughly 10% of all individual models across all specifications contained the path from *Stress* to *Negative Affect*. We do not show paths that were included in less than 5% of all models. Overall, the results of the multiverse are very similar to the reference fit. Temporal paths occurred less frequently than contemporaneous ones. Each possible effect was included at least once in any model across the multiverse.

To compare individual path estimates, we computed the absolute difference of all paths of all individual fitted models within a certain specification relative to the reference model. We then calculated the average of all differences and of all non-zero differences across all individuals. The average mean non-zero difference across all specifications was 0.125 (SD = 0.031), while the average mean difference was 0.012 (SD = 0.005). The average non-zero edge weight in the reference model was 0.320. This means

Figure 1

Specification Curve of Adjacency Differences for the Personality Dataset.



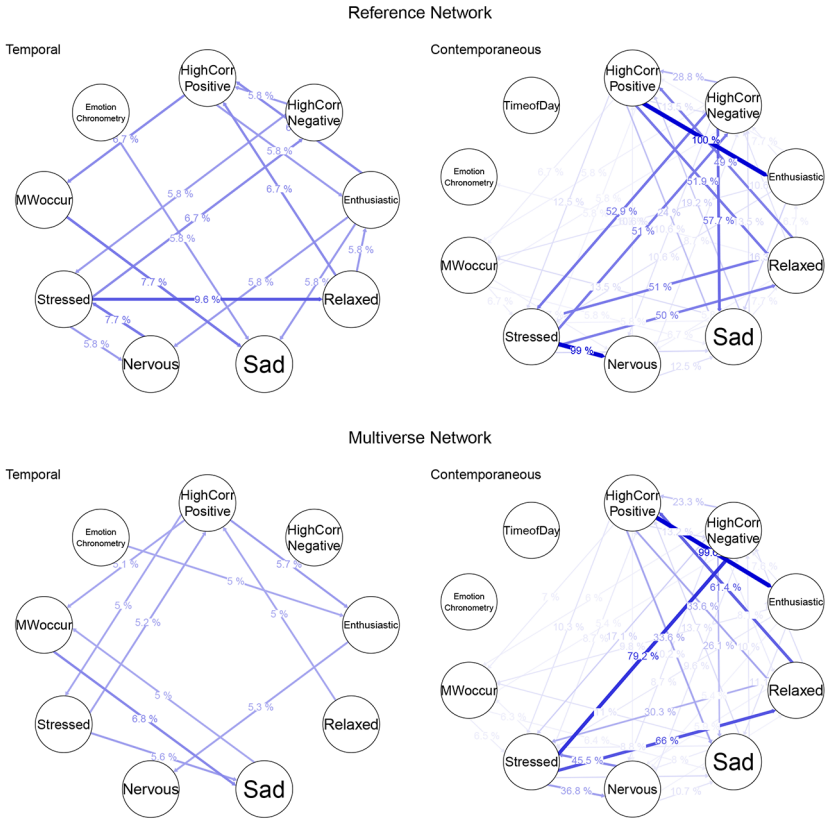
Note. The upper panel is ordered by the size of the mean of the summed differences. The lower panel shows the color-coded levels of different specifications. The verbal labels correspond to different numerical values for each of the 7 factors in the lower panel (for details, see the Method section).

that when an edge weight in any multiverse specification was different from the edge weight in the reference fit, they differed by an average absolute value of 0.125.

To aid interpretation, we use network density (the sum of all absolute edge weights) as a summary statistic and compare it against the reference fit. A clear trend emerged: The more cutoffs are required for convergence (coded here as 'strict'), the denser the resulting networks were. A similar, although weaker, trend emerged for the four fit index cutoffs. We present a specification curve analysis as well as detailed results for out-strength centrality and different fit statistics are provided in the supplement (Siepe

Figure 2

Multiverse Network for the Personality Dataset



Note. The upper row of the plot shows the percentage of non-zero individual paths in the reference fit (heterogeneity across individuals). The lower row shows the percentage of non-zero edges across all specifications and individual models (indicating heterogeneity across individuals and specifications). Autoregressive paths are omitted. The sign of an effect is ignored. Thickness and transparency of the arrows are scaled to the maximum edge size.

& Heck, 2025a). For some individuals (12 for the temporal and 6 for the contemporaneous network), the most central node was identical to the reference model in less than one-third of all specifications.

Emotion Dataset

Similar to the original analyses, all but one individual fitted model in the reference analysis converged successfully. In the reference model for the emotion dataset, only one effect (contemporaneous effect from *Happy* to *Enthusiastic*) surpassed the threshold to be included as a group effect. In general, individual temporal networks were relatively sparse compared to the contemporaneous networks. Besides autoregressive effects (included in all models by default), the most common temporal effect was the path from *Stressed* to *Relaxed* (estimated for 10 participants). Two subgroups with 53 and 51 members were obtained with five and four contemporaneous paths, respectively. One path was not shared among the two subgroups, whereas the other four contemporaneous paths connected the same variable pairs, but in different directions. *Happy* (temporal network) and *Stressed* (both networks) were the most frequent central nodes.

Multiverse: Group Level

In 1,335 specifications, one individual-level model did not converge (the same one as in the original analysis). In 63 specifications, the models of two individuals did not converge. In all other specifications, all models converged. In about half of the specifications, one group-level effect was estimated, matching the reference model. In the other half of the specifications, four or five group-level effects were estimated (each with roughly the same proportion). With stricter group cutoffs, these effects often became subgroup effects, which can be explored in our shiny application. Homogeneity of the multiverse models ranged between 6.4% and 65.4% (M = 24.2%). The reference model had a homogeneity of 7.9%, so most of the multiverse specifications resulted in a higher homogeneity.

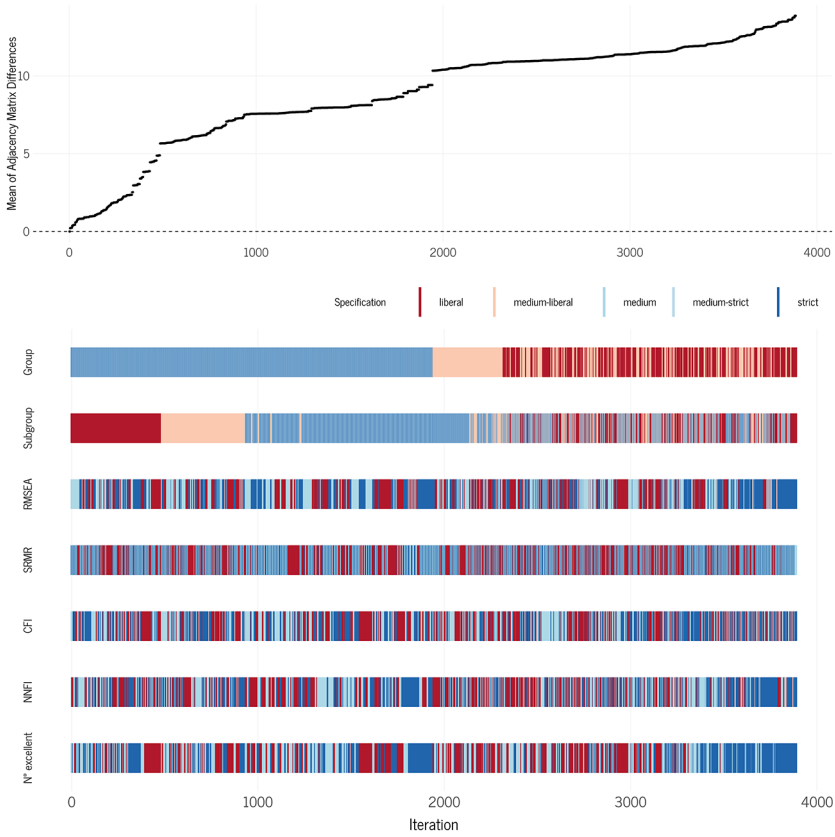
Multiverse: Subgroup Level

Three subgroups were found in all specifications, but three different subgroup configurations emerged. A subgroup comprised of a single individual emerged in all specifications. The other two subgroups were either of size 31 and 73, 65 and 39, or 53 and 51 (as in the original study), respectively. Each of these alternatives occurred 972 times. Therefore, the ARI and VI took three distinct values. We only report the ARI here due to its interpretability. In half of the specifications, it was 1, indicating perfect similarity. In the other specifications, it was either 0.09 or 0.12. As stated before, an ARI of at least 0.80 is sometimes used as a threshold for similarity (Gates et al., 2019). Hence, roughly half of the subgroup solutions are very dissimilar to the original study.

Due to an oversight, we initially used a subgroup threshold of 50% instead of the original value of 51%. This small change led to 128 paths being different compared to the original study, as the addition of a subgroup effect had downstream effects. All results reported here refer to the analysis using 51%.

Figure 3

Specification Curve of Adjacency Differences for the Emotion Dataset



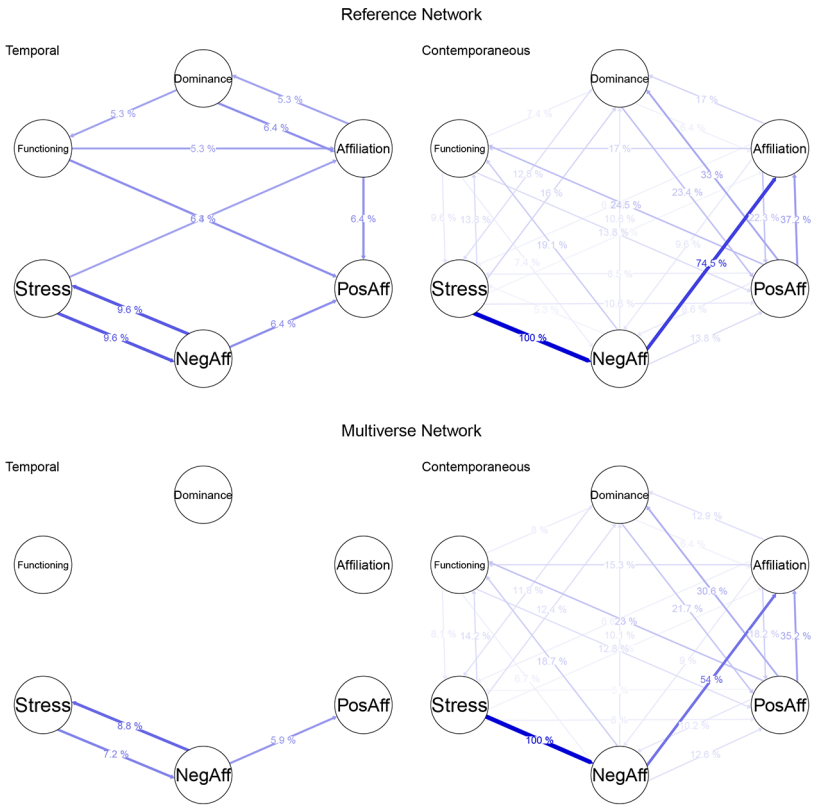
Note. Panel A is ordered by the size of the differences of adjacency matrices per specification. Panel B shows the color-coded levels of different specifications. The y-axis has a different scale than Figure 1.

Multiverse: Individual Level

The average absolute individual adjacency matrix difference was 8.872 (*SD* = 3.286). This means that, on average, almost nine paths differed in their presence between the multiverse and the reference model. Figure 3 shows a specification curve plot illustrating these differences. Larger differences on the right side of the plot indicate larger average deviations. A clear effect of group thresholds emerged, where more liberal group thresholds resulted in larger average differences. In contrast, more liberal subgroup cutoffs resulted both in relatively strong positive and negative differences (right and left side of Figure 3).

Figure 4

Multiverse Network for the Emotion Dataset



Note. See Figure 2 for an explanation of the plot. 'HighCorrPositive' and 'HighCorrNegative' refer to the aggregate Positive and Negative Affect variables, respectively, to keep the notation consistent with the original paper.

Figure 4 shows a multiverse network plot illustrating the frequency with which different paths were found across the multiverse. The plot indicates some discrepancies between the multiverse specifications compared to the reference model, especially regarding the directionality of the contemporaneous paths. As for the personality dataset, paths were included more frequently in the contemporaneous than in the temporal network.

We again calculated the average absolute differences in edge weights. The average difference when excluding paths that are zero across all specifications was 0.202 (*SD* = 0.059), while the average mean difference of all paths (including zero paths) was

0.030 (*SD* = 0.011), both larger than in the personality dataset. This indicates that, when an edge was different for an individual, it differed by roughly 0.2 on average. Given the average absolute nonzero edge weight of 0.358 in the reference model, this is a substantial difference.

For network density, we found that while most specifications are within 10% of the reference density, some differences are close to 30%. The clearest pattern emerged for the number of fit indices exceeding a cutoff for convergence, where a stricter setting led to denser networks. A similar tendency emerged for stricter settings of the RMSEA and the CFI. A visualization of network density differences as well as results for centrality and fit indices are available in the supplement (Siepe & Heck, 2025a). For 30 (temporal network) and 29 (contemporaneous network) individuals, the most central node was identical in less than one-third of the specifications.

Discussion

We introduced multiverse analysis to multivariate psychological time series modeling and used it to evaluate the robustness of previous empirical research to alternative plausible model specifications. Results were generally robust at the group level. However, fitted individual-level networks were more sensitive to alternative specifications.

Implications for Applied Research

Users can already change group and subgroup thresholds when using the standard GIMME package (Lane et al., 2019). In our reanalysis, this choice had no relevant impact on the personality dataset. In the emotion dataset, however, changing the (sub-)group thresholds had a visible impact on the model outcome. Even a slight change from 50% to 51% had a small but notable effect on the estimated model. The average number of observations in the emotion dataset was relatively small. This is consistent with evidence from simulation studies (Lane et al., 2019, Nestler & Humberg, 2021) showing that results of GIMME with relatively few time points should be interpreted cautiously.

In both data sets, varying the number of fit indices needed to establish convergence from one to three had a substantial influence on the results. For example, the density of networks, a popular network characteristic (Bringmann et al., 2022), differed substantially depending on the fit index convergence criterion. The most central node, sometimes used as an indicator for therapeutic targets (Bringmann et al., 2022), also differed across many specifications for some individuals. Even for summary statistics that may be interpreted as relatively stable across the multiverse (e.g., differences in adjacency matrices), we found considerable interindividual differences in this stability. While model results may be robust overall, they may differ strongly for some individuals. The implications of our results therefore depend on the level of analysis that one is interested in.

Results at the group level generally seem to be more robust than those at the individual level. Especially if models such as GIMME are interpreted at the individual

level, for instance, for developing individualized treatment rules (Ong et al., 2022), the inherent uncertainty in model fitting should be considered. A multiverse analysis not only guards against the overinterpretation of chance findings but may also increase the trustworthiness of results if conclusions remain robust.

In practice, researchers could thus pre-specify the preferred main analysis and the main outcome(s) of interest together with certain robustness criteria. For example, a primary goal of a study may concern the subgrouping of individuals into clusters. In this case, researchers should focus on the robustness of subgroup solutions by defining the percentage of specifications in the multiverse yielding identical or highly similar subgroups. The primary interest may alternatively be in estimating the association of a network metric such as network density with an external covariate or outcome. To assess the robustness of this association across the multiverse, researchers can adapt the changes we have made to the GIMME package. Of course, one can also modify existing analysis options or other pre-processing choices considered to be relevant. This recommendation applies not only to GIMME analyses but also to other dynamic network analyses whenever multiple decisions need to be made in the modeling process. Instead of exploring a very large multiverse as in this article, researchers can start small with one or two parameters of interest and then expand based on the computational resources available.

Implications for Methodological Research

The interpretation of models such as GIMME, containing many parameters at multiple levels, can already be challenging when obtaining only a single model output. Fitting a few thousand such models increases the difficulties in interpretation. We proposed several approaches for aggregating, visualizing, and summarizing relevant parameters at the group, subgroup, and individual levels. In doing so, we provide guidance and inspiration for future multiverse analyses in multivariate time series analyses.

Our results have implications for methodological research on GIMME more broadly. We showed that slightly changing arbitrary model-search criteria can markedly alter the resulting estimates. While default GIMME has shown good performance in simulation studies (e.g., Gates et al., 2017, Hoekstra et al., 2022, Lane et al., 2019), substantive conclusions based on GIMME may not always be robust to relatively small changes in the algorithm. The simulation study of some selected scenarios in the supplementary material (Siepe & Heck, 2025a) illustrates that the impact of these arbitrary modeling choices on the performance of GIMME can be small and depends on context. In some scenarios, such as data sets with few observations, more liberal cutoffs might be justified to increase sensitivity. Hence, we have no principled reasons to generally prefer one set of algorithmic decisions for GIMME.

This highlights a larger point: Even if future simulations provide evidence about the performance of particular model specifications, researchers should not overestimate

the certainty of recommendations derived from simulation studies with a limited range of settings (Siepe et al., 2023). In simulation studies, researchers always have to focus on a subset of data-generating scenarios (e.g., a specific number of observations or the strength of a true effect) which are then repeatedly used to simulate data sets with random noise. Such simulations often require idealized assumptions, such as multivariate normality or some forms of homogeneity and independence. In many applied settings, especially when dealing with complex time-series data, it is often unclear to what extent such assumptions are met and how much their violation would affect the results of a study. Multiverse analyses provide a remedy for these limitations by using the data at hand with all their peculiarities. Instead of relying on simulation results on how certain analysis decisions affect the results in general, one can directly check whether different preprocessing steps and alternative modeling decisions matter for the analysis at hand. While simulation studies can thus answer broad questions about average performance in reasonable settings, multiverse analyses can show the robustness of results for a specific data set. Exploring the robustness of results and conclusions for specific empirical data sets will therefore remain an important task.

Limitations

Beyond the choices we studied, there are many other important decisions in modeling time series data. For example, the variables in the original personality analysis were apparently not detrended. As an exploratory analysis, we applied a commonly used detrending procedure to the dataset. This led to the detection of seven instead of three subgroups in the original analysis, indicating that other choices in the analysis pipeline may be consequential. However, the decision of whether and how to detrend is not arbitrary and therefore not necessarily appropriate for a multiverse analysis.

Our multiverse analysis here was a case study using a single modeling framework on two datasets that differed in a number of characteristics. Nevertheless, the results of our analysis do not necessarily generalize to new datasets or other modeling approaches. This again emphasizes the utility of future multiverse analyses.

Besides the multiverse approach, GIMME analyses themselves have several limitations (e.g., Kullar et al., 2024, Wright et al., 2019). These include the required resources in terms of length and frequency of assessments, the selection of relevant variables for individuals, and the assumption of stationarity. Also, while there is great promise in using models such as GIMME in applied clinical settings (Wright & Woods, 2020), it is not yet clear which parameters or summary statistics will be most useful (Bringmann et al., 2022).

Conclusion

We conducted a multiverse analysis to re-analyze two previously published analyses that used GIMME. Overall, we found satisfactory robustness of group-level results against

alternative model specifications. However, some individual-level results varied substantially between different specifications. Overall, multiverse analysis is a valuable tool for investigating researchers' degrees of freedom in complex time series analyses.

Funding: The authors have no funding to report.

Acknowledgments: We thank Monica Kullar and Aidan Wright for sharing their data and code online and permitting us to reuse it. We thank Lea Schumacher and Matthias Kloft for their helpful comments on an earlier version of this manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Author Contributions: *Björn S. Siepe*: Conceptualization, Methodology, Formal Analysis, Investigation, Software, Visualization, Writing—Original Draft Preparation, Writing—Review & Editing. *Daniel W. Heck*: Conceptualization, Methodology, Supervision, Writing—Review & Editing

Data Availability: The preregistration and all code, data, and supplementary materials are available at Siepe and Heck (2025a)

Supplementary Materials

For this article, the following Supplementary Materials are available:

- Code. (Siepe & Heck, 2025a)
- Data. (Siepe & Heck, 2025a)
- Preregistration. (Siepe & Heck, 2025a)
- Study materials. (Siepe & Heck, 2025a)

References

Beltz, A. M., & Gates, K. M. (2017). Network mapping with GIMME. *Multivariate behavioral research, 52*(6), 789–804. https://doi.org/10.1080/00273171.2017.1373014

Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., Robinaugh, D. J., Perugini, M., Dalege, J., Costantini, G., Isvoranu, A.-M., Wysocki, A. C., van Borkulo, C. D., van Bork, R., & Waldorp, L. J. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers, 1*(1), Article 58. https://doi.org/10.1038/s43586-021-00055-w

Bringmann, L. F., Albers, C., Bockting, C., Borsboom, D., Ceulemans, E., Cramer, A., Epskamp, S., Eronen, M. I., Hamaker, E., Kuppens, P., Lutz, W., McNally, R. J., Molenaar, P., Tio, P., Voelkle, M. C., & Wichers, M. (2022). Psychopathological networks: Theory, methods and practice. *Behaviour Research and Therapy, 149*, Article 104011. https://doi.org/10.1016/j.brat.2021.104011

Bringmann, L. F., Elmer, T., Epskamp, S., Krause, R. W., Schoch, D., Wichers, M., Wigman, J. T., & Snippe, E. (2019). What do centrality measures measure in psychological networks? *Journal of Abnormal Psychology, 128*(8), 892–903. https://doi.org/10.1037/abn0000446

Dejonckheere, E., Mestdagh, M., Houben, M., Erbas, Y., Pe, M., Koval, P., Brose, A., Bastian, B., & Kuppens, P. (2018). The bipolarity of affect and depressive symptoms. *Journal of Personality and Social Psychology, 114*(2), 323–341. https://doi.org/10.1037/pspp0000186

Del Giudice, M., & Gangestad, S. W. (2021). A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science, 4*(1). https://doi.org/10.1177/2515245920954925

Gates, K. M., Fisher, Z. F., Arizmendi, C., Henry, T. R., Duffy, K. A., & Mucha, P. J. (2019). Assessing the robustness of cluster solutions obtained from sparse count matrices. *Psychological Methods, 24*(6), 675–689. https://doi.org/10.1037/met0000204

Gates, K. M., Lane, S. T., Varangis, E., Giovanello, K., & Guiskewicz, K. (2017). Unsupervised classification during time-series model building. *Multivariate Behavioral Research, 52*(2), 129–148. https://doi.org/10.1080/00273171.2016.1256187

Gates, K. M., & Molenaar, P. C. M. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage, 63*(1), 310–319. https://doi.org/10.1016/j.neuroimage.2012.06.026

Hall, B. D., Liu, Y., Jansen, Y., Dragicevic, P., Chevalier, F., & Kay, M. (2022). A survey of tasks and visualizations in multiverse analysis reports. *Computer Graphics Forum, 41*(1), 402–426. https://doi.org/10.1111/cgf.14443

Hoekstra, R. H. A., Epskamp, S., & Borsboom, D. (2022). Heterogeneity in individual network analysis: reality or illusion? *Multivariate Behavioral Research, 58*(4), 762–786. https://doi.org/10.1080/00273171.2022.2128020

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193–218. https://doi.org/10.1007/BF01908075

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

Kullar, M., Carter, S., Hitchcock, C., Whittaker, S., Wright, A. G. C., Dalgleish, T. (2024). Patterns of emotion-network dynamics are orthogonal to mood disorder status: An experience sampling investigation. *Emotion, 24*(1), 116–129. https://doi.org/10.1037/emo0001245

Lane, S. T., Gates, K. M., Pike, H. K., Beltz, A. M., & Wright, A. G. (2019). Uncovering general, shared, and unique temporal patterns in ambulatory assessment data. *Psychological Methods, 24*(1), 54–69. https://doi.org/10.1037/met0000192

McNeish, D., & Wolf, M. G. (2023). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods, 28*(1), 61–88. https://doi.org/10.1037/met0000425.supp

Meilă, M. (2007). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis, 98*(5), 873–895. https://doi.org/10.1016/j.jmva.2006.11.013

Nestler, S., & Humberg, S. (2021). Gimme's ability to recover group-level path coefficients and individual-level path coefficients. *Methodology, 17*(1), 58–91. https://doi.org/10.5964/meth.2863

Ong, C. W., Hayes, S. C., & Hofmann, S. G. (2022). A process-based approach to cognitive behavioral therapy: A theory-based case illustration. *Frontiers in Psychology, 13*, Article 1002849. https://doi.org/10.3389/fpsyg.2022.1002849

R Core Team. (2023). *R: A language and environment for statistical computing* (Version 4.3.2). R Project for Statistical Computing. https://www.R-project.org/

Siepe, B. S., Bartoš, F., Morris, T., Boulesteix, A.-L., Heck, D. W., & Pawel, S. (2023). *Simulation studies for methodological research in psychology: A standardized template for planning, preregistration, and reporting*. PsyArXiv Preprints. https://doi.org/10.31234/osf.io/ufgy6

Siepe, B. S., & Heck, D. W. (2025a). *Dynamic network multiverse.* [Preregistration, Code, Data, Supplementary Materials]. OSF. https://osf.io/xvrz5/

Siepe, B. S., & Heck, D. W. (2025b). *Network Multiverse Shiny App.* [Shiny App: Hosted]. Shinyapps.io. https://tinyurl.com/netshiny

Siepe, B. S., & Heck, D. W. (2025c). *Network Multiverse Shiny App.* [Shiny App: Download]. GitHub. https://tinyurl.com/git-shiny

Siepe, B. S., & Heck, D. W. (2025d). *Fork of Group Iterative Multiple Model Estimation for multiverse analysis.* [R package: GIMME modified package]. GitHub. https://github.com/bsiepe/mv-gimme

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour, 4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

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

Weermeijer, J., Lafit, G., Kiekens, G., Wampers, M., Eisele, G., Kasanova, Z., Vaessen, T., Kuppens, P., & Myin-Germeys, I. (2022). Applying multiverse analysis to experience sampling data: Investigating whether preprocessing choices affect robustness of conclusions. *Behavior Research Methods, 54*(6), 2981–2992. https://doi.org/10.3758/s13428-021-01777-1

Wright, A. G. C., Gates, K. M., Arizmendi, C., Lane, S. T., Woods, W. C., & Edershile, E. A. (2019). Focusing personality assessment on the person: Modeling general, shared, and person specific processes in personality and psychopathology. *Psychological Assessment, 31*(4), 502–515. https://doi.org/10.1037/pas0000617

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

Wright, A. G. C., & Zimmermann, J. (2019). Applied ambulatory assessment: Integrating idiographic and nomothetic principles of measurement. *Psychological Assessment, 31*(12), 1467–1480. https://doi.org/10.1037/pas0000685

Methodology is the official journal of the European Association of Methodology (EAM).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.