Original Article

Block-Wise Model Fit for Structural Equation Models With Experience Sampling Data

Julia Norget and Axel Mayer

Faculty of Psychology and Sport Science, Bielefeld University, Germany

Abstract: Common model fit indices behave poorly in structural equation models for experience sampling data which typically contain many manifest variables. In this article, we propose a block-wise fit assessment for large models as an alternative. The entire model is estimated jointly, and block-wise versions of common fit indices are then determined from smaller blocks of the variance-covariance matrix using simulated degrees of freedom. In a first simulation study, we show that block-wise fit indices, contrary to global fit indices, correctly identify correctly specified latent state-trait models with 49 occasions and N = 200. In a second simulation, we find that block-wise fit indices cannot identify misspecification purely between days but correctly rejects other misspecified models. In some cases, the block-wise fit is superior in judging the strength of the misspecification. Lastly, we discuss the practical use of block-wise fit evaluation and its limitations.

Keywords: structural equation modeling, fit indices, latent state-trait theory, experience sampling

https://econtent.hogrefe.com/doi/pdf/10.1027/2151-2604/a000482 - Monday, April 29, 2024 9:05:11 PM - IP Address:3.146.37.35

In psychological research, we often measure people's affect, behavior or cognition in different situations. Changes in measures from one occasion to another may reflect a change of the attribute in question, the different situations in which it was assessed, or be due to measurement error. With several measurement occasions, latent state-trait theory and its revised version (LST-R theory; Steyer et al., 1999, 2015) allows researchers to distinguish between occasion-specific (state residual) and stable (trait) influences on the observed attribute. State residuals reflect the influence of a specific situation and the person-situation interaction on the observed variable. A trait is an attribute of the person at the time of measurement (Steyer et al., 2015).

When we research states that fluctuate over short periods, experience sampling (ES) studies can be useful. In ES studies, participants respond about their behavior or thoughts several times a day during one or more weeks (Mehl et al., 2011), leading to large datasets. LST-R theory can also be applied to ES datasets. Eid et al. (2012) give an overview of models for ES data. These models include autoregressive effects to account for short time lags and can be defined in the LST-R framework (Eid et al., 2017).

There are multiple other approaches to assessing the (in)stability of constructs with structural equation models, for example, the single indicator STARTS model (Kenny & Zautra, 1995, 2001), the integrated state-trait model (Hamaker et al., 2007), the random intercept cross-lagged panel model (RI-CLPM; Hamaker et al., 2015), or multilevel approaches such as dynamic structural equation models (DSEM; e.g., Asparouhov et al., 2018; Zhang et al., 2008).

LST-R Theory

In this article, we focus on LST-R models for ES data. LST-R theory is an extension of classical test theory (CTT) for longitudinal data. While CTT can differentiate between person ("trait") effects and measurement error, LST-R theory also considers the influences of the situation and person-situation interaction. A revised version (LST-R theory; Steyer et al., 2015) recognizes that a person changes with experience and thus that traits can change over time.

Each observed variable (indicator) is denoted as Y_{it}, where i (i = 1, 2, 3,...) stands for the indicator and t (t = 1, 2, 3,...) for the time point. Each indicator can be decomposed into a latent state variable (τ_{it}) and measurement error. The latent state variables are defined as the expected value of Y_{it} given the person-at-time-t and the situation-at-time-t. The measurement error variable (ϵ_{it}) is the difference between Y_{it} and τ_{it}. The latent state variable is further decomposed into the latent trait variable (ξ_{it}) and the state residual variable (ζ_{it}). The latent trait variable is defined as the expected value of Y_{it} given the person-at-time-t. The state residual variable is the difference between the latent

© 2022 The Author(s) Distributed as a Hogrefe OpenMind article under the license CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1027/2151-2604/a000482 - Monday, April 29, 2024 9:05:11 PM - IP Address:3.146.37.35

state and the latent trait variables. Overall, we obtain the following equation:

$$Y_{it} = \xi_{it} + \zeta_{it} + \epsilon_{it}. \quad (1)$$

The latent trait variable represents the person-at-time-t-specific influence on the measurement. Since the person can change with experience, we could also call the trait variable an occasion-specific disposition. The state residual variable represents the influences of the situation and person-situation interaction.

Based on this decomposition, LST-R theory defines three important coefficients for each indicator. Consistency is the proportion of variance due to the trait variable: $\mathrm{Con}(Y_{it}) = \mathrm{Var}(\xi_{it})/\mathrm{Var}(Y_{it})$. Occasion-specificity is the proportion of variance due to the state residual variable: $\mathrm{Spe}(Y_{it}) = \mathrm{Var}(\zeta_{it})/\mathrm{Var}(Y_{it})$. Reliability is the sum of both, or in other words, it is the proportion of variance not due to unsystematic measurement error: $\mathrm{Rel}(Y_{it}) = 1 - \mathrm{Var}(\epsilon_{it})/\mathrm{Var}(Y_{it})$.

With these definitions alone, it is not yet possible to estimate an LST-R model. Additional assumptions about the equivalence of latent state and trait variables need to be made to obtain an identified model. For a model with a single trait variable, the most restrictive equivalence assumptions (state- and trait-equivalence), assume that the state and trait variables are measured on the same scale with the same intercept, meaning that the intercept is zero and factor loadings are fixed to one. The model equation with these assumptions is $Y_{it} = \theta + \zeta_t + \epsilon_{it}$. There is one single trait variable θ for all occasions and several occasion-specific state residual variables ζ_t, as can be seen in the path model in Figure 1A. For a detailed overview of the definitions and additional assumptions in LST-R theory, see Steyer and colleagues (2015).

Models With Autoregressive Effects

In ES studies, time intervals between measurements are very short. Measures taken close together in time are more similar than measures taken further apart, and autoregressive effects are common in ES data (Bolger & Laurenceau, 2013). Eid and colleagues (2012) therefore propose different LST models with autoregression, which can be defined in the framework of LST-R theory (Eid et al., 2017). Autoregressive paths are added at the level of occasion-specific residual variables. The latent state variable is decomposed into the latent trait variable and an occasion factor (OCC_{ij}). Autoregressive paths are added between these occasion factors. The OCC_{ij} variables have a residual, which is the state residual variable ζ_{ij}. This means that the occasion factors are the current state residual plus a linear combination of all previous state residuals. The first occasion factor is identical to the first state residual. A model with autoregression is depicted in Figure 1B. It is also possible to add autoregressive paths between the latent states, but for most short-term longitudinal studies, autoregression between occasion-specific residual variables seems more suitable (Stadtbäumer et al., 2021).

Indicator- and Day-Specific Traits

While the models described above included a single trait, the constructs explored in ES research are often more dynamic and a single trait across the entire measurement period is not always realistic. Eid and colleagues (2012) describe models with indicator- and day-specific traits. If the indicators in the model are not homogeneous (e.g., positive and negative valence) indicator-specific traits can capture the specific components which are not shared. Given that the indicators are supposed to measure the same construct, indicator-specific traits should correlate highly. Indicator-specific LST-R models can also include indicator-specific equivalence assumptions, meaning that we assume state- and trait-equivalence separately for the manifest variables of each indicator. When the construct in question is stable within days but less stable across the entire measurement period, it is also possible to include day-specific traits. The day-specific trait variables can capture within-day stability, while the correlation between traits gives an indication of between-day stability. Day-specific models can also have day-specific state- and trait-equivalence assumptions, meaning that equivalence is assumed within each day. Day-specific and indicator-specific traits can also be combined. Some path models of indicator-specific and day-specific models can be found in the Electronic Supplementary Material 1 (ESM 1, Figure E1), which illustrates the design of the simulation study.

Model Fit Evaluation

In LST-R models for ES data, it is difficult to estimate model fit. Fit indices are less reliable for models with many manifest variables: they show inflated χ^2-values with rejection rates of up to 100% for correctly specified models (Moshagen, 2012). This so-called model size effect is largely influenced by the number of manifest variables (p) and the sample size (N) (Shi et al., 2019). The number of free parameters (q) has a smaller influence. Moshagen (2012) found no influence of q on inflated Type I error rates, but Shi and colleagues (2019) found such an effect. However, with a large number of manifest variables ($p \geq 60$) Type I error rates are dramatically inflated, independent of q and even with very large sample sizes (N = 2,000) (Shi et al., 2019). The model size effect disappears asymptotically (i.e., when N approaches infinity).

© 2022 The Author(s) Distributed as a Hogrefe OpenMind article under the license CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0)

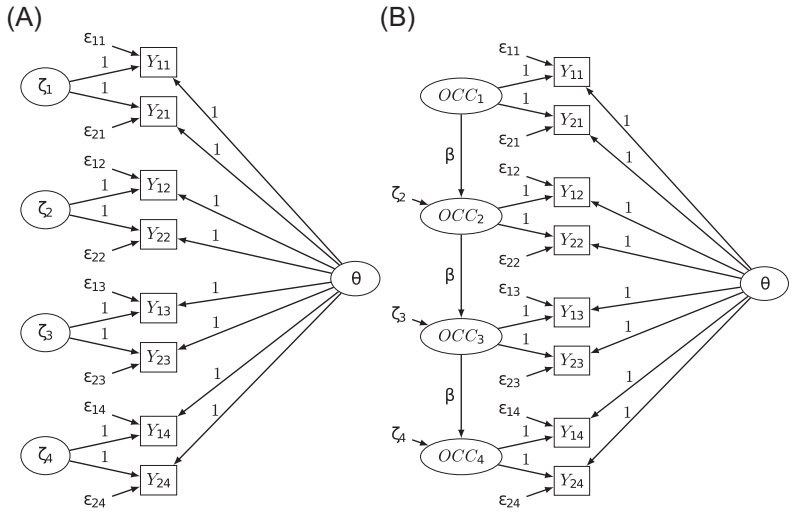

Figure 1. Path models of two LST-R single-trait models. (A) LST-R model with state residuals on the left side and a single trait on the right side. (B) LST-R model with autoregression.

There are different χ^2-corrections to counteract the model size effect, such as the ones by Bartlett (1950), Swain (1975), and Yuan and colleagues (2015). A comparison by Shi and colleagues (2018) showed that the correction by Yuan et al. (2015) performs best and results in acceptable Type I error rates, except with very large p (≥ 90) and small N (= 200). Yuan and colleagues (2015) multiply the empirical maximum likelihood χ^2 test statistic with the correction factor $e = [N - (2.381 + 0.361p + 0.006q)]/(N - 1)$.

Common fit indices such as Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA) are based on the χ^2-value and are also biased in larger models (Kenny & McCoach, 2003; Shi et al., 2019). In the case of misspecified dimensionality, CFI and TLI worsen with more manifest variables but improve when the data-generating model includes residual correlations which are omitted in the analysis. RMSEA values decrease with more manifest variables (Kenny & McCoach, 2003; Savalei, 2012; Shi et al., 2019). Guidelines for interpreting these values are based on studies with smaller models (e.g., 15 manifest variables in the study by Hu & Bentler, 1999). For models with ES data, relying on these fit indices may lead to the incorrect rejection of acceptable models.

Local Fit Evaluation

The bias of fit indices is associated with model size, so a more local evaluation of smaller model elements seems intuitive. Maydeu-Olivares and Shi (2017) suggest that the local source of misspecification can be visually detected through areas (for misspecified trait dimensionality) or rows (for misspecified secondary loadings) of high residual correlations. However, with two items measured at only 14 time points, a residual correlation matrix has 784 entries, making it difficult to detect meaningful patterns. The number of large residual correlations also increases with matrix size. If there are no obvious patterns, this approach may not tell us if model rejection is due to model size or legitimate model misfit and may not be helpful for judging the fit of ES models.

Another approach to local fit evaluation is testing individual implications of the proposed model. Thoemmes and colleagues (2018) suggest conditional independence test for implications of the model structure and tetrad tests if latent variables are involved. The number of conditional independence constraints equals the degrees of freedom (*df*), and the number of tetrad constraints is large even with few latent variables. For models with ES data, there will be thousands of tests, making it difficult to derive what they imply for the model structure.

Recently, Rosseel and Loh (2021) presented the Structural After Measurement (SAM) framework, where parameters of the measurement part are estimated first, followed by the parameters of the structural part. The measurement part can be estimated as (1) a single measurement block containing all latent variables, (2) separate measurement blocks for each latent variable, or (3) several measurement blocks which can contain more than one latent variable. There may not be equality constraints, cross-loadings, or correlated residuals between indicators in different blocks. Fit indices are derived for each measurement block and the structural part. A special case of SAM is step-wise factor score regression, where the measurement models of each latent variable are estimated independently, and their relationships are modeled with factor scores (Devlieger, 2019). For ES LST-R models, measurement models with two or three indicators are too small for factor score regression or option 2 of the SAM framework, and option 3 does not work for models with a single trait, indicator-specific traits,

https://econtent.hogrefe.com/doi/pdf/10.1027/2151-2604/a000482 - Monday, April 29, 2024 9:05:11 PM - IP Address:3.146.37.35

© 2022 The Author(s) Distributed as a Hogrefe OpenMind article
under the license CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1027/2151-2604/a000482 - Monday, April 29, 2024 9:05:11 PM - IP Address:3.146.37.35

measurement invariance over time or other equivalence assumptions. The first option, however, is not recommended and offers little benefit over SEM.

While local fit assessment has typically been recommended as a follow-up analysis, Rosseel and Loh (2021) show that it can also be useful as an alternative to global fit evaluation. Unfortunately, for the evaluation of most ES LST-R models, the SAM framework provides little added benefit. In this article, we will thus show that a new approach to local fit assessment can be a viable alternative to global evaluation for ES LST-R models.

We propose an approach where the full variance-covariance matrix is first estimated for the global model based on all postulated relationships. Then, local versions of fit indices are determined for each day (or other blocks) based on the global variance-covariance matrix using simulated block-wise *df*. This approach can take all kinds of relationships across measurement models and days into account and provides familiar fit indices. We will first show how local block-wise fit indices can be estimated. We then show in two simulation studies under which conditions they provide a more reliable fit assessment than global fit measures and discuss the implication of our results for the evaluation of large SEMs.

Block-Wise Model Fit Indices

In the past decades, a variety of fit indices have been developed to examine how well a theoretical model is supported by empirical data. Some of the most common indices are the χ^2, CFI, TLI, and RMSEA. In this section, we explain how these indices can be computed for blocks (e.g., days) of LST-R models for ES data in three steps: (1) estimating the overall model, (2) extracting blocks from $\hat{\Sigma}$ and S, and (3) calculating fit indices from these blocks.

First, the model including all latent constructs and postulated relationships is specified and estimated with maximum likelihood, yielding a model implied variance-covariance matrix $\hat{\Sigma}$. In the second step, a number of substantively meaningful blocks is chosen, such as the days in an ES study. All manifest variables are uniquely associated with one block, and all blocks contain the same number of manifest variables $p_k = \frac{p}{K}$, where p is the total number of manifest variables and K is the number of blocks. We use the subscript k for all block-specific parameters, with $k = 1,\ldots, K$. Then, the (co)variances of the manifest variables of each block are extracted from $\hat{\Sigma}$ and S. This results in K $p_k \times p_k$ model-implied ($\hat{\Sigma}_k$) and observed (S_k) (co)variance matrices (i.e., $\hat{\Sigma}_1$ for block 1, etc.). The block-wise matrices $\hat{\Sigma}_k$ and S_k must be invertible so that block-wise χ^2_k-values can be determined. In step three, model fit indices are determined for each block based on $\hat{\Sigma}_k$ and S_k with the regular formulas adapted for block-wise use.

This block-wise approach can be applied to large LST-R or other longitudinal models, where we can identify a substantively meaningful number of blocks. The block-wise approach allows for a day-specific evaluation of models that include restrictions across days, such as a single trait or measurement invariance over time.

Block-Wise χ^2

In a structural equation model, the χ^2-test evaluates the discrepancy between $\hat{\Sigma}$ and S, with the null hypothesis that $\hat{\Sigma}$ is identical to the (co)variances in the population from which the sample is drawn. An insignificant χ^2 value at an α-level of .05 is often used as an indicator of good fit. The χ^2-value is the product of the fitting function and sample size. The most common estimator to minimize the fitting function is the maximum likelihood (Bollen, 1989). Adapted for block-wise use, we get the formula:

$$\chi^2_k = \big(\log|\hat{\Sigma}_k| + \mathrm{tr}(\hat{\Sigma}_k^{-1}S_k) - \log|S_k| - p_k + (\bar{x}_k - \hat{\mu}_k)^T\hat{\Sigma}_k^{-1}(\bar{x}_k - \hat{\mu}_k)\big)\cdot(N-1), \quad (2)$$

where p_k is the number of observed variables per block, $\bar{x}_k$ the vector of sample means, and $\hat{\mu}_k$ the vector of model-implied means, both for the items in block k. N is the sample size. Although the sample estimates for χ^2 and other fit indices include $N - 1$ in the formulas, both lavaan and MPlus use N instead. For the sake of consistency, we therefore used N in the computations for the simulation study. With large sample sizes, the χ^2-test yields significant p-values even for models with a minor misfit. This is one of the reasons which have inspired the development of different fit indicators including RMSEA, CFI, and TLI.

Block-Wise Degrees of Freedom

In order to test the null hypothesis and to calculate other fit indices, we need the *df*. In SEM, *df* are the difference between the number of empirical parameters (means, variances, and covariances of the manifest variables) and estimated parameters. All estimated parameters are involved in computing the implied (co)variances in $\hat{\Sigma}$. However, not all estimated parameters are uniquely associated with only one block. Time-invariant factor loadings, or the variance and mean of a single trait, affect the calculation of co(variances) in more than one $\hat{\Sigma}_k$ matrix.

© 2022 The Author(s) Distributed as a Hogrefe OpenMind article under the license CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1027/2151-2604/a000482 - Monday, April 29, 2024 9:05:11 PM - IP Address:3.146.37.35

Since it is unclear how estimated parameters can be split among blocks, we suggest simulating block-wise df_k. Under the null hypothesis, empirical χ^2 values follow a χ^2-distribution with $df = E(\chi^2)$. Thus, we can approximate the df by simulating many datasets from the true model and computing the mean of the block-wise χ^2-values $M(\chi_k^2)$.

Simulation of Block-Wise df_k

To test this behavior, we simulated data for 4 ES LST-R models (single-trait and day-specific model with and without autoregressive effect, with 7 occasions per day and 2 indicators per occasion), differing numbers of days (1, 2, 7), and sample sizes (200, 10,000). For each condition, 1,000 datasets were created and analyzed with the data-generating model. Block-wise χ_k^2 values were calculated for 2 or 7 blocks, conferring to the number of days. We examined global df, $M(\chi^2)$, and the distribution of the χ^2 values, as well as block-wise $M(\chi_k^2)$ and the distribution of all χ_1^2, that is, the χ^2 values of the first block.

Simulation results show that global $M(\chi^2)$ for 2 or 7 days and N = 200 are overestimated, for example, the autoregressive model with day-specific traits for 7 days has df = 4,852, but $M(\chi^2)$ = 5,965.70. With N = 10,000, the χ^2 inflation almost disappeared (for the same model: $M(\chi^2)$ = 4,867.79). For models with 1 day, $M(\chi^2)$ closely resembles df (e.g., for autoregressive model with day-specific traits, N = 200: df = 109, $M(\chi^2)$ = 112.62). We also computed Kolmogorov-Smirnov (KS) distances, which express the maximum difference on a scale from 0 to 1 between the observed distribution of the χ^2-values and their theoretical distribution with df degrees of freedom and checked which proportion of simulated χ^2-values fall within each decile of a χ^2-distribution with $df = M(\chi^2)$. Both approaches indicate that the simulated values are χ^2-distributed with $df = M(\chi^2)$. A table with all results is included in ESM 2.

Overall, the block-wise χ_k^2 values are approximately χ^2-distributed with $df = M(\chi_k^2)$. Based on these results, we recommend simulating datasets based on $\hat{\Sigma}$ and the model-implied means with the actual sample size, computing block-wise χ^2 values for all datasets with Formula 2 and using $M(\chi_k^2)$ as an approximation of block-wise df_k. This approximation can then be used for the calculation of other fit indices. As an alternative, one can directly simulate the distribution of the test statistic and use its empirical distribution to test for significance. However, in this case, the other fit indices cannot be computed.

Block-Wise Absolute Fit Indices

Absolute fit indices such as the RMSEA can better be understood as measures of misfit, where small values indicate little misfit. The RMSEA is based on the χ^2 statistic but corrects for model complexity. The block-wise version for each block k can be calculated as follows:

$$\mathrm{RMSEA}_k = \sqrt{\max\left(0, \frac{\chi_k^2 - df_k}{df_k \cdot (N-1)}\right)}. \quad (3)$$

Although fit indices were developed to judge the extent of model (mis)fit, they are also affected by other influences such as the strength of factor loadings (Heene et al., 2011) or, as discussed before, the number of manifest variables. Common rules of thump should thus be used with great caution, and several indices need to be taken into consideration to judge model fit. According to Hu and Bentler (1999), RMSEA values < .06 indicate good fit. Another common rule of thumb is that RMSEA values $\leq$ .05 indicate close fit, and values $\leq$ .08 indicate reasonable fit (Browne & Cudeck, 1992). Another absolute fit index is the SRMR. This fit index does not depend on the χ^2-test statistic, but we provide information on block-wise SRMR in ESM 1.

Block-Wise Incremental Fit Indices

Incremental fit indices (e.g., CFI and TLI; Bentler, 1990; Tucker & Lewis, 1973) do not use the χ^2-statistic directly but compare the proposed model to the worst possible (null) model. The null model only includes variances for the observed variables, but no relationships are modeled. For block-wise CFI and TLI, the null model is computed for each block based on the manifest variables from the block in question, p_k. The block-wise null model has $p_k(p_k - 1)/2$ df, and a χ^2-value is estimated according to Formula 2. The block-wise Bentler Comparative Fit Index (CFI_k) can then be calculated as follows (Shi et al., 2019):

$$\mathrm{CFI}_k = \frac{\max\left(d_k(\text{Null Model}), 0\right) - \max\left(d_k(\text{Proposed Model}), 0\right)}{\max\left(d_k(\text{Null Model}), 0\right)}, \quad (4)$$

where $d = \chi_k^2 - df_k$ for the null and proposed model. CFI_k can range between 0 and 1. The block-wise TLI_k is calculated as follows:

$$\mathrm{TLI}_k = \frac{\chi_k^2/df_k(\text{Null Model}) - \chi_k^2/df_k(\text{Proposed Model})}{\chi_k^2/df_k(\text{Null Model}) - 1}. \quad (5)$$

Since the TLI is not normed, TLI > 1 or negative values are possible. For both CFI and TLI, values $\geq$ .97 indicate a good fit between model and data, but values between .95 and .97 are considered acceptable (Hu & Bentler, 1999; Schermelleh-Engel et al., 2003).

© 2022 The Author(s) Distributed as a Hogrefe OpenMind article
under the license CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1027/2151-2604/a000482 - Monday, April 29, 2024 9:05:11 PM - IP Address:3.146.37.35

Simulation Studies

We have shown that common advice for judging model fit is not suitable for models with many manifest variables (e.g., Kenny & McCoach, 2003; Moshagen, 2012; Savalei, 2012; Shi et al., 2019) and have proposed block-wise evaluation. In order to demonstrate that block-wise evaluation is a viable alternative for models with many manifest variables, we conduct two simulation studies. We first simulate correctly specified data for ES LST-R models and evaluate the effect of model size and sample size on global and block-wise fit indices. Here, we expect that global fit indices will incorrectly reject models with more days and a smaller sample size, which is common in ES studies. We expect that block-wise fit indices can correctly identify these models. In a second simulation study, we generate data for the same ES LST-R models but with different misspecifications and evaluate the effects of model size, sample size, and misspecification on global and block-wise fit. Block-wise fit evaluation is based on (co)variances within each day, so we expect that block-wise fit indices will correctly reject models which are misspecified within days, but fail to identify models which are misspecified purely between days.

Study 1: Correctly Specified Models

Method

In the first simulation study, model fit is evaluated for two different ES LST-R models, with varying model and sample size. Overall, we have a 2 (models) × 2 (model size) × 2 (sample size) design. The analysis models are (1) an autoregressive multistate–singletrait model, where a single trait is assumed across all measurements, and (2) an autoregressive multistate–multitrait model with day-specific traits. We included both models because the multistate–singletrait model is most common in applications of LST(-R) theory, but the model with day-specific traits is suitable for many applications with ES data. We did not have different hypotheses for these two models. The models include 2 ndicators for each occasion and 7 occasions for each day. LST-R models can include more indicators, but ES studies typically include as few questions as possible to keep the strain on participants low. Models with 2–3 indicators thus seem realistic for ES LST-R models. Both models have η-equivalent and θ-equivalent measures within each day. This implies that all factor loadings are set to 1, all intercepts are 0, and all (state) residual variances are equal within each day ($\mathrm{Var}(\epsilon_t) = \mathrm{Var}(\epsilon_u)$ and $\mathrm{Var}(\zeta_t) = \mathrm{Var}(\zeta_u)$, $t \neq u$). Autoregressive effects are restricted to be equal between all occasion-specific factors. Parameters in the data-generating population models were $\mathrm{Var}(\zeta_t) = .3$, $\mathrm{Var}(\theta_{(u)}) = .3$ for the trait in the single-trait model and all day-specific trait variables θ_u in the day-specific traits model, $\mathrm{Var}(\epsilon_{it}) = .4$, $\mathrm{Cov}(\theta_u, \theta_V, u \neq v) = 0.21$ (corresponding to a correlation of $r = .7$), $M(\theta_{(u)}) = 2.2$, and autoregressive effects $\beta = .1$. This implies equal occasion-specificity and consistency, with item reliabilities between .60 and .61. These values are approximately based on an empirical application with ratings of perceived conflict of interest in social situations (Norget et al., 2021). The trait and state residual variances are adjusted to be equal because many constructs assessed in longitudinal studies have both stable and occasion-specific aspects (Geiser, 2021). Please refer to Figure E1 (ESM 1) for path models of the single-trait and the day-specific traits model. Data was generated for models with 2 or 7 days (i.e., 28 or 98 manifest variables) and sample sizes of 200 or 1,000. Typical data situations in ES studies include sample sizes around or smaller than $N = 200$ and data collection on several days, often one or two weeks. For each condition of the study, we estimate global fit indices as well as block-wise fit indices for each day. For comparison, we also simulated global *df* in the same way as we described for the block-wise *df* and computed all global fit indices using these simulated *df*. Additionally, we computed the Yuan et al. (2015) corrected χ^2-estimates to compare rejection rates and χ^2/df-ratios.

For each of the 8 conditions, 500 datasets were generated and analyzed. Block-wise (and global) *df* were simulated for the first dataset in each condition. These estimates were then used for all 500 datasets in the same condition. In a test phase, we simulated the block-wise *df* several times for the same condition and found very small deviations between the estimates. The simulation study was conducted in R (R Core Team, 2020; RStudio Team, 2019) using the packages SimDesign (Chalmers & Adkins, 2020), lavaan (Rosseel, 2012), lsttheory (Mayer, 2020), and MASS (Venables & Ripley, 2002). We discuss χ^2-rejection rates at $\alpha = .05$, KS distances and mean CFI, TLI, and RMSEA values for the different conditions and point out the most important aspects of this visual analysis. Results are shown in Figure 2 and Figures E2 and E3 (ESM 1).

Results

There were (almost) no differences between the two models. We present the results for the day-specific model here and provide results for the other model in Figure E2 (ESM 1). We will refer to the global fit indices as implemented in common SEM software as "global" χ^2, CFI, TLI, and RMSEA. Estimates based on simulated global *df* are "simulated global" values, and Yuan and colleagues (2015) corrected values are "Yuan-corrected".

χ^2-Rejection and Kolmogorov-Smirnov Distance

For correctly specified models, χ^2-rejection rates at $\alpha = .05$ should be around 5%. As shown in Figure 2A, χ^2-rejection

© 2022 The Author(s) Distributed as a Hogrefe OpenMind article under the license CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0)

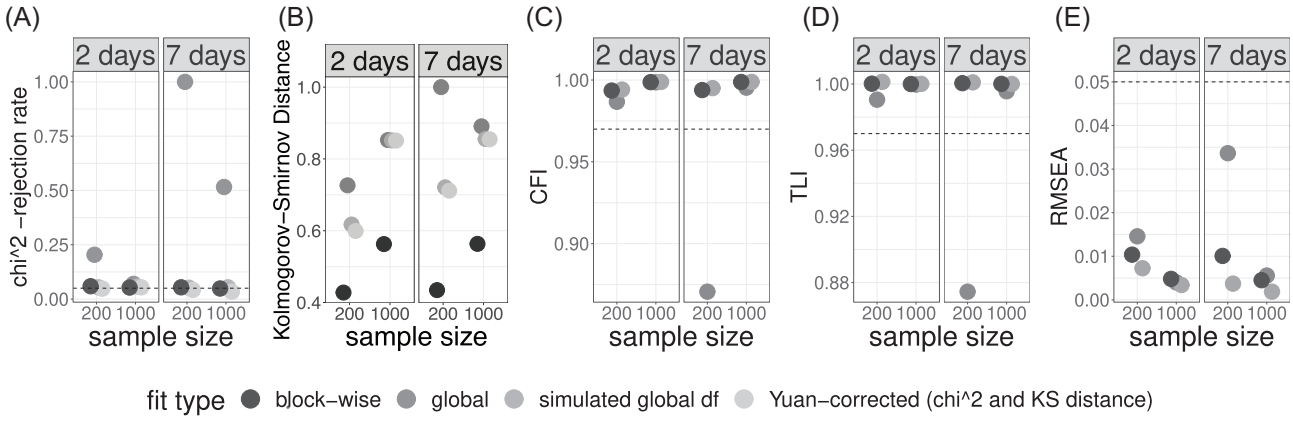

Figure 2. Results of Study 1 for the day-specific model (single-trait figures are included in Figure E2 (ESM 1). (A) χ^2-rejection rates at α = .05; (B) KS distance single-trait model; (C) CFI values; (D) TLI values; (E) RMSEA values.

rates for global evaluation, with globally simulated *df*, and Yuan-corrected χ^2 for models with 2 days and N = 1,000 are close to the expected rejection rate. With smaller sample sizes and more days, global rejection rates increase up to 100% for models with 7 days and N = 200. The KS distances (Figure 2B) show that the distribution of global, simulated global, and Yuan-corrected χ^2 values differs more strongly from their theoretical distribution than block-wise χ^2_k. Global χ^2 values are most strongly overestimated. Again, this difference is especially large for models with 7 days and N = 200. Overall, global χ^2, as implemented in most software, highly overestimates the test statistic and too often rejects correctly specified models, especially in the most likely data scenario in ES studies, while block-wise χ^2_k performs much better.

Comparative Fit Index and Tucker-Lewis Index

For CFI and TLI most models yield estimates $\geq$ .97, indicating a good model fit. However, we can see in Figures 2C and 2D that global CFI and TLI indicate worse fit for smaller samples (N = 200 vs. N = 1,000) and for larger models (7 vs. 2 days). In the most likely data scenario in ES studies (7 days and N = 200) global indices reject the correctly specified model (CFI = .88, TLI = .88). However, simulated global CFI and TLI, and block-wise CFI_k and TLI_k correctly indicate a good fit in all cases.

Root Mean Square Error of Approximation

All types of $RMSEA_{(k)}$ correctly identify a good fit in all four scenarios (see Figure 2E). For models with N = 1,000, global, simulated global, and block-wise $RMSEA_{(k)}$-values are very small (.003–.006). For models with N = 200, $RMSEA_{(k)}$ values are slightly higher, and global values indicate worse fit than simulated global or block-wise values. RMSEA indicates good fit in all scenarios, but global RMSEA is noticeably worse for N = 200 and 7 days (.034) compared to all other conditions. Block-wise $RMSEA_{(k)}$ clearly indicates a better fit in this case. Simulated global values indicate better fit in all scenarios. Since all $RMSEA_{(k)}$ correctly indicate a good fit, block-wise evaluation may offer less benefit over global evaluation in the case of RMSEA compared to other fit indices. However, block-wise and simulated global $RMSEA_{(k)}$ still correctly indicate a better fit than global RMSEA.

Discussion

Overall, correctly specified models were correctly identified by block-wise χ^2_k, CFI_k, TLI_k, and $RMSEA_k$, but not always by their global counterparts. Simulated global indices behave similarly to block-wise indices. The biases in global fit are in line with previous simulation studies (Kenny & McCoach, 2003; Moshagen, 2012). Especially in the most likely data scenario with experience sampling data, models for 7 days (49 occasions), and sample sizes of N = 200, block-wise fit evaluation seems to offer a good alternative to global evaluation.

Study 2: Misspecified Models

Method

While Study 1 showed that block-wise fit correctly identifies correctly specified models, it is also important to consider under which conditions block-wise fit can correctly reject misspecified models. Since block-wise fit is based on the (co)variances of each block, we expect that misspecifications within blocks should be identified correctly, while misspecifications purely between blocks should be undetectable for block-wise fit indices. In Study 2, we generated data with different misspecifications in a 2 (models) × 2 (model size) × 2 (sample size) × 6 (misspecification) design. Analyzing ES LST-R models were the same as

https://econtent.hogrefe.com/doi/pdf/10.1027/2151-2604/a000482 - Monday, April 29, 2024 9:05:11 PM - IP Address:3.146.37.35

© 2022 The Author(s) Distributed as a Hogrefe OpenMind article
under the license CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1027/2151-2604/a000482 - Monday, April 29, 2024 9:05:11 PM - IP Address:3.146.37.35

described in Study 1: a single-trait and a day-specific model. Again, we generated data for either 2 or 7 days, with sample sizes of 200 or 1,000. Global CFI and TLI worsen with more manifest variables and misspecified dimensionality, but improve with omitted residual correlations (Shi et al., 2019), so we included models with omitted residual correlations between and within days, as well as structural misspecifications similar to those in Shi et al. (2019). Pathmodels are provided in Figure E1 (ESM 1). The misspecified models include (1) small or (2) large residual correlations between days, that is, both items measured on occasion 1, 2, and so forth on each day are correlated with the same item measured on the same occasion on other days. Residual correlations are small (r = .15) or large (r = .40); (3) small (r = .15) or (4) large (r = .40) residual correlations within days, that is, the residuals of item 1 on all occasions within the same day are correlated, and likewise for item 2; (5) small or (6) large structural error, that is, each trait is split into two indicator-specific traits in the population model, with correlations of r = .90 (small error) or r = .60 (larger error). Other population values are identical to Study 1. We expected that block-wise fit would detect the structural error and the omitted residual correlations within days but not between days.

For the χ^2, we discuss rejection rates at α = .05 and provide further analysis for the ratio between χ^2 and *df*. χ^2/df = 1 indicates perfect fit. For CFI, TLI, and RMSEA, we analyze their global, simulated global, and block-wise values using analyses of variance (ANOVAs) with the respective fit index as the outcome and the four predictors: (1) model (single-trait/day-specific traits), (2) number of days (2/7), (3) sample size (200/1,000), and (4) the type of the fit index (global/simulated global/block-wise). All predictors are coded as factors, and we use Type III sum of squares and sum to zero contrasts. We used the R package car (Fox & Weisberg, 2019) to fit the ANOVAs and the package effectsize (Ben-Shachar et al., 2020) for effect sizes. Normal distribution of the residuals and variance homogeneity were visually checked, and the assumptions were met sufficiently.

Results

χ^2-Rejection Rates and χ^2/df Ratios

The χ^2-rejection rates at α = .05 are displayed in Figure 3A. Colored figures can be found in Figures E4a–E4e (ESM 1). Most models have rejection rates of around 100%. Block-wise χ^2 cannot detect the omitted residual correlations between days and incorrectly indicates perfect fit (i.e., rejection rates around 5%). For small misspecifications and N = 200, block-wise, and to a lesser degree also simulated global and Yuan-corrected $\chi^2_{(k)}$ sometimes have rejection rates notably lower than 100%; global χ^2 for N = 200 and 2 days as well, but with higher rejection rates than the other types.

The ANOVA for the χ^2/df ratio revealed substantial main effects of the misspecification, $F(5, 179{,}876) = 412{,}102$, $p < .001$, $\eta^2 = .26$, and sample size, $F(1, 179{,}876) = 952{,}053$, $p < .001$, $\eta^2 = .12$, meaning that all χ^2/df ratios are for a large part similarly affected by these two influences. Since we are more interested in the differences between the types of fit, we will focus on interaction effects with the type of fit measure. A table with the complete ANOVA results is included in Table E1 (ESM 1).

First, there is considerable two-way interaction between the type of fit measure and the misspecification, $F(15, 179{,}876) = 105{,}156.1$, $p < .001$, $\eta^2 = .20$. Block-wise χ^2_k/df_k ratios are higher (i.e., indicate worse fit) than global, simulated global, or Yuan-corrected ratios for models with large structural misspecification and omitted residual correlations within days. However, block-wise χ^2_k/df_k ratios indicate a perfect fit for the models with omitted residual correlations between days.

Second, there is an interaction effect between type of fit and sample size, $F(3, 179{,}876) = 578{,}95.2$, $p < .001$, $\eta^2 = .02$). Looking at the types of fit separately, the effect of sample size remains substantial for all, with lower ratios for N = 200 than N = 1,000. The difference between sample sizes is larger for block-wise ratios ($M_{1000} - M_{200} = 3.48$) than for global ($M_{1000} - M_{200} = 1.72$), simulated global ($M_{1000} - M_{200} = 1.84$) and Yuan-corrected ($M_{1000} - M_{200} = 1.84$) ratios.

Furthermore, there is a 3-way interaction between type of fit, misspecification, and sample size, $F(15, 179{,}876) = 58{,}386.4$, $p < .001$, $\eta^2 = 0.11$. Figure 3B shows that for block-wise χ^2_k/df_k, and to a lesser extend for global, simulated global, and Yuan-corrected ratios, the difference between N = 200 and N = 1,000 is larger with strongly misspecified models compared to their less strongly misspecified counterparts.

Contrary to our expectations, there was no noteworthy interaction between the type of fit and the number of days, $F(3, 179{,}876) = 20{,}434.2$, $p < .001$, $\eta^2 = .008$, or main effect of the number of days, $F(1, 179{,}876) = 107{,}115.8$, $p < .001$, $\eta^2 = .01$.

Comparative Fit Index and Tucker-Lewis Index

The results for CFI and TLI barely differ, and results are reported together. A figure with the TLI results is included in Figure E4d (ESM 1). Most effects are significant in the ANOVAs, and we focus on those with notable effect sizes. All main effects, except for the effect of the model (single-trait vs. day-specific), are noteworthy and interact with the type of fit. We will focus on these interactions here since we are mostly interested in how global and block-wise fit are differently affected by other influences. Full results are included in Tables E2 and E3 (ESM 1).

© 2022 The Author(s) Distributed as a Hogrefe OpenMind article under the license CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1027/2151-2604/a000482 - Monday, April 29, 2024 9:05:11 PM - IP Address:3.146.37.35

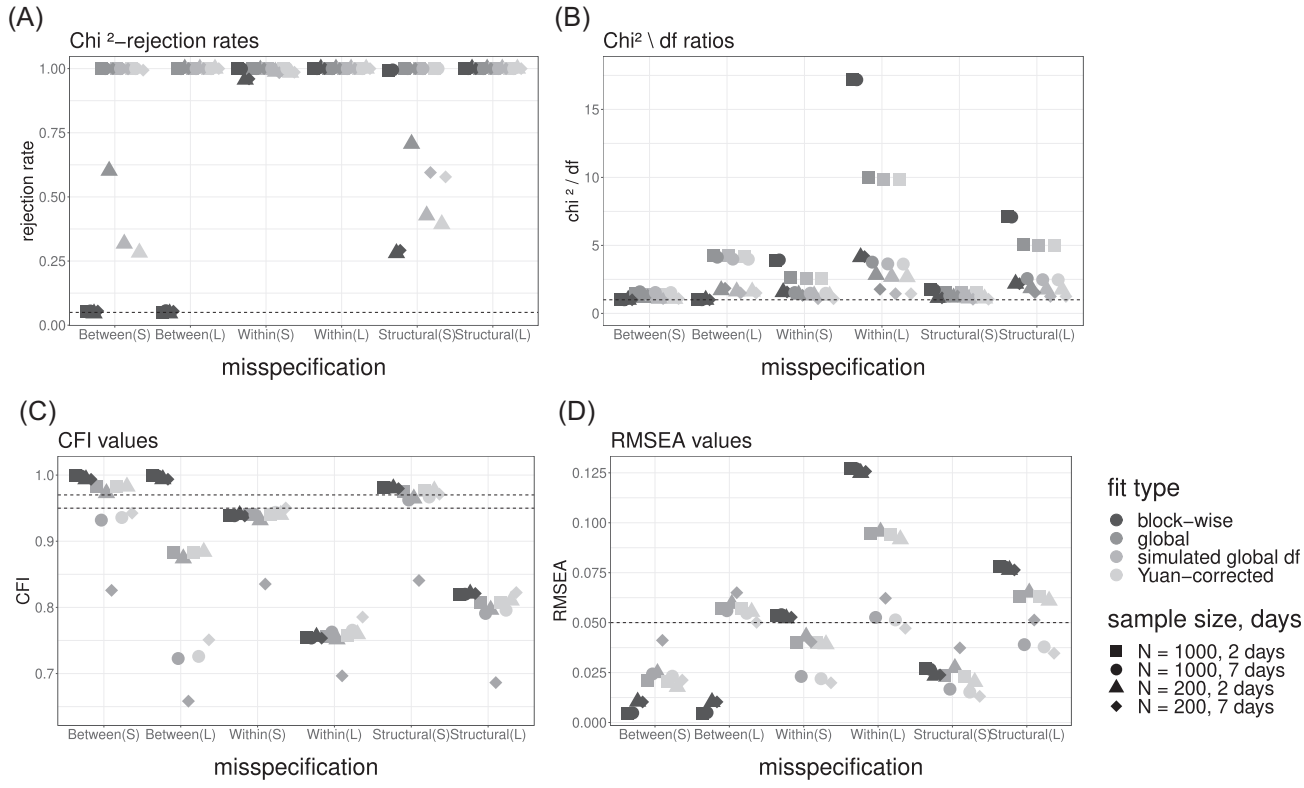

Figure 3. Overview of results of Study 2 for each fit index. There are six misspecifications on the *x*-axis. Between(S): omitted residuals correlations between days (r = .15); Between(L): likewise but with r = .40; Within(S): omitted residual correlations within days (r = .15); Within(L): likewise but with r = .40, Structural(S): data is generated with correlated indicatorspecific traits (r = .90); Structural(L): likewise but with r = .60. (A) χ^2-rejection rates at α = .05; (B) χ^2/df ratios; (C) CFI values (TLI values look almost identical); (D) RMSEA.

Most notably, there is a substantial interaction between the misspecification and type of fit (CFI: $F(10, 155{,}903) = 36{,}904.1$, $p < .001$, $\eta^2 = .13$; TLI: $F(10, 155{,}903) = 38{,}277.6$, $p < .001$, $\eta^2 = .11$). This effect is largely due to the models with omitted residual correlations between days. Here, block-wise CFI_k and TLI_k indicate perfect fit, while global and simulated global CFI and TLI can identify the misspecification.

The number of days also interact with the type of fit (CFI: $F(2, 155{,}903) = 42{,}835.7$, $p < .001$, $\eta^2 = .03$; TLI: $F(2, 155{,}903) = 72{,}257.6$, $p < .001$, $\eta^2 = .04$). Global CFI and TLI values are lower for models with 7 than 2 days (CFI: $t(23{,}394) = 67.04$, $p < .001$, $d = .87$; TLI: $t(22{,}675) = 77.21$, $p < .001$, $d = 1.00$), to a lesser extend this is also true for simulated global indices (CFI: $t(23{,}783) = 24.44$, $p < .001$, $d = .32$; TLI: $t(23{,}666) = 24.58$, $p < .001$, $d = .32$) but for block-wise CFI_k and TLI_k there is no notable difference between 2 and 7 days (CFI: $t(38{,}858) = 0.60$, $p = .55$; TLI: $t(38{,}741) = 0.03$, $p = .98$).

There is a smaller interaction between the sample size and type of fit (CFI: $F(2, 155{,}903) = 19{,}282.2$, $p < .001$, $\eta^2 = .01$; TLI: $F(2, 155{,}903) = 33{,}428.6$, $p < .001$, $\eta^2 = .02$). Block-wise CFI_k and TLI_k are barely affected by sample size (CFI_k: $t(107{,}980) = -2.66$, $p = .008$, $d = -0.02$; TLI_k: $t(107{,}980) = 4.07$, $p < .001$, $d = 0.02$), and the effect on simulated global indices is also small (CFI: $t(23{,}939) = 6.63$, $p < .001$, $d = 0.09$; TLI: $t(23{,}326) = 18.45$, $p < .001$, $d = 0.24$). Here, models with $N = 200$ fit better than with $N = 1{,}000$. Global CFI and TLI are generally worse for smaller samples (CFI: $t(23{,}739) = -40.2$, $p < .001$, $d = -0.52$) TLI: $t(23{,}611) = -40.2$, $p < .001$, $d = -0.52$).

Type of fit also interacts with number of days and misspecification (CFI: $F(10, 155{,}903) = 6{,}071.9$, $p < .001$, $\eta^2 = .02$; TLI: $F(10, 155{,}903) = 7{,}123.1$, $p < .001$, $\eta^2 = .02$). The interaction between days and misspecification remains noteworthy for global (CFI: $F(5, 23{,}988) = 1{,}965.8$, $p < .001$, $\eta^2 = .06$), and simulated global (CFI: $F(5, 23{,}988) = 6{,}697.9$, $p < .001$, $\eta^2 = .09$) but not for block-wise fit (CFI: $F(5, 107{,}988) = 3.8175$, $p = 002$, $\eta^2 < .001$). Figure 3C shows that especially for models with omitted residual correlations between days, global and simulated global CFI and TLI for 7 days (circle and diamond shape) is smaller than for 2 days (square and triangle). This difference is smaller for other misspecifications.

Another interesting three-way interaction is between the type of fit, number of days, and sample size (CFI:

© 2022 The Author(s) Distributed as a Hogrefe OpenMind article
under the license CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1027/2151-2604/a000482 - Monday, April 29, 2024 9:05:11 PM - IP Address:3.146.37.35

$F(1, 155{,}903) = 13{,}693.5$, $p < .001$, $\eta^2 = .01$; TLI: $F(2, 155{,}903) = 21{,}879.2$, $p < .001$, $\eta^2 = .01$). This effect can easily be understood when we look at Figure 3C: For $N = 200$ and 7 days the global values (i.e., medium-gray diamond) are systematically lower than global and block-wise values for other combinations of the three predictors.

Root Mean Square Error of Approximation

Most effects on the RMSEA are statistically significant, and we only discuss those with notable effect sizes. Complete results are included in Table E4 (ESM 1). In terms of main effects, the misspecification accounts for the majority of variance in all $\text{RMSEA}_{(k)}$ values, $F(5, 155{,}903) = 192{,}939.9$, $p < .001$, $\eta^2 = .53$, and there is a small main effect of the number of days, $F(1, 155{,}903) = 27{,}321.8$, $p < .001$, $\eta^2 = .01$.

Again, the strongest interaction is between the type of fit and misspecification, $F(10, 155{,}903) = 53{,}426.9$, $p < .001$, $\eta^2 = .29$. Figure 3D shows that block-wise RMSEA_k generally indicates worse fit than global and simulated global RMSEA, except in the case of omitted residual correlation between days, which cannot be detected by block-wise fit.

Additionally, there is an interaction between type of fit, misspecification, and the number of days, $F(10, 155{,}903) = 31{,}80.7$, $p < .001$, $\eta^2 = .02$. The interaction between misspecification and number of days remains notable for global RMSEA, $F(5, 23{,}988) = 6{,}241$, $p < .001$, $\eta^2 = .14$, and simulated global RMSEA, $F(5, 23{,}988) = 11{,}129$, $p < .001$, $\eta^2 = .12$, but there is no interaction for block-wise RMSEA_k, $F(5, 107{,}988) = 2.06$, $p = .67$. A look at the medium- and light-gray shapes in Figure 3D reveals that (simulated) global RMSEA indicates better fit for models with 7 than 2 days in the case of omitted residual correlations within days or large structural misspecification. The figure also shows that global and simulated RMSEA tend to assess strongly misspecified models with 7 days as acceptable, while block-wise fit can identify them as fitting badly.

There is also a small interaction effect of the type of fit with number of days, $F(1, 155{,}903) = 11{,}240.8$, $p < .001$, $\eta^2 = .01$. While the RMSEA values with 7 days are lower than with 2 day for global, $t(19{,}600) = 32.44$, $p < .001$, $d = 0.42$, and simulated global RMSEA, $t(19{,}234) = 58.15$, $p < .001$, $d = 0.75$, there is no difference between the number of days for block-wise RMSEA_k, $t(38{,}839) = -0.32$, $p = .75$.

Discussion

In general, all fit types indicate a less-than-perfect fit for misspecified models and stronger misfit for more strongly misspecified models. As expected, block-wise fit cannot identify the misspecification between blocks because block-wise fit indices are based on the (implied and observed) (co)-variances of items associated with the same block.

In line with previous research (Kenny & McCoach, 2003; Moshagen, 2012), global χ^2 is strongly affected by sample size. The same remains true for block-wise χ^2_k, χ^2 evaluation with simulated *df* and Yuan et al. (2015) corrected χ^2. Contrary to the correctly specified models, number of days (and thus number of manifest variables) does not affect the χ^2-tests for misspecified models. This could be due to the fact that the number of misspecified covariances also increases with model size.

CFI and TLI behave practically identical in our simulation study. Globally, they are sensitive to the number of days in the model, to a lesser extend also when they are estimated globally with simulated *df*. Their block-wise counterparts are not affected by the number of days. Regular global CFI and TLI indicate worse fit for all models with 7 days and $N = 200$, which is a likely ES data scenario. Block-wise CFI_k and TLI_k, and to a lesser extend global CFI and TLI with simulated *df*, generally indicate better fit than regular global indices. Contrary to Kenny and McCoach (2003) and Shi and colleagues (2019), global CFI and TLI also worsened with more days for models with omitted residual correlations. In previous studies, the number of misspecified covariances remained stable with more manifest variables in the model, and the proportion of misspecified covariances decreased with model size. In our study, the number of misspecified covariances also increased with model size, explaining our different results. In fact, for the model with omitted residual correlations within days, the proportion of misspecified covariances is larger for 2 days than 7 days, but CFI and TLI indicate a worse fit for 7 days. This demonstrates that these global indices are indeed strongly affected by the number of manifest variables.

Global RMSEA and global RMSEA based on simulated *df* indicate a slightly better fit for models with more days (i.e., more manifest variables), while the number of days does not affect block-wise RMSEA_k. Especially in the case of strongly misspecified models for 7 days, global RMSEA would still let us erroneously conclude that these models are acceptable, while block-wise evaluation can identify them as fitting badly. Block-wise RMSEA_k generally indicates a worse fit than both global indices, which is desirable in misspecified models. The behavior of global RMSEA is largely in line with previous research (Kenny & McCoach, 2003; Savalei, 2012; Shi et al., 2019).

Global Discussion

In this article, we introduced block-wise model fit evaluation for LST-R models with experience sampling data. We performed two simulation studies to compare block-wise fit evaluation to traditional global evaluation. We also included Yuan and colleagues (2015) corrected χ^2 estimates

© 2022 The Author(s) Distributed as a Hogrefe OpenMind article under the license CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1027/2151-2604/a000482 - Monday, April 29, 2024 9:05:11 PM - IP Address:3.146.37.35

and global fit indices derived with simulated degrees of freedom for comparison. In Study 1, we investigated if the different fit indices properly identify correctly specified models. Results show that traditional global fit evaluation too often leads to the rejection of correctly specified models, especially with realistic sample and model sizes in ES studies. Block-wise fit evaluation, global fit derived from simulated *df*, and Yuan et al. (2015) corrected χ^2 correctly identified that these models fit well.

In the second study, we investigated under which conditions block-wise fit indices correctly reject misspecified models. As expected, if models are misspecified purely between days, block-wise fit cannot identify the misfit. Furthermore, traditional global CFI and TLI generally indicate a worse fit for the most realistic ES data scenario (7 days and N = 200) compared to other models and sample sizes. Block-wise and global indices with simulated *df* do not share this bias. Block-wise RMSEA_k can more often identify (strongly) misspecified models than both types of global RMSEA.

Practical Usage

Based on the simulation results, we recommend using block-wise fit indices for LST models with many measurement occasions (e.g., several occasions on each day for one or more weeks) and sample sizes around 200 or smaller. With large sample sizes of around 1,000, there is less benefit in using block-wise fit evaluation. However, sample sizes around or under 200 are much more common in empirical research, so block-wise fit is useful for data from a typical ES study.

When using block-wise fit, researchers need to find logical blocks in their data. This decision should be based on the (LST) model and the study design. For example, blocks can correspond to days. When data was collected over several weeks with fewer occasions per day, blocks may better correspond to weeks.

Compared with existing corrections for the model size effect, such as Yuan et al. (2015), the block-wise evaluation provides valuable additional information about each block. In empirical data, it may happen that some blocks indicate acceptable fit, while others do not. This information about the source of misfit can be used to reflect on the model and data collection. For example, were there any structural differences between the days of assessment, such as weekdays and weekends being on the same days of the study for all participants? The R-function to determine block-wise fit indices is available in ESM 3.

Limitations and Future Research

The main limitation of the block-wise fit approach is evident from Study 2: misspecifications between blocks cannot be detected. We have proposed a block-wise fit for each block, but it is usually not possible to determine a block-wise fit between two blocks. To calculate block-wise fit indices, we extract blocks from $\hat{\Sigma}$ which only contain the (implied) (co)variances between items of the same block. Theoretically, it would be possible to extract the sections containing only covariances of items from two different blocks. However, if we assume any kind of measurement invariance between the blocks, the section of $\hat{\Sigma}$ which contains only the implied covariances between two blocks i and j, $i \neq j$ contains identical (and thus linearly dependent) vectors. The determinant of such a matrix is zero, log(0) is not defined, and a block-wise χ^2_{ij} cannot be determined. As a consequence, the block-wise fit is not informative about misfit between blocks. In future studies, the block-wise approach could be extended to include information between blocks. It should be possible to extract (co)variances of two consecutive or non-consecutive days and estimate a block-wise indices from these blocks. Blocks of different sizes could also be an option, for example, if a researcher is interested in morning- and evening-blocks with different numbers of measurements.

Also, the influence of differing numbers of indicators per block on block-wise fit indices was not assessed, and we cannot give advice on the number of manifest variables per block. Studies on which common advice for interpreting fit indices are based might serve as an orientation. For example, Hu and Bentler (1999) used 15 indicators.

Furthermore, missing data is common in ES studies. In the article, we have not yet discussed how Full Information Maximum Likelihood (FIML), a common missing data strategy, could be applied for block-wise fit indices. To date, the approach we introduced works with multiply imputed datasets. For practical reasons, it will be helpful to extend this approach to FIML. We have also focused on the χ^2-test, CFI, TLI and RMSEA, but the block-wise approach can be extended to other fit indices.

Electronic Supplementary Material

The electronic supplementary materials are available with the online version of the article at https://doi.org/10.1027/2151-2604/a000482

ESM 1. Formula of the Standardized Root Mean Square Residual (SRMR), design, and results of Studies 1 and 2. Figure E1: Design of Studies 1 and 2. Figure E2: Results of Study 1 for the single-trait model. Figure E3: Results of Study 1 for the model with day-specific traits. Figure E4: Results of Study 2 for each fit index, (a)–(e). Tables E1–E4: ANOVA tables showing the effects on χ^2/df ratios, CFI, TLI, and RMSEA.

© 2022 The Author(s) Distributed as a Hogrefe OpenMind article under the license CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0)

https://econtent.hogrefe.com/doi/pdf/10.1027/2151-2604/a000482 - Monday, April 29, 2024 9:05:11 PM - IP Address:3.146.37.35

ESM 2. Results of the block-wise degrees of freedom pilot simulation study.
ESM 3. The R-function to compute block-wise fit indices based on a fitted lavaan object.

References

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(3), 359–388. https://doi.org/10.1080/10705511.2017.1406803

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Statistical Psychology, 3*(2), 77–85.

Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software, 5*(56), Article 2815. https://doi.org/10.21105/joss.02815

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. Guilford Press.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Chalmers, R. P., & Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology, 16*(4), 248–280. https://doi.org/10.20982/tqmp.16.4.p248

Devlieger, I. (2019). *Factor score regression* (Doctoral dissertation). Ghent University.

Eid, M., Courvoisier, D. S., & Lischetzke, T. (2012). Structural equation modeling of ambulatory assessment data. In M. R. Mehl & T. S. Connor (Eds.), *Handbook of research methods for studying daily life* (pp. 384–406). Guilford Press.

Eid, M., Holtmann, J., Santangelo, P., & Ebner-Priemer, U. (2017). On the definition of latent-state-trait models with autoregressive effects. *European Journal of Psychological Assessment, 33*(4), 285–295. https://doi.org/10.1027/1015-5759/a000435

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Geiser, C. (2021, June). *Are psychological constructs traits or states? Preliminary findings from a review of applied latent state-trait studies*. Paper presented at the EAPA Digital Event 2021 Conference. https://www.eapa.science/services/ability-assessment-1-1-1

Hamaker, E. L., Nesselroade, J. R., & Molenaar, P. C. M. (2007). The integrated state-trait model. *Journal of Research in Personality, 41*, 295–315. https://doi.org/10.1016/j.jrp.2006.04.003

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods, 20*(1), 102–116. https://doi.org/10.1037/a0038889

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods, 16*(3), 319–336.

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

Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling, 10*(3), 333–351. https://doi.org/10.1207/S15328007SEM1003_1

Kenny, D. A., & Zautra, A. (1995). The trait-state-error model for multiwave data. *Journal of Consulting and Clinical Psychology, 63*(1), 52–59. https://doi.org/10.1037/0022-006X.63.1.52

Kenny, D. A., & Zautra, A. (2001). Trait-state models for longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *Decade of behavior. New methods for the analysis of change* (pp. 243–263). American Psychological Association. https://doi.org/10.1037/10409-008

Maydeu-Olivares, A., & Shi, D. (2017). Effect sizes of model misfit in structural equation models. *Methodology, 13*(S1), 23–30. https://doi.org/10.1027/1614-2241/a000129

Mayer, A. (2020). *Lsttheory: Latent state-trait theory*. [R package version 0.2-1.002]. https://github.com/amayer2010/lsttheory

Mehl, M. R., Conner, T. S., & Csikszentmihalyi, M. (2011). *Handbook of research methods for studying daily life* (1st ed.). Guilford Press.

Moshagen, M. (2012). The model size effect in sem: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling, 19*(1), 86–98. https://doi.org/10.1080/10705511.2012.634724

Norget, J., Columbus, S., Mayer, A., & Balliet, D. (2021, June). *Latent state-trait models of subjective interdependence*. Paper presented at the EAPA Digital Event 2021 Conference. https://www.eapa.science/services/ability-assessment-1-1-1

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. http://www.jstatsoft.org/v48/i02/

Rosseel, Y., & Loh, W. W. (2021). *A structural after measurement (SAM) approach to SEM*. https://osf.io/pekbm/

RStudio Team. (2019). *Rstudio: Integrated development environment for R*. RStudio. http://www.rstudio.com/

Savalei, V. (2012). The relationship between root mean square error of approximation and model misspecification in confirmatory factor analysis models. *Educational and Psychological Measurement, 72*(6), 910–932. https://doi.org/10.1177/0013164412452564

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), 23–74.

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement, 79*(2), 310–334. https://doi.org/10.1177/0013164418783530

Shi, D., Lee, T., & Terry, R. A. (2018). Revisiting the model size effect in structural equation modeling. *Structural Equation Modeling, 25*(1), 21–40. https://doi.org/10.1080/10705511.2017.1369088

Stadtbäumer, N., Kreissl, S., & Mayer, A. (2021). *Comparing reformulated latent state-trait models with autoregressive effects*. Manuscript submitted for publication.

Steyer, R., Mayer, A., Geiser, C., & Cole, D. A. (2015). A theory of states and traits-revised. *Annual Review of Clinical Psychology, 11*, 71–98. https://doi.org/10.1146/annurev-clinpsy-032813-153719

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality, 13*(5), 389–408. https://doi.org/10.1002/(SICI)1099-0984(199909/10)13:5<389::AID-PER361>3.0CO;2-A

© 2022 The Author(s) Distributed as a Hogrefe OpenMind article under the license CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0)

Swain, A. J. (1975). *Analysis of parametric structures for variance matrices* (Doctoral dissertation). Adelaide.

Thoemmes, F., Rosseel, Y., & Textor, J. (2018). Local fit evaluation of structural equation models using graphical criteria. *Psychological Methods, 23*(1), 27–41. https://doi.org/10.1037/met0000147

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*(1), 1–10. https://doi.org/10.1007/BF02291170

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer. http://www.stats.ox.ac.uk/pub/MASS4/

Yuan, K.-H., Tian, Y., & Yanagihara, H. (2015). Empirical correction to the likelihood ratio statistic for structural equation modeling with many variables. *Psychometrika, 80*(2), 379–405. https://doi.org/10.1007/s11336-013-9386-5

Zhang, Z., Hamaker, E. L., & Nesselroade, J. R. (2008). Comparisons of four methods for estimating a dynamic factor model. *Structural Equation Modeling, 15*(3), 377–402. https://doi.org/10.1080/10705510802154281

History
Received April 30, 2021
Revision received October 9, 2021
Accepted November 2, 2021
Published online February 2, 2022

Funding
Open access publication enabled by Bielefeld University.

ORCID
Julia Norget
https://orcid.org/0000-0002-3388-8873

Julia Norget
Faculty of Psychology and Sport Science
Bielefeld University
Universitätsstraße 25
33615 Bielefeld
Germany
julia.norget@uni-bielefeld.de

https://econtent.hogrefe.com/doi/pdf/10.1027/2151-2604/a000482 - Monday, April 29, 2024 9:05:11 PM - IP Address:3.146.37.35

© 2022 The Author(s) Distributed as a Hogrefe OpenMind article
under the license CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0)