STRUCTURAL EQUATION MODELING
2026, VOL. 33, NO. 3
ISSN: 1070-5511 | E-ISSN: 1532-8007

Model Specification Search in Correlated Factor Models Using Bee Swarm Optimization

Tim Trautwein , Florian Scharf , and Ulrich Schroeders

University of Kassel

ABSTRACT

We introduce a Bee Swarm Optimization (BSO), a metaheuristic algorithm for correlated factor models that optimizes test construction by simultaneously determining the dimensionality and the optimal item set of a measure. In an extensive simulation study, we systematically varied the number of factors and indicators, the structural complexity, the magnitude of factor correlations, cross-loadings, and the extent of noise items. We benchmarked the BSO performance against four EFA–CFA pipelines. Performance was assessed via multiple criteria, including global and local model fit, maximal factor correlation, and item retention. BSO consistently recovered optimal models, achieving higher factor-retention accuracy while removing problematic cross-loading and noise items. In 97.2% of datasets and conditions, BSO outperformed the EFA–CFA pipelines. Based on a second simulation, we provide recommendations for hyperparameters. Overall, BSO offers a robust, data-driven alternative to prevalent approaches for developing measures with a correlated factor structure that are interpretable and parsimonious.

KEYWORDS

BSO; item selection; metaheuristics; model specification search; test development

Metaheuristics are increasingly used in the construction of psychological measures, particularly for developing short scales (e.g., Achaa-Amankwaa et al., 2025; Jing et al., 2022; Raborn et al., 2020; Raborn & Leite, 2018). These methods encompass various approaches, many of which are inspired by natural processes. For example, Ant Colony Optimization (Dorigo & Stützle, 2010) mimics the foraging behavior of ants, whereas Genetic Algorithms (Marcoulides & Schumacker, 2001) draw on evolutionary principles such as mutation, crossover, and selection (e.g., Schroeders et al., 2016). A defining characteristic of metaheuristics is that they find near-optimal solutions by iteratively selecting and evaluating item sets in a non-exhaustive search, making them adaptive and non-deterministic optimization procedures. In psychological research, metaheuristics have primarily been used to create short scales that meet goals by optimizing multiple criteria simultaneously, or to optimize model specification in structural equation modeling, where the goal is to identify the best-fitting or most appropriate model without evaluating all possible combinations.

The extensive metaheuristic literature dealing with model specification search typically relies on a fixed number of items and focuses on determining their assignment to a fixed set of factors (Marcoulides et al., 1998, 2012; Marcoulides & Drezner, 2003; Marcoulides & Falk, 2018; Marcoulides & Leite, 2013; Marcoulides & Schumacker, 2001). However, this approach has several limitations in applied research. First, definite knowledge about the assignment of items to factors may be limited, which means that test developers are often faced with an exploratory rather than a confirmatory factor analytic task. Second, the factor structure can be ambiguous with respect to the number of factors and their substantive interpretation. Third, depending on the stage of test development, the quality of the item pool may be suboptimal, necessitating the exclusion of certain items to identify the optimal structure. To this end, Schroeders et al. (2024) developed a Bee Swarm Optimization (BSO) algorithm, which was designed for bifactor models to identify the optimal factor structure while simultaneously performing item selection, and showcased its effectiveness on two empirical datasets. Although their empirical examples yielded plausible solutions, a systematic investigation of the quality of BSO solutions under varying conditions is still pending.

The present study extends previous research on the BSO algorithm in several respects: First, we extend the BSO algorithm for correlated factor models, which requires changes to the algorithm's architecture, including restructured bee operations and the development of new optimization criteria to address the distinct challenges of correlated factor models (e.g., inflated factor correlations and cross-loadings). Second, we modify the algorithm by introducing new nest-site scout bees for a more sophisticated initial model search and better performance in larger factor models. Third, we conduct a comprehensive simulation study to systematically evaluate the performance of the BSO algorithm (e.g., factor retention and consistency of results) across conditions of varying complexity and to compare it with that of simpler

analytical approaches based on a combination of exploratory and confirmatory factor analysis (EFA–CFA), which rely on standard procedures such as Parallel Analysis and cutoff criteria for item exclusion. Finally, we examine the influence of hyperparameter choices in a second simulation study. In the following, we explain the specific changes to the BSO algorithm in the context of correlated factor models before reporting a simulation study assessing the performance of BSO under known data-generating mechanisms.

1. Extending the Bee Swarm Optimization Algorithm to Correlated Factor Models

The Bee Swarm Optimization (BSO) algorithm was inspired by the foraging patterns of bees, whose division of labor enables more complex and organized behavior than that observed in other eusocial insects such as ants (Schroeders et al., 2024; Seeley, 2011). Originally designed for model specification search in bifactor models, there are good reasons to extend its application to correlated factor models: First, correlated factor models are the most commonly used models in social and behavioral science (DiStefano & Hess, 2005), because they align closely with the typical conceptualization of constructs as relatively distinct but related latent traits. While bifactor models presume a unidimensional core by positing a general trait alongside domain-specific traits, correlated factor models are more appropriate when such a common cause is not theoretically justified. Second, bifactor models are particularly prone to absorbing residual correlations into the general factor and may therefore exhibit apparently good model fit even when the model is theoretically inappropriate. This tendency to produce good fit may help explain their recent "rediscovery" (Reise et al., 2016). In contrast, correlated factor models typically yield a more distinct and interpretable structure, at least if a Thurstonian simple structure can be achieved.

Several changes to the initial BSO algorithm were necessary to account for the properties of the correlated factor model: (1) we introduced so-called nest-site scout bees to generate plausible starting solutions; (2) we adapted which changes the scout and onlooker bees can propose to the measurement model; and (3) we added several criteria to the objective function that BSO optimizes, thereby specifying what constitutes an adequate measurement model. In the following, we explain the functioning of BSO and the changes introduced in more detail.

The BSO algorithm mimics the sophisticated division of labor observed in honeybee colonies during foraging and nest site selection (Seeley, 2011; Seeley & Visscher, 2004). The revised version of BSO builds on three distinct roles: (a) *nest-site scout bees*, which explore potential hive locations before building the colony, (b) *scout bees*, which search for promising food sources in the vicinity of the hive, and (c) *onlooker bees*, which follow the scout bees to exploit these sources (see Figure S1 in the online supplement for an overview of the revised workflow). In the context of model specification search, nest-site scout bees identify promising initial factor solutions to narrow down the large search space. Scout bees conduct global searches through major structural modifications to the factor structure, while onlooker bees perform local searches through minor refinements of the item-factor assignment to optimize the measurement model. This sophisticated search strategy, which balances between diversification and intensification, enables BSO to efficiently navigate the solution space of possible factor model configurations (Blum & Roli, 2003).

1.1. Nest-Site Scout Bees

By introducing nest-site scouts as a warm-start initialization, the revised BSO can navigate the vast search space of possible models more efficiently. The original BSO version relied on a cold-start procedure with purely random starting solutions. However, this approach becomes increasingly problematic as the number of candidate items and the number of potential factors grows. For instance, with 56 items and factor models ranging from 1 to 8 factors, there are approximately forty-seven quindecillion ($= 4.70 \times 10^{48}$) possible model configurations. Random sampling is unlikely to identify sufficiently promising initial solutions within this enormous search space without increasing the number of bees to an infeasible level. The new nest-site scout bees use simple exploratory factor analysis (EFA) to identify promising initial solutions. These bees systematically explore different factor structures using various rotation methods (geominQ, equamax, promax), different numbers of factors, and item sets within the specified range. Each EFA solution is then converted into a simple structure CFA model by assigning all items to their highest-loading factor (with random assignment in case of ties), providing a data-informed warm start for the BSO optimization.

1.2. Modifications of Scout and Onlooker Bees

Using BSO for correlated factor models also requires changes to the modifications that scout and onlooker bees can propose to the measurement model. Once promising initial models have been identified, scout bees make major changes to the factor structure of the best candidate solutions. More specifically, scout bees can (a) add a factor (*Scout 1*), (b) split an existing factor into two (*Scout 2*), (c) remove a factor (*Scout 3*), or (d) collapse two factors (*Scout 4*). Each scout bee randomly applies one of these four operations to the best previous solutions. When multiple operations are available, for instance, when more than one factor can be split, the algorithm decides between them at random. For illustration, the major model changes may affect a 20-item scale as follows; an "×" indicates that an item is removed from the item pool, and a number denotes the assignment of the item to a specific factor:

[1 1 1 × 2 2 2 2 × 3 3 3 4 4 × 4 4 4 4] (*Best previous solution*)

[1 1 1 5 2 2 2 2 5 3 3 3 4 4 5 4 4 4] (*Scout 1: Add factor*)

[1 1 1 × 2 2 2 2 × 3 3 3 4 4 × 4 5 5 5] (*Scout 2: Split factor*)

[1 1 1 × ×××× × 3 3 3 4 4 × 4 4 4 4] (*Scout 3: Remove factor*)

[1 1 1 × 2 2 2 2 × 2 2 2 4 4 × 4 4 4 4] (*Scout 4: Merge factors*)

Onlooker bees introduce minor changes to the measurement model, which are also different between bifactor and correlated factor models. In the original bifactor implementation, items could be added to the model, reassigned to the general factor only, removed from the item set, or moved between specific factors. In contrast, in the correlated factor model, all items (including ambiguous items) are either assigned to a specific factor or completely removed from the item pool. Consequently, onlooker bees may introduce the following item-factor changes to the model: (a) add an item to a factor (if it was not previously assigned; *Onlooker 1*), (b) swap an item between factors (*Onlooker 2*), or (c) remove an item from the model entirely (*Onlooker 3*). The decision about which operation to execute and which item to target is made at random. For illustration, the minor model changes may affect a 20-item scale as follows:

[1 1 1 × 2 2 2 2 × 3 3 3 4 4 × 4 4 4 4] (*Best previous solution*)

[1 1 1 1 2 2 2 2 × 3 3 3 4 4 × 4 4 4 4] (*Onlooker 1: Adding item*)

[4 1 1 × 2 2 2 2 × 3 3 3 4 4 × 4 4 4 1] (*Onlooker 2: Swapping items*)

[1 1 1 × × 2 2 2 × 3 3 3 4 4 × 4 4 4 4] (*Onlooker 3: Removing item*)

1.3. Additional Optimization Criteria

Beyond these changes to the BSO algorithm, we propose additional optimization criteria to address distinct challenges of correlated factor models, particularly in handling cross-loadings. In bifactor models, cross-loadings can be absorbed by the general factor, allowing the model to accommodate more ambiguous structures (Bonifay et al., 2025; Reise et al., 2016), without significantly affecting factor correlations. In contrast, in correlated factor models, omitted cross-loadings typically manifest as inflated factor correlations (Asparouhov & Muthén, 2009; Marsh et al., 2013). Although this shift of misfit may improve global fit indices, excessively high factor correlations (e.g., $r > 0.80$) often result in factors collapsing (Sass & Schmitt, 2010; Scharf & Nestler, 2019a). To counteract this effect, factor correlations can be penalized (Scharf & Nestler, 2019b), which motivated our introduction of a penalty for inflated factor correlations.

An important feature of BSO is its capability to perform item selection. To guide this selection effectively, BSO requires explicit criteria that define which model parameter to favor or penalize. One key threat to the generalizability of measurement models is hidden cross-loadings—items that load on factors beyond those specified in the theoretical model (Marsh et al., 2009; Morin et al., 2013). Such misspecifications can distort factor meanings and compromise construct validity. When systematic misspecifications are concentrated within individual factors, these factors may lose theoretical interpretability even when the model achieves acceptable global fit. Since hidden cross-loadings are usually not detectable from global fit indices alone, a local fit criterion is needed to detect and penalize such misspecification. This concern is particularly prominent for correlated factor models, because their flexibility in accommodating factor correlation patterns makes them particularly vulnerable to fitting non-genuine local associations (Sass & Schmitt, 2010; Scharf & Nestler, 2019a). To address this vulnerability and to balance local and global aspects of model fit, the newly introduced local fit criterion penalizes the highest 10% of absolute values in the standardized residual covariance matrix, thereby encouraging item selection toward measurement models with a clean simple structure.

2. Advantages of BSO in Model Specification Search

The BSO method offers several advantages in refining psychological measures. First, BSO integrates item selection directly into the optimization process rather than treating model refinement and item selection as separate steps. Because of its non-greedy nature, BSO is less prone to converging on local optima resulting from the premature exclusion of items (Schroeders et al., 2024). In contrast, manual EFA–CFA pipelines risk measurement artifacts because item selection is based on loadings estimated before any deletions are made; once items are removed, loadings and factor correlations can shift, potentially leading to suboptimal solutions (e.g., Olaru et al., 2015; Schroeders et al., 2016).

Second, unlike other metaheuristics such as ACO, which require items to be assigned to specific factors prior to estimation, BSO does not impose predefined assumptions about item-factor assignment, which allows for a purely data-driven optimization process. Such flexibility is particularly beneficial in the early stages of test development when theoretical knowledge about the measure is still limited. BSO's capacity to identify an optimal structure without strong theoretical assumptions reduces the likelihood of prematurely restricting the solution space and arriving at suboptimal solutions.

Third, when applied to a correlated factor model, BSO facilitates the identification of clear and unambiguous simple structures that better generalize across samples. Unlike EFA or Exploratory Structural Equation Modeling (ESEM) approaches (Marsh et al., 2010; Morin et al., 2013), which allow cross-loadings that may reflect sample-specific noise and thereby increase the risk of overfitting, the distinct and purified factors resulting from a simple structure are easier to interpret and more likely to withstand confirmatory testing in new samples (see also Sass & Sanchez, 2023). Simple-structure models are particularly suitable for measures where factors represent distinct constructs that are adequately captured by their indicators. However, for measures with many ambiguous items, where cross-loadings reflect the genuine complexity of a construct, alternative approaches such as regularized (E)SEM may be preferable to avoid eliminating theoretically important items or item-factor relationships (Jacobucci et al., 2016).

Fourth, metaheuristics in general allow for the simultaneous optimization of multiple criteria (Raborn et al., 2020). In scale shortening, these criteria include (global) model fit, reliability, and measurement invariance and other quantifiable indicators (e.g., Jankowsky et al., 2020; Zimny et al., 2024). In contrast, traditional methods of item selection

often optimize a single criterion (e.g., reliability), potentially leading to suboptimal solutions (see also Kruyen et al., 2013; Steger et al., 2023). By optimizing multiple objectives concurrently, BSO avoids one-sided optimization that may compromise other important aspects (Dorigo & Stützle, 2010; Schroeders et al., 2016).

Finally, in typical EFA-CFA pipelines, item selection mainly relies on indicators of local model misfit (e.g., modification indices, Whittaker, 2012), whereas the identification of a suitable factor structure relies on global model fit statistics. Because adjustments to one aspect often have unpredictable consequences for the other, the balancing process usually proceeds iteratively until an acceptable compromise is reached. In contrast, metaheuristics such as BSO enable the simultaneous optimization of local and global fit. Local model fit can be improved, for example, by penalizing large item residuals, while global model fit can be addressed by optimizing global fit indices and penalizing excessively high factor correlations. These penalty terms promote parsimony by limiting the number of free parameters, ensuring statistically sound and easy to interpret models (e.g., Bader & Moshagen, 2025; Preacher, 2006).

3. The Present Study

Despite these presumed advantages and the promising preliminary results (Schroeders et al., 2024), a systematic evaluation of BSO in simulated data with known data-generating mechanisms, as well as a comparison with more conventional EFA–CFA-based approaches is still pending. To evaluate the extended BSO algorithm, we therefore conducted Monte Carlo simulations systematically varying the (a) number of factors, (b) number of indicators, (c) factor correlations, (d) presence and strength of cross-loadings, and (e) the presence of noise items that load on a substantive factor and an additional noise factor. Specifically, we examined the extent to which BSO identified the correct number of factors, allocated items to the intended factors, and excluded noise items. In addition, we investigated the quality and stability of the solution as a function of the BSO hyperparameters. For this purpose, we varied (a) the number of bees (i.e., the number of solutions tested simultaneously), (b) the number of nest-site scout bees used for initial exploration, (c) the proportion of scout bees (which explore different factor structures) to onlooker bees (which explore different item-factor assignments), and (d) the number of top solutions explored in the next iteration. We assessed the influence of these hyperparameters on the effectiveness and consistency of the metaheuristic results.

4. Method

4.1. Data Generating Models and Hyperparameter Setting

We conducted two simulation studies to evaluate BSO's performance under varying levels of structural complexity and hyperparameter configurations. Table 1 provides an overview of all conditions, data-generation parameters, hyperparameters, and distributions used in both simulations. Simulation 1 was designed to assess whether BSO can reliably recover the correct factor structure while simultaneously selecting items under increasingly challenging conditions. To this end, we generated data across 60 distinct conditions by systematically manipulating the number of factors, the number of indicators per factor, the size of factor correlations, and the complexity of the factor loading pattern matrix across five scenarios (see Table 1). These five scenarios involved: (1) a near-simple structure condition (SS) in which every cross-loading was non-zero but small (i.e., between 0.02 and 0.20; see also Sass & Schmitt, 2010; Schmitt & Sass, 2011); (2) a complex structure condition (CS) in which 10% of indicators exhibited stronger cross-loadings (between 0.20 and 0.40); (3) a simple noise condition (SN), where an additional noise factor with three indicators was introduced with cross-loadings on all items (between 0.20 and 0.35); (4) a complex noise condition (CN) that combined the previous two manipulations, making it a more challenging scenario because neither condition complies with a Thurstonian simple structure; and (5) a massive noise condition (MN) with approximately 35% noise items across two noise factors (i.e., an additional major noise factor with cross-loadings between 0.30 and 0.45 on items of 1–2 substantive factors), which was the most demanding condition. To investigate to what extent BSO tends to collapse substantially correlated factors, we manipulated the degree of correlations among latent factors. In the low-correlation condition, the first three factors correlated at $\rho = 0.30$, with all additional factors set to weak correlations ($\rho = 0.10$). In the high-correlation condition, the first two factors correlated at $\rho = 0.60$, whereas all other correlations were set to $\rho = 0.10$. This design maintained positive definite factor correlation matrices across all conditions.

Simulation 1 was conducted with a fixed set of hyperparameters informed by previous findings (Schroeders et al., 2024). These settings were chosen to ensure that the BSO algorithm operates reliably across different levels of structural complexity—even if not in the most computationally efficient way, as fewer tested models (i.e., fewer bees) would likely suffice in simpler scenarios. The following parameters were used: (a) 500 bees, (b) 200 nest-site scout bees for initial model search, (c) 25% of them designated as scout bees, and (d) the top 10% of solutions retained for further exploration (see Table 2). In contrast, we systematically varied the hyperparameter settings in Simulation 2 across 54 unique combinations: the number of bees (200, 500, 1000), nest-site scout bees (200, 400), proportion of scout bees (10%, 25%, 40%), and percentage of top solutions retained (10%, 20%, 30%). We therefore tested these hyperparameter settings under the three conditions considered among the most challenging in Simulation 1.

4.2. Optimization Criteria, Dependent Measures, and Implementation

The core of each metaheuristic is the objective function, which specifies the criteria to be optimized and their relative

Table 1. Data generation and simulation settings for simulations 1 and 2.

	Simulation 1	Simulation 2
Loading Patterns	**Values**	
All Patterns	Substantive factors: primary loadings 0.50–0.75	
Simple Structure (SS)	All cross-loadings ≤ 0.20, realistic minimums ($\lambda \geq 0.015$)	
Complex Structure (CS)	SS + 10% items with high cross-loadings on 1 substantive factor $\lambda \in [0.20, 0.40]$	
Simple Noise (SN)	SS + 1 minor noise factor (3 items)	
Complex Noise (CN)	CS + 1 minor noise factor (3 items)	
Massive Noise (MN)	SS + 2 noise factors (1 minor + 1 major; ~35% noise items)	
Data-Generating Parameters	**Distributions/Values**	
Primary factor loadings (substantive)	$\mathcal{U}(0.50, 0.75)$[a]	
Primary factor loadings (minor noise)	$\mathcal{U}(0.35, 0.45)$	
Primary factor loadings (major noise)	$\mathcal{U}(0.40, 0.55)$	
Cross-loadings (regular)	$\mathcal{N}_{trunc}(\hat{\mu}, \hat{\sigma}; 0.015, 0.20)$[b]	
Cross-loadings (higher, CS/CN)	$\mathcal{U}(0.20, 0.40)$ for 10% of items	
Cross-loadings (minor noise, SN/CN/MN)	$\mathcal{U}(0.20, 0.35)$	
Cross-loadings (major noise, MN)	$\mathcal{U}(0.30, 0.45)$ for 1–2 substantive factors	
Factor correlation structure (low)	$\rho_{12} = \rho_{13} = \rho_{23} = 0.30$; others $= 0.10$[c]	
Factor correlation structure (high)	$\rho_{12} = 0.60$; others $= 0.10$[c]	
Maximum communality constraint	$h^2 \leq 0.99$	
Simulation Conditions	**Values**	**Values**
Number of latent factors (k)	3, 5	(3, 5)
Items per factor (p)	4, 5, 7	(4, 5)
Factor correlation magnitude (ρ)	low, high	high
Loading pattern complexity (k, p)	SS, CS, SN, CN, MN	CN (3,4), MN(3,4), MN(5,5)
Total conditions	60	3
Datasets per condition	30	30
Sample size per dataset (N)	500	500
Seeds per dataset	20	20
BSO Hyperparameters	**Values**	**Values**
Number of bees	500	200, 500, 1000
Proportion of scout bees (%)	25	10, 25, 40
Top solutions retained (%)	10	10, 20, 30
Nest-site scout bees	200	200, 400
Hyperparameter combinations	1	54

Note. SS: Simple Structure; CS: Complex Structure; SN: Simple Noise; CN: Complex Noise; MN: Massive Noise. $\mathcal{U}$ (a,b) = uniform distribution; $\mathcal{N}_{trunc}$ (μ,σ;a,b) = truncated normal distribution. The values selected for primary and cross-loadings align with commonly reported factor-analytic models in the literature (e.g., Sass & Schmitt, 2010). The values in parentheses in Simulation 2 give the number of factors and indicators.

[a]Upper bound reduced to 0.70 when $k \geq 5$ to ensure sufficient budget for cross-loadings.

[b]Parameters $\hat{\mu}$ and $\hat{\sigma}$ are adaptively calculated based on available communality budget using the formula $\hat{\mu} = 0.25 \times \text{mean(available\_budget)}$, $\hat{\sigma} = 0.15 \times \text{mean(available\_budget)}$.

[c]For $k > 3$, additional factors maintain weak correlations ($\rho = 0.10$) to prevent eigenvalue collapse while ensuring positive definite correlation matrices.

Table 2. Hyperparameters of the BSO algorithm.

Argument	Explanation	Value
`n_bees`*	Number of bees in the hive	500
`percent_scouts`*	Proportion of scout bees	0.25
`top_best`*	Number of top solutions to be examined	50
`nest_site_bees`*	Number of nest-site scout bees used for initial model search	200
`max_iter`	Maximum number of iterations without improvement	10
`min_fac`	Minimum number of (correlated) factors	1
`max_fac`	Maximum number of (correlated) factors	8
`depletion`	Number of iterations after which the food source is depleted	5

Note. In Simulation 1 the hyperparameters were fixed to the values given in the last column. In Simulation 2, the arguments marked with an asterisk (*) were systematically varied.

weights. In the context of model specification search, the objective function defines what constitutes an adequate measurement model. We included five criteria in the BSO algorithm, which remained consistent across all conditions and simulations: (1) global model fit, (2) local model fit, (3) maximum factor correlation, (4) the number of retained items, and (5) a minimum threshold for item retention. All criteria were rescaled to a 0–1 metric with an inverse logit transformation centered on the respective cutoffs (see Appendix for details). On this scale, values closer to 1 indicate better fulfillment, and values closer to 0 indicate increasingly severe violations. All criterion values are combined into an overall nectar value ν, named after the Greek word nectar ($\nu\acute{\varepsilon}\kappa\tau\alpha\rho$). A more detailed functional specification of these optimization criteria, including weighting details, is provided in the Appendix.

The first three criteria (i.e., global model fit, local model fit, and the maximum factor correlation) primarily guided the dimensionality aspect within the model selection process. Global model fit was evaluated using the Comparative Fit Index (CFI) and Root Mean Square Error of Approximation (RMSEA), with cutoffs of CFI $\geq$ 0.95 and

RMSEA $\leq$ 0.06 indicating good model fit (Hu & Bentler, 1999). Local model fit was considered sufficient if the highest 10% of values in the standardized residual variance-covariance matrix were on average below 2.58 (corresponding to $\alpha = 0.01$). The maximum factor correlation was constrained to $r \leq 0.80$.

The final two criteria (i.e., number of retained items and minimum item retention) guided the item selection process. The fourth criterion required that at least 75% of the original items were retained, representing a desirable (though not strictly mandatory) target for preserving scale length and favoring scales with more items. The fifth criterion imposed a hard minimum acceptable scale length, ensuring that at least 67% of the initial items remained. Models falling below this threshold were treated as insufficient and discarded. In the massive noise condition, this minimum was relaxed to 50% to account for the higher proportion of construct-irrelevant items and to allow for stronger item selection. These requirements ensured that scales retained sufficient length to support reliable measurement and adequate substantive coverage, rather than achieving good fit simply by discarding large numbers of items and narrowing dimensionality to fewer factors.

All simulations were conducted in R (v. 4.5.0; R Core Team, 2025) using the *mvtnorm* package (Genz & Bretz, 2009) to generate multivariate normal data. Each condition was replicated across 30 datasets. The sample size was fixed at $N = 500$ across all conditions. This sample size provides sufficient estimation accuracy in line with recommendations for confirmatory factor analysis (Wolf et al., 2013), while introducing realistic sampling variation. To avoid local optima within the optimization process, BSO was executed using 20 different random seeds per dataset. The extended BSO algorithm for correlated factor models is openly available on GitHub: https://github.com/timtrautwein/BSO-Correlated-Factor-Models. All CFA models were estimated in R using *lavaan* (v. 0.6.19; Rosseel, 2012); exploratory factor analyses and the parallel analyses were conducted with the *psych* package (v. 2.5.3; Revelle, 2024). All syntax and data files necessary to reproduce the analyses of the simulations are available online: https://osf.io/3vuth.

4.3. Conventional EFA–CFA Procedures as Benchmark

To provide a point of comparison for BSO, we also processed every simulated data set with four model modification pipelines based on a combination of EFA and CFA, mimicking popular manual modification strategies by researchers. To ensure comparability with BSO while still reflecting applied practice, we implemented pragmatic "one-shot" EFA–CFA pipelines as benchmarks, that is, non-fitting items were removed based on the initial EFA solution, and the remaining item pool and structure was re-estimated once using a CFA. Although sequential re-estimation after each item deletion can partially mitigate shifts in loadings, it also introduces path dependence and sequence effects (where final solutions depend on deletion order) inherent to greedy approaches (e.g., Castle et al., 2011).

All pipelines included the following four steps: First, parallel analysis determined the number of factors, using either the 95th-percentile or the mean-eigenvalue rule as threshold. Second, a maximum-likelihood EFA with geomin rotation was fitted, relying on either the default ε used in Mplus or a fixed $\varepsilon = 0.50$ to avoid inflated factor correlations due to cross-loadings (Marsh et al., 2009, 2014). Third, items were deleted from the item pool when their primary loading was too low (either $\lambda \leq 0.30$ or $\lambda \leq 0.50$) or when a cross-loading was too high (either $\lambda \geq 0.30$ or $\lambda \geq 0.20$). A minimum of three items was required for factor retention; factors below this threshold were dissolved (Costello & Osborne, 2005; Fabrigar & Wegener, 2012). Fourth, the resulting structure was re-estimated in a CFA assuming simple structure (with factor assignment based on primary loadings) so that the same criteria could be extracted that were used in the BSO algorithm. The standard benchmark (*EFA Standard*) combined the 95th-percentile rule, default ε (2 Factors = 0.0001/ 3 Factors = 0.001/$\geq$ 4 Factors = 0.01), and moderate item selection (minimum main loading/maximum cross-loading: 0.30/0.30). The other three pipelines varied one of these decisions: the alternative rotation benchmark, *EFA Alternative-Rotation*, changed ε to 0.50, thereby tolerating stronger cross-loadings; the mean-PA benchmark (*EFA Mean-PA*) replaced the percentile rule with the mean-eigenvalue criterion; and the strict benchmark (*EFA Strict*) applied stricter item selection thresholds (0.50/0.20). These pipelines enabled a comparison of BSO with EFA-CFA procedures that followed standard practices in the literature (see also Goretzko et al., 2021).

4.4. Evaluation Criteria for the Simulation Studies

In Simulation 1, we evaluated the effectiveness of BSO across several key aspects. First, we examined its *criteria optimization performance*, defined as the proportion of datasets in which the best model met the criteria concerning the model specification part (i.e., global and local model fit, and maximum factor correlation), excluding criteria concerning item selection. In this context, the best model was defined as the one with the highest overall nectar value ν within a dataset across different seeds. A criterion was considered satisfied if the model parameter was equal to or exceeded the predefined threshold (i.e., CFI $\geq$ 0.95, RMSEA $\leq$ 0.06, max factor correlation $\leq$ 0.80), which corresponds to an inverse logit-transformed ν-value equal or above 0.5. Second, we examined *factor retention*, comparing the number of factors in the selected best models to the ideal target solution, as well as factor collapse rates (assessed by whether items from the highly correlated first two factors were assigned to the same factor). Third, we investigated *model purification*, assessing how frequently noise items and items with high cross-loadings were excluded. Fourth, we assessed BSO's *consistency across seeds*, examining how often the best model was obtained across different datasets and seeds. Finally, we evaluated BSO's *competitive performance* relative to the EFA pipelines as a benchmark, including the

performance on criteria optimized across conditions, factor retention and model purification.

Simulation 2 investigated to what extent hyperparameters affect the performance of BSO (see Table 1). To manage computational requirements of the simulation, we focused on three of the most challenging conditions based on the results of Simulation 1: (1) the complex noise condition, (2) the massive noise condition with 3 factors (4 indicators each, high factor correlations), and (3) the massive noise condition with 5 factors (5 indicators per factor, high factor correlations). To evaluate the results of the simulation, we calculated the proportion of best models within each dataset that met all criteria related to model specification (i.e., global and local model fit, and maximum factor correlation), as well as the proportion that met the individual criteria. For both simulations, key determinants of performance metrics were identified via generalized η^2 extracted from ANOVAs and visualized in the corresponding figures.

5. Results

In the following, we will first report results on the performance of the BSO algorithm, followed by a comparison with the EFA-CFA pipelines (Simulation 1). Afterwards, we will report the results of Simulation 2 in which we investigated the influence of BSO hyperparameter choices.

5.1. Evaluation of BSO Performance

Figure 1 summarizes the key determinants of BSO's performance with respect to (a) optimization performance, (b) factor retention, (c) model purification, and (d) consistency across seeds.

5.1.1. Optimization Performance

Optimization performance directly reflects the proportion of datasets in which the best model meets all criteria. Across best models, criteria for global fit (CFI, RMSEA) and the maximum factor correlation were always satisfied, with the local fit criterion being the only one that was occasionally not fulfilled. As shown in Figure 1, Panel A, optimization performance depended strongly on the combination of loading patterns and the number of factors. The performance decreased with increasing structural complexity in the loading patterns (i.e., more cross-loadings on substantial items combined with noise items)—from 96% (near-simple) to 59% (complex noise). Counterintuitively, the massive noise condition attained the highest optimization performance (98%), because BSO effectively removed ambiguous items, while leaving enough items with a clear allocation to satisfy local fit and the other optimization criteria. By contrast, when many cross-loadings remained even after item selection, sampling variance could make local fit unattainable without violating other criteria (e.g., via inflated factor correlations). The loading-pattern × dimensionality interaction was most pronounced in the complex noise condition, where optimization performance dropped strongly from three to five factors (76%–42%), whereas simpler patterns showed less sensitivity toward dimensionality. Detailed results for all metrics and conditions are provided in Table S1 of the online supplement.

5.1.2. Factor Retention

As shown in Figure 1, Panel B, the *factor retention* rate refers to the proportion of datasets in which the best BSO model recovers the target factor count. Importantly, model fit values can only meaningfully be interpreted when the correct factor structure is identified and the factor loadings are sufficiently high (Heene et al., 2011). Factor retention was primarily driven by the factor correlation pattern and interacted substantially with dimensionality. With low factor correlations ($\rho_{12} = \rho_{13} = \rho_{23} = 0.30$; others $= 0.10$), BSO almost always recovered the simulated number of factors. With higher correlations ($\rho_{12} = 0.60$; others $= 0.10$), however, factor retention deteriorated on average from about 70% with 3 factors to 10% with 5 factors because the highly correlated factors often collapsed in the best models. Factor collapsing and the factor retention were almost perfectly inversely related ($r = -0.97$, $p < .001$). This pattern reflects a general identification challenge in factor analysis rather than a BSO-specific limitation (see also the comparison to the EFA-CFA benchmark model below): unmodeled cross-loadings inflate factor correlations and promote conflation of highly correlated factors (Asparouhov & Muthén, 2009; Scharf & Nestler, 2019a).

5.1.3. Model Purification

Model purification quantifies the degree to which items with high cross-loadings and noise items were removed, aiming for a model solution that comes close to a Thurstonian simple structure. Figure 1, Panel C, shows that BSO effectively removed problematic items including noise items and those with high cross-loadings. Model purification was primarily driven by the factor loading pattern and the number of factors, and it was slightly better in models with more factors. In conditions with high cross-loadings (i.e., complex structure, complex noise), the best BSO models achieved a purification rate of 79%. Items with the largest cross-loadings were typically eliminated, whereas items with moderate cross-loadings ($\approx$ 0.20–0.25) were removed in about 50% of cases. For noise items, purification was generally high and higher than for items with cross-loading. In the simple noise condition, minor noise factors were almost always fully removed by BSO. Even in complex and massive noise conditions, the average purification rate remained high at 88%, indicating that ambiguous noise items were removed reliably. More detailed results on model purification are provided in Figure S2 in the online supplement.

5.1.4. Consistency Across Seeds

To assess *consistency* across seeds of the BSO algorithm, we determined the proportion of random seeds per dataset and condition for which the best model was recovered (i.e., the model with the highest overall ν for a given dataset). In

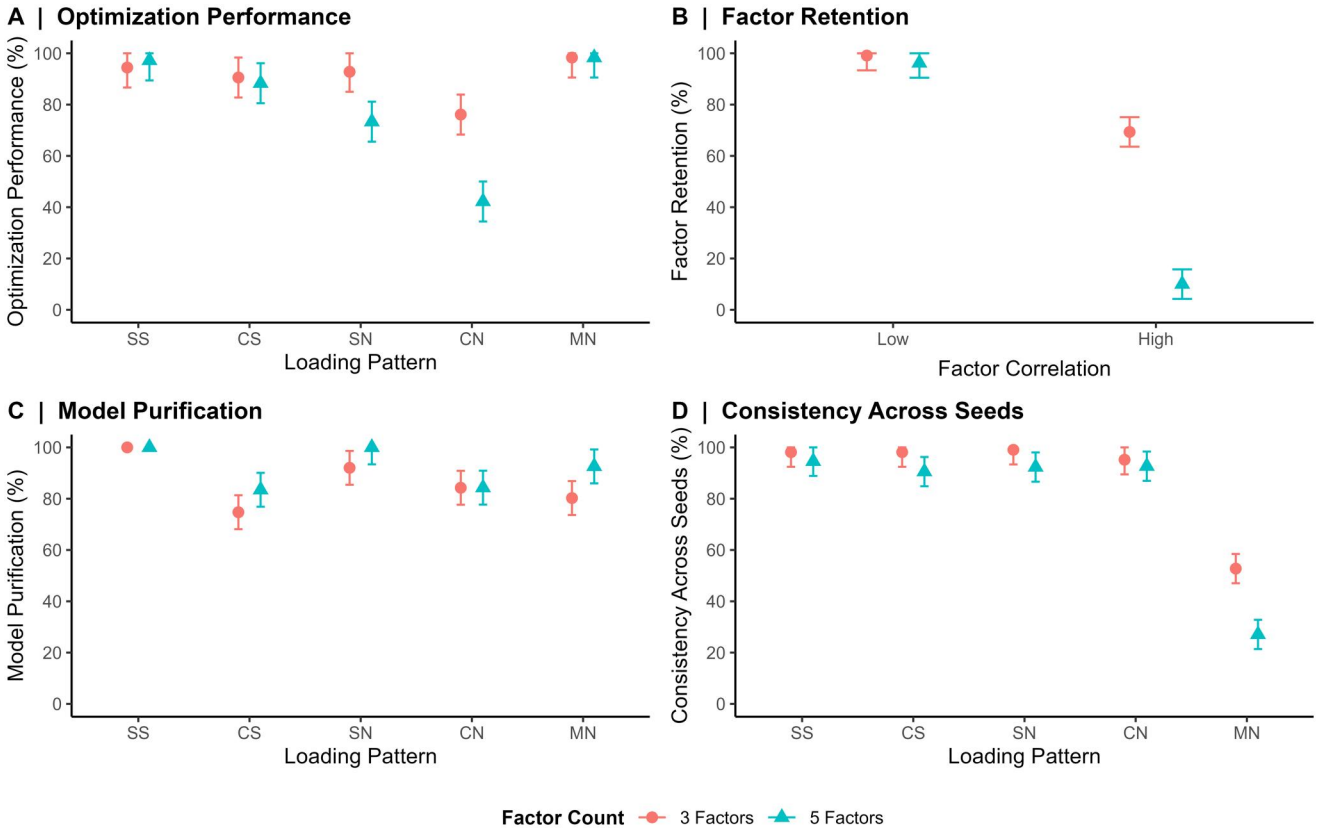

Figure 1. Model performance across the most influential Simulation factors.
Note. Panel A: Estimated marginal means (EMMs) of optimization performance (%) by loading pattern × factor count (F). *Panel B*: EMMs of factor retention (%) by factor correlation (ρ) × factor count. *Panel C*: EMMs of model purification (%) by loading pattern × factor count. *Panel D*: EMMs of consistency across seeds (%) by loading pattern × factor count. Points show estimated marginal means (EMMs) with 95% CIs for each performance metric at the most influential main effects/interactions, as identified by an ANOVA across conditions. EMMs are averaged over the remaining condition factors not displayed in each panel. *Optimization performance* = *Local-fit fulfillment* = % of best models meeting the all criteria, whereas local fit was the only criterion not always fulfilled ($\nu \geq .50$). *Factor retention* = % of best models matching the true number of substantial factors. *Model purification* = % of problematic items (noise or high cross-loading items) removed from the model. *Consistency* = agreement of best models across random seeds. Loading patterns: SS: Simple Structure; CS: Complex Structure; SN: Simple Noise; CN: Complex Noise; MN: Massive Noise; F: number of factors; ρ: factor correlation with two levels: *Low* (first three factors $\rho = 0.30$; others $\rho = 0.10$) and *High* (first two factors $\rho = 0.60$; others $\rho = 0.10$).

general consistency was high at around 95% across most simulation settings and depended on both the factor loading pattern and the number of factors (see Figure 1, Panel D). Only in the massive noise condition, consistency declined substantially due to the many ambiguous items, which created multiple plausible item assignments. After removing noise items, the changed factor loading matrix still allowed several comparably well-fitting solutions. For detailed results across conditions, see Table S1 and Figure S3 in the online supplement.

5.2. Performance of BSO Relative to Conventional EFA-CFA Benchmarks

Overall, BSO successfully optimized the criteria given in its objective function. Compared with simpler EFA-CFA analytical pipelines, BSO results were consistently better with respect to optimization performance, factor retention, and model purification. As shown in Figure 2, Panels A–B, BSO achieved acceptable and less variable global model fit values, whereas conventional procedures exhibited on average poorer and much more variable results—likely due to non-trivial changes in the factor loadings and factor correlations after item removal. Panel C of Figure 2 shows that BSO also achieved better and less variable local fit values. Panel D further demonstrates that BSO reliably kept factor correlations below the prespecified threshold that many conventional approaches frequently exceeded (with *EFA Strict* as a notable exception due to more aggressive item selection). Taken together, in the combined overall ν score metric that also includes the item retention criteria, BSO outperformed current EFA–CFA practice in 97% of cases across all datasets and conditions.

5.2.1. Factor Retention Compared to EFA-CFA Pipelines

Table 3 summarizes *factor retention* of BSO in comparison with the four EFA-CFA pipelines. BSO achieved the highest factor retention accuracy of 69% across all datasets, compared to 39%–55% for the EFA-CFA pipelines. BSO showed lower rates of both over- and underextraction; overextraction was limited to noise-item conditions where some noise items could be combined to an additional factor after other items were removed. With low factor correlations ($\rho_{12} = \rho_{13} = \rho_{23} = 0.30$; others = 0.10), BSO nearly always recovered the correct number of factors, substantially outperforming EFA methods. However, in conditions with highly correlated factors, all methods struggled with underextraction—collapsing highly correlated factors—though BSO achieved higher accuracy than

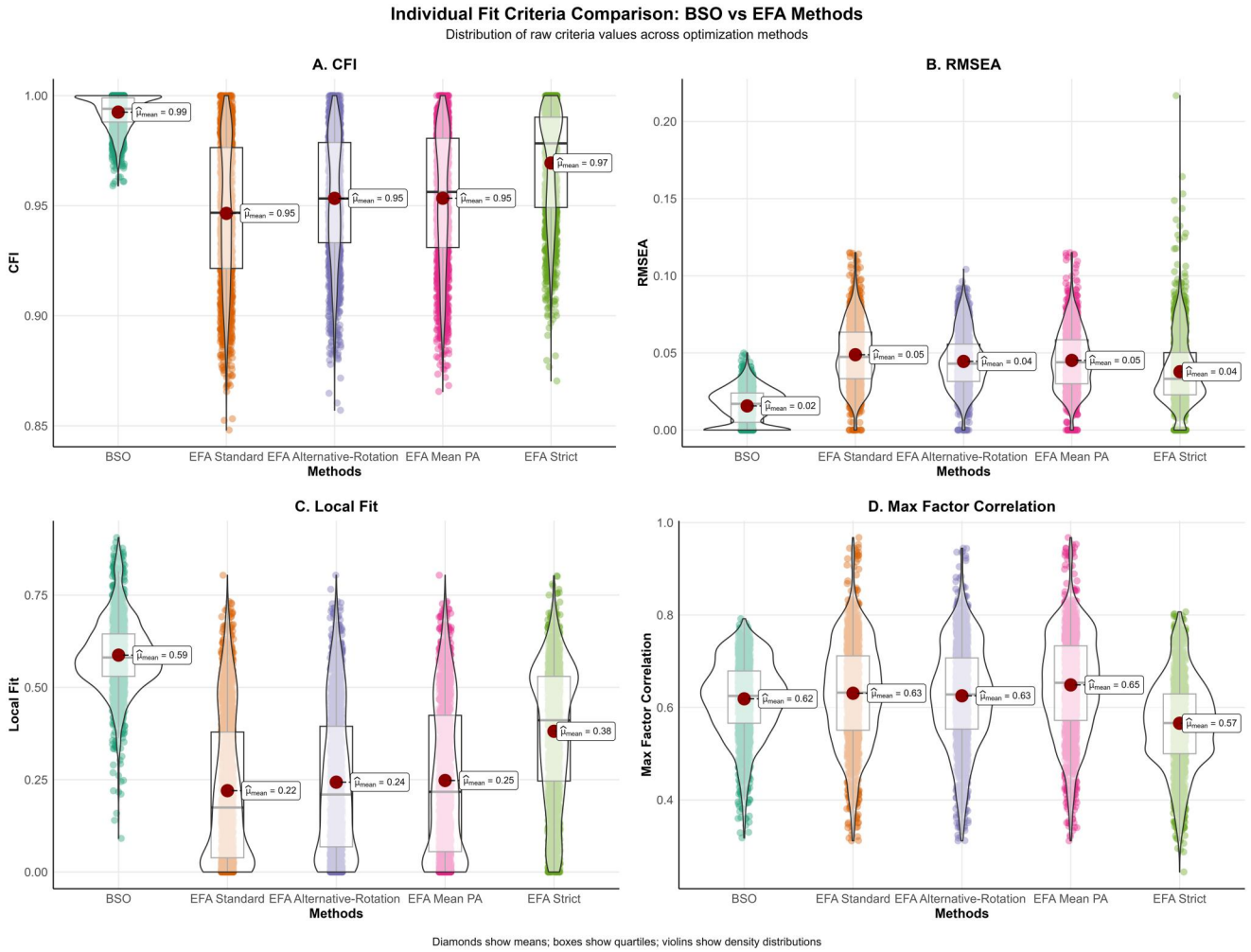

Figure 2. Distribution of individual criteria fulfillment of BSO against EFA benchmarks (Simulation 1).
Note. Each violin shows the full distribution of raw values from 1,800 simulated data sets; boxes mark the interquartile range, horizontal lines the median, and red diamonds the mean. *Higher CFI* and *local fit*, along with *lower RMSEA* and *maximum factor correlation* $r < 0.8$, indicate better solutions. All EFA–CFA benchmarks followed the same four-step pipeline: (1) factor retention by parallel analysis (95th-percentile rule) or the mean-eigenvalue rule; (2) maximum-likelihood EFA with geomin rotation using either Mplus' default ε or $\varepsilon = 0.50$; (3) item selection based on primary loadings ($\leq$0.30 or $\leq$0.50) and cross-loadings ($\geq$0.30 or $\geq$0.20) while retaining $\geq$3 items per factor; and (4) re-estimation of the retained structure in a simple structure CFA to extract the same criteria as BSO. The *EFA Standard* condition used the 95th-percentile rule, default ε, and moderate item selection (0.30/0.30); *EFA Alternative Rotation* changed only ε to 0.50; *EFA Mean-PA* used the mean-eigenvalue rule; and *EFA Strict* tightened item selection to 0.50/0.20. CFI: comparative fit index; RMSEA: root-mean-square error of approximation; Max Factor Correlation = highest latent-factor correlation in the model; Local Fit = criterion penalizing the 10% largest standardized residuals in the variance–covariance matrix, higher values indicate better local fit.

EFA-CFA pipelines and showed narrower dispersion in factor count distributions. In condition-by-condition comparisons with simple EFA-CFA pipelines, BSO achieved the best factor retention in 48% of conditions, tied in 32%, and was outperformed in 20% (where the best EFA–CFA pipeline varied by setting). Even in datasets where an EFA–CFA pipeline retained the targeted number of factors more accurately, BSO still obtained models with a higher overall ν, indicating that some underextracted solutions were preferred by the optimization criteria. This pattern of results illustrates that multiple model specifications can yield essentially equivalent performance on the optimization criteria. For a visualization of factor retention across methods, see Figure S4 in the online supplement.

5.2.2. Model Purification Compared to EFA-CFA Pipelines

Regarding model purification, BSO and *EFA Strict* both successfully removed most problematic items. *EFA Strict* showed slightly better noise-item purification (94% vs. 88% for BSO), whereas both methods yielded comparable cross-loading purification rates (80%). In contrast, the remaining EFA-CFA pipelines performed markedly worse (e.g., 32% cross-loading purification rate) and exhibited considerably broader dispersion. However, similar purification rates do not imply similar models: item removal changes factor loadings and factor correlations, so items that appear problematic in the full set may fit well within reduced item sets. For instance, in one massive noise condition dataset (3 factors, 4 indicators, high factor correlation), BSO and *EFA Strict* models both showed a model purification rate of 87.5%, yet BSO achieved notably better global and local fit (CFI = 0.99 vs. 0.93; RMSEA = 0.03 vs. 0.07; local fit = 0.51 vs. 0.32). This helps explain why BSO and *EFA Strict* differ on global and local fit values (see Figure 2) despite similar purification rates, as BSO's iterative CFA re-estimation addresses these effects. Full purification distributions are provided in the online supplement (see Figure S2).

Table 3. Comparison of numbers of extracted factors across methods.

		Factors extracted							
Condition	Method	1	2	3	4	5	6	Underextracted	Overextracted
Low ρ	BSO	—	—	**99.1**	0.7	0.2	—	—	CN, MN
3 Factors	EFA Standard	5.6	16	**75.3**	3.1	—	—	CN, CS, MN, SN	MN
	EFA Alter.-Rotation	—	3.8	**85.3**	10.9	—	—	CN, CS, MN, SN	MN
	EFA Mean PA	5.1	15.3	**76.0**	3.6	—	—	CN, MN, SN	MN
	EFA Strict	16.9	15.6	**67.5**	—	—	—	ALL	—
High ρ	BSO	—	23.1	**69.3**	7.3	0.2	—	ALL	CN, MN, SN
3 Factors	EFA Standard	0.4	69.8	**28.0**	1.8	—	—	ALL	MN
	EFA Alter.-Rotation	—	64.0	**31.8**	4.2	—	—	ALL	MN
	EFA Mean PA	0.4	54.2	**42.4**	2.9	—	—	ALL	MN
	EFA Strict	2.4	72.7	**24.9**	—	—	—	ALL	—
Low ρ	BSO	—	—	—	3.8	**96.2**	—	CS, MN	—
5 Factors	EFA Standard	0.9	2.0	6.9	20.9	**68.9**	0.4	ALL	MN
	EFA Alter.-Rotation	—	0.2	6.9	20	**68.4**	4.4	ALL	MN
	EFA Mean PA	0.2	0.2	1.8	9.3	**88.0**	0.4	ALL	MN
	EFA Strict	1.1	6.2	10.9	20.0	**61.8**	—	ALL	—
High ρ	BSO	—	—	—	89.6	**10.0**	0.4	ALL	CN, MN
5 Factors	EFA Standard	—	0.7	3.1	90.7	**5.6**	—	ALL	—
	EFA Alter.-Rotation	—	—	0.9	91.1	**8.0**	—	ALL	—
	EFA Mean PA	—	0.2	2.4	82.7	**14.7**	—	ALL	—
	EFA Strict	0.2	2.7	14.7	79.8	**2.7**	—	ALL	—

Note. Numbers represent the percentage of datasets where each method extracted the corresponding number of factors. Bold percentages indicate correct extraction if the number of simulated items in the complete itemset is considered the true number. Underextracted and overextracted. columns list loading patterns where under or overextraction occurred. BSO: Bee Swarm Optimization; EFA: Exploratory Factor Analysis. The *EFA Standard* condition used the 95th-percentile rule, default ε, and moderate item selection (0.30/0.30); *EFA Alternative Rotation* changed only ε to 0.50; *EFA Mean-PA* used the mean-eigenvalue rule; and *EFA Strict* tightened item selection to 0.50/0.20. Em dash (—) indicates no cases. ρ = factor correlation with two levels: *Low* (first three factors $\rho = 0.30$; others $\rho = 0.10$) and *High* (first two factors $\rho = 0.60$; others $\rho = 0.10$).

5.3. BSO Hyperparameters

We systematically varied BSO's hyperparameters—number of bees, nest-site scout bees, proportion of scouts, and percentage of top solutions—to assess their effects under the three most challenging conditions in terms of optimization performance and consistency across seeds (see Table 1 for more details). Overall, hyperparameter changes only had a small effect on optimization performance, factor retention, and consistency. The only hyperparameter with a substantial and consistent impact was the number of bees, which mainly affected consistency across seeds while having only small effects on optimization performance and factor retention. Its impact on consistency was particularly pronounced in the massive noise condition with 5 factors, 5 indicators per factor, and high factor correlations. In this condition, the number of bees explained approximately 29% of the variance in results (Table S2). Increasing the swarm size to 1,000 bees raised average consistency to approximately 42% for the best hyperparameter combination (1,000 bees, 200 nest-site bees, 40% scouts, 10% top), whereas using 200 bees reduced consistency to 4% (200 bees, 200 nest-site bees, 40% scouts, 30% top). In less complex structural conditions with fewer ambiguous items, gains by using more bees were small, because 200 bees were already sufficient. In comparison, local fit showed minimal variance and remained a challenging criterion to achieve. Consequently, optimization performance varied little across runs (see Figure 3), whereas factor retention decreased slightly with more bees, because larger swarms more often discovered parsimonious solutions in which the highly correlated factor collapsed. More detailed results across all hyperparameter combinations are presented in Figure S5 (Panels A–D) and Table S2 in the online supplement.

6. Discussion

With this extensive simulation study, we systematically evaluated the performance of BSO. Overall, BSO successfully optimized the criteria entered in its objective function and showed high consistency across seeds; BSO reliably identified factor models with acceptable global fit, distinct factors, and—under many conditions—the correct number of factors, while effectively removing ambiguous and noise items. As expected, BSO achieved fewer optimal solutions under the most challenging conditions (e.g., loading patterns with many strong cross-loadings or high factor correlations), in which highly correlated factors were frequently merged, and the resulting models often had substantial remaining local misfit. In comparison, simpler EFA–CFA pipelines were less effective in identifying measurement models that fulfilled the full set of criteria defined in the objective function, indicating that BSO's systematic optimization offers advantages over conventional procedures. Although some EFA–CFA pipelines performed well on selected aspects (e.g., EFA Strict was also successful at removing ambiguous items), BSO outperformed them on the performance metrics, attaining higher overall ν values on the objective function, and producing more consistent results across samples.

The examination of hyperparameters in Simulation 2 revealed that the number of bees was the most influential hyperparameter, with larger swarms achieving better results. All other hyperparameters only had negligible effects on the

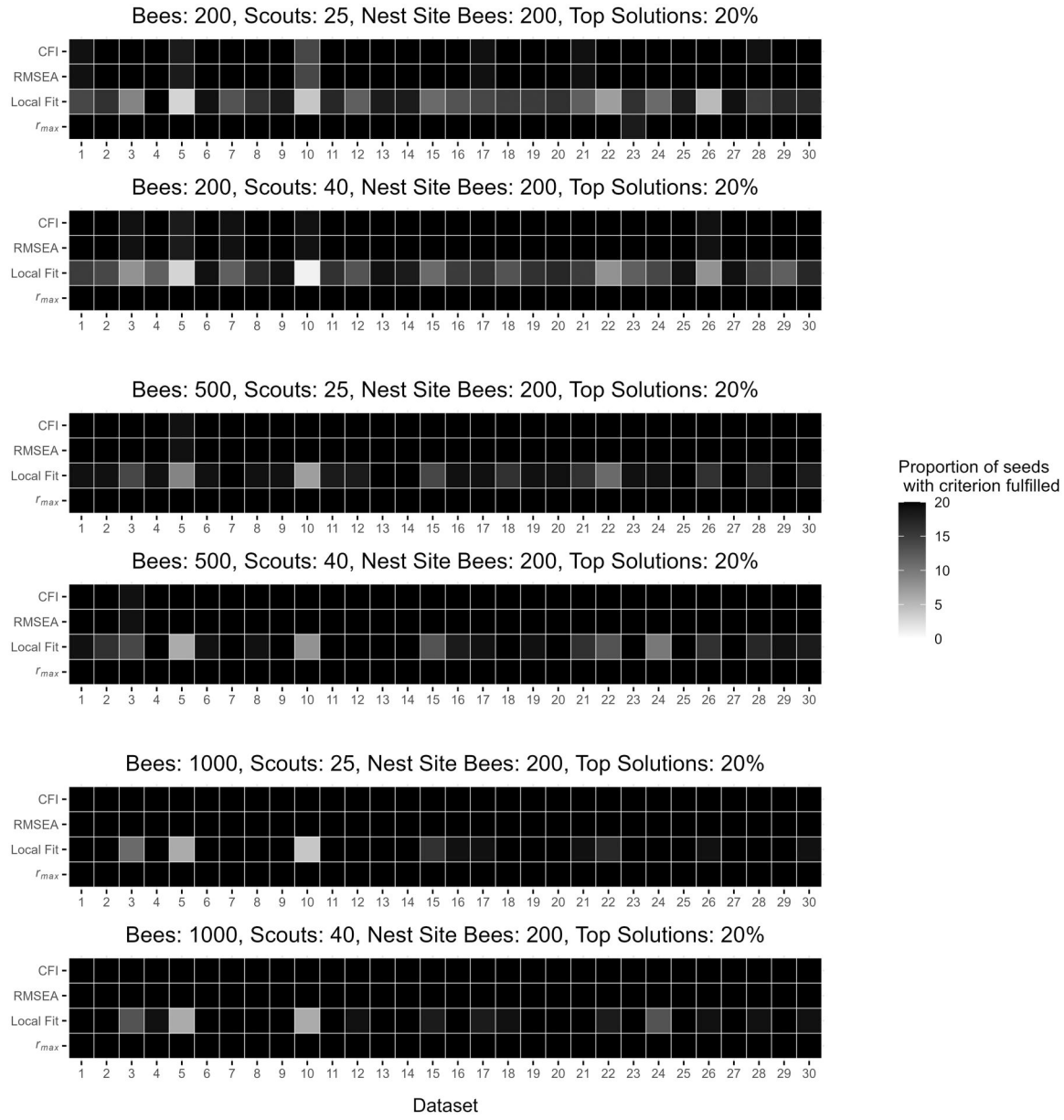

Figure 3. Proportion of criteria met across datasets for various hyperparameters.
Note. This figure shows how frequently models met specific criteria across different hyperparameter variations, with the number of bees and the percentage of scouts varying while the percentage of top solutions remains constant. Each square indicates the proportion of 20 seeds per dataset in the massive noise condition with 5 factors, 5 indicators per factor and $\rho = 0.6$ in which the criterion was met (i.e., nectar value $\nu > 0.5$). Black squares indicate that the criterion was met in all seeds; white squares indicate that the criterion was never met ($\nu < 0.5$). Gray shades reflect partial fulfillment. The criteria include CFI $\geq$ 0.95; RMSEA $\leq$ 0.06; Local Fit (the top 10% standardized values in residualized variance covariance matrix averaging below 2.58); maximal factor correlation (c) $\leq$ 0.8.

performance metrics. Increasing the number of bees also improved consistency across seeds, especially in settings with many ambiguous (noise) items. This is expected because more bees explore more candidate models, which increases the likelihood of identifying an optimal solution. Although the default settings used in Simulation 1 performed well across many conditions, they were less efficient and resulted in longer computation times than settings with fewer bees. From a practical perspective, we therefore recommend using at least five seeds with at least 500 bees, allocating 25% as scout bees, using 200 nest-site scout bees, and exploring the top 20% of solutions intensively in the next iteration. When substantial noise and ambiguity are expected—for example, in early stages of test development or when the item pool is large (more than 40 items)—the number of bees should be increased to 1,000 and the number of seeds set to 10 in order to improve the likelihood of finding an optimal model. Increasing the number of nest-site scouts is particularly useful when a wide range of factors must be considered or when the model search fails to yield satisfactory initial solutions.

BSO's performance advantage is presumably due to re-estimating the model during its non-greedy search process and continuously evaluating all criteria for each candidate model. This procedure prevents factor loadings and correlations from shifting due to item removal, an issue that affects

conventional EFA–CFA approaches because their item selection is based on initial loadings. Although sequential EFA–CFA pipelines that refit the model after each deletion can partially mitigate loading shifts, they remain greedy, introducing path dependence and sequence effects.

6.1. Criteria for Good Measurement Models in a CFA Context

The BSO algorithm performed exactly as intended from a technical perspective—it optimized the predefined criteria as far as possible. Specifying the optimization criteria is a non-trivial task because it requires careful consideration of what constitutes a good measurement model, how the relevant aspects can best be operationalized, and how they are weighed against each other. Depending on the purpose of the measurement, a good measurement model may include several aspects such as content coverage, model fit and parsimony, interpretability, and construct validity. In the present work, we primarily focus on model fit. Evaluating model fit—and, in particular, balancing global and local fit—is challenging and has been discussed from multiple angles in the recent literature (Lance et al., 2016; McNeish & Hancock, 2018; Thoemmes et al., 2018; Zhang & Wu, 2024). It has repeatedly been shown that the sensitivity of global fit measures depends on model complexity, model size, and the unique variance structure (Heene et al., 2011, 2012; Moshagen, 2012), particularly with non-adaptive cut-off values (Groskurth et al., 2024; McNeish & Wolf, 2023). However, global fit indices do not indicate whether model misfit is concentrated in specific parts of the model or distributed evenly across all model parameters. Therefore, we also incorporated local model fit into the automated model specification search by penalizing large, standardized residuals, which helps differentiate between models that appear equivalent based on global indices alone.

Beyond global and local fit, the quality of measurement models often depends on additional, sometimes competing criteria that warrant consideration in applied settings, though they were not explicitly evaluated in our simulations. These include the size of factor loadings, the reliability of factors, and construct coverage. Items with low factor loadings—often removed in some simple item selection strategies—contribute little to explaining factor variance (Brown & Moore, 2012), may negatively impact factor saturation (McDonald, 1999), and affect the usefulness of certain fit indices (e.g., CFI, SRMR; Heene et al., 2011). Defining a minimum standardized factor loading threshold (e.g., $\lambda \geq 0.3$), as already implemented in some applied metaheuristic-based studies, can mitigate such issues (e.g., Olaru & Jankowsky, 2022; Schroeders et al., 2024). However, an exclusive focus on high factor loadings and internal consistency may compromise construct coverage and the validity of the measure due to the attenuation paradox (e.g., Loevinger, 1954; Schroeders et al., 2016). Thus, model specification searches may balance these competing objectives by maintaining adequate loadings while also penalizing high inter-item correlations which may indicate redundant content and narrow the construct's representation (Steger et al., 2023). BSO's flexible objective function allows researchers to incorporate such criteria according to their specific requirements.

6.2. Additional Considerations in Deciding Between Competing Models

In metaheuristic model specification search, multiple candidate models are identified that yield almost identical results with respect to the overall fit function. Although these models are not mathematically equivalent—that is, they do not have the same model-implied covariance matrices (e.g., Raykov & Marcoulides, 2001)—they become essentially equivalent with respect to the selected optimization criteria. This quasi-equivalence introduces the challenge of determining which model to select among several competing measurement models. A practitioner may argue that the decision should be made on theoretical grounds which solution to prefer. Technically, this means adding some kind of information that was not part of the optimization process. One advantage of model specification search is that such implicit criteria can be made explicit and incorporated as an additional aspect of measurement quality. For example, lexical difficulty can introduce systematic bias in personality assessment (Altgassen et al., 2025). In this context, semantic embeddings, which capture the linguistic complexity (Wulff & Mata, 2025), could be integrated into the optimization function. Moreover, transformer-based sentence embedding models can predict the empirical associations between survey items without empirical testing (Hommel et al., 2025). As a proximate measure of construct validity they can be directly taken into account during the optimization process using metaheuristics. This example illustrates that metaheuristics enable psychologists to incorporate multiple aspects of measurement quality simultaneously.

Parsimony is another central aspect when comparing quasi-equivalent models. Parsimony is often assessed by the number of estimated parameters in relation to the degrees of freedom and has been widely discussed in the SEM literature (e.g., Mulaik, 2001; Partsch et al., 2025; Preacher, 2006; Raykov & Marcoulides, 1999). However, the underlying rationale—Occam's razor—is based on philosophical and epistemic rather than mathematical considerations. In factor model selection, an emphasis on parsimony can be placed by facilitating the merging of factors with very high correlations, but critics may argue that less restrictive models, such as ESEM which allow cross-loadings, better capture multidimensional constructs and more accurately reflect the relation between indicators and latent factors (Asparouhov & Muthén, 2009; Marsh et al., 2013; Morin et al., 2013). The decision between a parsimonious and a more complex model is ultimately a theoretical one and depends on the practical implications of the measurement. Notably, BSO is not restricted to a specific model, but can be extended to other methods. In the present study, we opted for a simple and interpretable CFA model by jointly selecting an item set for which simple structure is reasonably well satisfied, but if the

primary aim were to maximize predictive performance, the underlying modeling in BSO optimization would be different.

6.3. Limitations and Future Outlook

Several limitations of BSO warrant consideration. First, although performance has been validated across various complexity levels, it remains uncertain how these findings generalize to real-world scenarios involving substantially more items and factors, where the search space expands dramatically. The newly introduced nest-site scout bees represent a first step toward addressing this challenge by providing informed initial solutions through EFA rather than random search. This systematic warm start is particularly valuable in large search spaces, where identifying promising starting points through random sampling alone is extremely inefficient. Second, as with many metaheuristics and data-driven methods in general, the optimal model is tailored to the specific dataset, raising overfitting concerns, particularly when used exploratorily or with small sample sizes. Although BSO consistently identifies optimal solutions for a specific dataset, this within-sample optimality does not guarantee generalizability. Future studies could consider additional validation procedures (e.g., cross-validation) to ensure that the model generalizes well to new samples.

Several promising extensions could further strengthen the BSO framework. First, instead of using warm starts through nest-site scout bees, future implementations could allow users to specify theoretically motivated starting solutions that reflect substantive expectations. This option would enable BSO to function as a theory-inspired refinement tool, discovering alternative models that better satisfy the optimization criteria than the initial theoretically derived model. In this sense, BSO implements a similar idea as multiverse (Steegen et al., 2016) and specification-curve analyses (Simonsohn et al., 2020). Instead of exhaustively exploring theoretically questionable variants, the algorithm would efficiently diversify a specific part of the solution space, returning a manageable set of well-fitting alternatives.

Second, BSO may still be prone to local misspecifications that can propagate through the model and affect other parameters when SEM estimation is used. For example, neglected residual correlations between indicators can result in inflated factor correlations. More robust alternatives discussed in the literature—such as Model Implied Instrumental Variables (MIIVs; Bollen, 2019; Bollen et al., 2024) and the Structural After Measurement approach (SAM; Rosseel & Loh, 2024)—could help mitigate this propagation. Thus, combining such approaches with BSO may be a promising direction to accommodate local misfit.

Finally, BSO currently relies on a weighted aggregation of the individual optimization criteria into a single objective-function score. This approach can bias the model specification search toward specific regions of the solution space and requires explicit assumptions about the relative importance of the criteria. A Pareto-front approach would avoid such aggregation by treating the criteria separately. Pareto-based optimization identifies multiple non-dominated solutions that reflect different balances between model fit, parsimony, and other criteria (Ngatchou et al., 2005). Transitioning from the current weighted optimization to a Pareto-based multi-objective framework could therefore increase BSO's flexibility in exploring tradeoffs among criteria. Although implementing such a framework would require a substantial reconceptualization of the metaheuristic, it would allow researchers to explore a diverse set of optimal models with different criteria tradeoffs (Agrawal et al., 2008; Ngatchou et al., 2005).

7. Conclusion

We introduced an extension of the Bee Swarm Optimization algorithm for correlated factor models and demonstrated through a simulation study that it reliably finds optimal model specifications. The results confirm that BSO not only operates as intended but also efficiently navigates the expansive search space inherent in model specification problems with a variable number of factors and a simultaneous item selection procedure. Importantly, across all evaluated performance metrics, BSO consistently outperformed simpler EFA–CFA pipelines, underscoring its practical advantage over established approaches. Furthermore, our investigation of hyperparameter settings revealed that more bees lead to better performance in both consistency and accuracy, while the other hyperparameters only had a minor impact. Overall, these results establish BSO as a promising, data-driven tool for model specification, with potential avenues for further refinement and application in more complex, real-world scenarios.

Acknowledgements

We thank Anna Nikolei for her assistance in deriving the formula used to calculate the number of possible models.

Declaration of AI usage

ChatGPT (OpenAI, GPT-5) was used for language editing and proofreading.

Disclosure statement

We have no conflicts of interest to disclose. We confirm that the work conforms to Standard 8 of the American Psychological Association's Ethical Principles of Psychologists and Code of Conduct.

References

Achaa-Amankwaa, P., Trautwein, T., Lenhard, W., & Schroeders, U. (2025). Balancing between categorical and dimensional assessment in short-scale construction using ant colony optimization. *European*

Journal of Psychological Assessment. Advanced online publication. https://doi.org/10.1027/1015-5759/a000892

Agrawal, S., Dashora, Y., Tiwari, M. K., & Son, Y.-J. (2008). Interactive particle swarm: A Pareto-adaptive metaheuristic to multiobjective optimization. *IEEE Transactions on Systems, Man, and Cybernetics A*, *38*, 258–277. https://doi.org/10.1109/TSMCA.2007.914767

Altgassen, E., Schittenhelm, C., & Wilhelm, O. (2025). Words don't come easy: How lexical difficulty of items and vocabulary of subjects (not) affect personality assessment. *Journal of Personality and Social Psychology*, *129*, 910–932. https://doi.org/10.1037/pspp0000572

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*, 397–438. https://doi.org/10.1080/10705510903008204

Bader, M., & Moshagen, M. (2025). Assessing the fitting propensity of factor models. *Psychological Methods*, *30*, 254–270. https://doi.org/10.1037/met0000529

Blum, C., & Roli, A. (2003). Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Computing Surveys*, *35*, 268–308. https://doi.org/10.1145/937503.937505

Bollen, K. A. (2019). Model implied instrumental variables (MIIVs): An alternative orientation to structural equation modeling. *Multivariate Behavioral Research*, *54*, 31–46. https://doi.org/10.1080/00273171.2018.1483224

Bollen, K. A., Gates, K. M., & Luo, L. (2024). A model implied instrumental variable approach to exploratory factor analysis (MIIV-EFA). *Psychometrika*, *89*, 687–716. https://doi.org/10.1007/s11336-024-09949-6

Bonifay, W., Cai, L., Falk, C. F., & Preacher, K. J. (2025). Reassessing the fitting propensity of factor models. *Psychological Methods*. Advanced online publication. https://doi.org/10.1037/met0000735

Brown, T. A., & Moore, M. T. (2012). Confirmatory factor analysis. *In* R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 361–379). Guilford Press.

Castle, J. L., Doornik, J. A., & Hendry, D. F. (2011). Evaluating automatic model selection. *Journal of Time Series Econometrics*, *3*, 8. https://doi.org/10.2202/1941-1928.1097

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*(7), 1–9. https://doi.org/10.7275/jyj1-4868

DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment*, *23*, 225–241. https://doi.org/10.1177/073428290502300303

Dorigo, M., & Stützle, T. (2010). Ant colony optimization: Overview and recent advances. *In* M. Gendreau & J.-Y. Potvin (Eds.), *Handbook of metaheuristics* (pp. 227–263). Springer US. https://doi.org/10.1007/978-1-4419-1665-5_8

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. Oxford University Press.

Genz, A., & Bretz, F. (2009). *Computation of multivariate normal and t probabilities*. Springer.

Goretzko, D., Pham, T. T. H., & Bühner, M. (2021). Exploratory factor analysis: Current use, methodological developments, and recommendations for good practice. *Current Psychology*, *40*, 3510–3521. https://doi.org/10.1007/s12144-019-00300-2

Groskurth, K., Bluemke, M., & Lechner, C. M. (2024). Why we need to abandon fixed cutoffs for goodness-of-fit indices: An extensive simulation and possible solutions. *Behavior Research Methods*, *56*, 3891–3914. https://doi.org/10.3758/s13428-023-02193-3

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*, 319–336. https://doi.org/10.1037/a0024917

Heene, M., Hilbert, S., Freudenthaler, H. H., & Bühner, M. (2012). Sensitivity of SEM fit indexes with respect to violations of uncorrelated errors. *Structural Equation Modeling*, *19*, 36–50. https://doi.org/10.1080/10705511.2012.634710

Hommel, B. E., Külpmann, A. I., & Arslan, R. C. (2025). The Synthetic Nomological Net: A search engine to identify conceptual overlap in measures in the behavioral sciences. PsyArXiv. https://doi.org/10.31234/osf.io/nfgmv_v1

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

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling*, *23*, 555–566. https://doi.org/10.1080/10705511.2016.1154793

Jankowsky, K., Olaru, G., & Schroeders, U. (2020). Compiling measurement invariant short scales in cross-cultural personality assessment using ant colony optimization. *European Journal of Personality*, *34*, 470–485. https://doi.org/10.1002/per.2260

Jing, Z., Kuang, H., Leite, W. L., Marcoulides, K. M., & Fisk, C. L. (2022). Model specification searches in structural equation modeling with a hybrid ant colony optimization algorithm. *Structural Equation Modeling*, *29*, 655–666. https://doi.org/10.1080/10705511.2021.2020119

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2013). On the shortcomings of shortened tests: A literature review. *International Journal of Testing*, *13*, 223–248. https://doi.org/10.1080/15305058.2012.703734

Lance, C. E., Beck, S. S., Fan, Y., & Carter, N. T. (2016). A taxonomy of path-related goodness-of-fit indices and recommended criterion values. *Psychological Methods*, *21*, 388–404. https://doi.org/10.1037/met0000068

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin*, *51*, 493–504. https://doi.org/10.1037/h0058543

Marcoulides, G. A., & Drezner, Z. (2003). Model specification searches using ant colony optimization algorithms. *Structural Equation Modeling*, *10*, 154–164. https://doi.org/10.1207/s15328007sem1001_8

Marcoulides, G. A., & Leite, W. (2013). Exploratory data mining algorithms for conducting searches in structural equation modeling: A comparison of some fit criteria. *In* J. J. McArdle & G. Ritschard (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences* (pp. 150–171). Routledge.

Marcoulides, G. A., & Schumacker, R. E. (2001). *New developments and techniques in structural equation modeling*. Psychology Press.

Marcoulides, G. A., Drezner, Z., & Schumacker, R. E. (1998). Model specification searches in structural equation modeling using tabu search. *Structural Equation Modeling*, *5*, 365–376. https://doi.org/10.1080/10705519809540112

Marcoulides, G. A., Ing, M., & Hoyle, R. (2012). Automated structural equation modeling strategies. *In* R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 690–704). Guilford Press.

Marcoulides, K. M., & Falk, C. F. (2018). Model specification searches in structural equation modeling with R. *Structural Equation Modeling*, *25*, 484–491. https://doi.org/10.1080/10705511.2017.1409074

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment*, *22*, 471–491. https://doi.org/10.1037/a0019227

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J., & Von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right—camouflaging misspecification with item parcels in CFA models. *Psychological Methods*, *18*, 257–284. https://doi.org/10.1037/a0032773

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*, 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students'

evaluations of university teaching. *Structural Equation Modeling*, *16*, 439–476. https://doi.org/10.1080/10705510903008220
McDonald, R. P. (1999). *Test theory: A unified treatment*. Psychology Press.
McNeish, D., & Hancock, G. R. (2018). The effect of measurement quality on targeted structural model fit indices: A comment on Lance, Beck, Fan, and Carter (2016). *Psychological Methods*, *23*, 184–190. https://doi.org/10.1037/met0000157
McNeish, D., & Wolf, M. G. (2023). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*, *28*, 61–88. https://doi.org/10.1037/met0000425
Morin, A. J. S., Marsh, H. W., & Nagengast, B. (2013). Exploratory structural equation modeling. *In* G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 395–436). Information Age.
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*, 86–98. https://doi.org/10.1080/10705511.2012.634724
Mulaik, S. A. (2001). The curve-fitting problem: An objectivist view. *Philosophy of Science*, *68*, 218–241. https://doi.org/10.1086/392874
Ngatchou, P., Zarei, A., & El-Sharkawi, M. A. (2005). Pareto multi objective optimization. *Proceedings of the 13th International Conference on Intelligent Systems Application to Power Systems*. https://doi.org/10.1109/ISAP.2005.1599245
Olaru, G., & Jankowsky, K. (2022). The HEX-ACO-18: Developing an age-invariant HEXACO short scale using ant colony optimization. *Journal of Personality Assessment*, *104*, 435–446. https://doi.org/10.1080/00223891.2021.1934480
Olaru, G., Schroeders, U., Hartung, J., & Wilhelm, O. (2019). Ant colony optimization and local weighted structural equation modeling: A tutorial on novel item and person sampling procedures for personality research. *European Journal of Personality*, *33*, 400–419. https://doi.org/10.1002/per.2195
Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: Testing competing models for designing short-scale Big-Five assessments. *Journal of Research in Personality*, *59*, 56–68. https://doi.org/10.1016/j.jrp.2015.09.001
Partsch, M. V., Sterner, P., & Goretzko, D. (2025). A simulation study on the interaction effects of underfactoring and nuisance parameters on model fit indices. *Structural Equation Modeling*, *32*, 1000–1015. https://doi.org/10.1080/10705511.2025.2514592
Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research*, *41*, 227–259. https://doi.org/10.1207/s15327906mbr4103_1
R Core Team. (2025). *R: A language and environment for statistical computing*. https://www.R-project.org/
Raborn, A. W., & Leite, W. L. (2018). ShortForm: An R package to select scale short forms with the ant colony optimization algorithm. *Applied Psychological Measurement*, *42*, 516–517. https://doi.org/10.1177/0146621617752993
Raborn, A. W., Leite, W. L., & Marcoulides, K. M. (2020). A comparison of metaheuristic optimization algorithms for scale short-form development. *Educational and Psychological Measurement*, *80*, 910–931. https://doi.org/10.1177/0013164420906600
Raykov, T., & Marcoulides, G. A. (1999). On desirability of parsimony in structural equation model selection. *Structural Equation Modeling*, *6*, 292–300. https://doi.org/10.1080/10705519909540135
Raykov, T., & Marcoulides, G. A. (2001). Can there be infinitely many models equivalent to a given covariance structure model? *Structural Equation Modeling*, *8*, 142–149. https://doi.org/10.1207/s15328007sem0801_8
Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively reweighted least squares to the Rosenberg self-esteem scale. *Multivariate Behavioral Research*, *51*, 818–838. https://doi.org/10.1080/00273171.2016.1243461
Revelle, W. (2024). psych: Procedures for psychological, psychometric, and personality research *(Version 2.5.3) [R package]*. Northwestern University. https://CRAN.R-project.org/package=psych
Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02
Rosseel, Y., & Loh, W. W. (2024). A structural after measurement approach to structural equation modeling. *Psychological Methods*, *29*, 561–588. https://doi.org/10.1037/met0000503
Sass, D. A., & Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research*, *45*, 73–103. https://doi.org/10.1080/00273170903504810
Sass, D. A., & Sanchez, M. A. (2023). A variable selection algorithm for creating replicable factor structures. *Multivariate Behavioral Research*, *58*, 48–70. https://doi.org/10.1080/00273171.2021.1948814
Scharf, F., & Nestler, S. (2019a). A comparison of simple structure rotation criteria in temporal exploratory factor analysis for event-related potential data. *Methodology*, *15*, 43–60. https://doi.org/10.1027/1614-2241/a000175
Scharf, F., & Nestler, S. (2019b). Should regularization replace simple structure rotation in exploratory factor analysis? *Structural Equation Modeling*, *26*, 576–590. https://doi.org/10.1080/10705511.2018.1558060
Schmitt, T. A., & Sass, D. A. (2011). Rotation criteria and hypothesis testing for exploratory factor analysis: Implications for factor pattern loadings and interfactor correlations. *Educational and Psychological Measurement*, *71*, 95–113. https://doi.org/10.1177/0013164410387348
Schroeders, U., Scharf, F., & Olaru, G. (2024). Model specification searches in structural equation modeling using bee swarm optimization. *Educational and Psychological Measurement*, *84*, 40–61. https://doi.org/10.1177/00131644231160552
Schroeders, U., Wilhelm, O., & Olaru, G. (2016). Meta-heuristics in short scale construction: Ant colony optimization and genetic algorithm. *PLOS One*, *11*, e0167110. https://doi.org/10.1371/journal.pone.0167110
Seeley, T. D. (2011). *Honeybee democracy*. Princeton University Press.
Seeley, T. D., & Visscher, P. K. (2004). Group decision making in nest-site selection by honey bees. *Apidologie*, *35*, 101–116. https://doi.org/10.1051/apido:2004004
Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*, 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*, 702–712. https://doi.org/10.1177/1745691616658637
Steger, D., Jankowsky, K., Schroeders, U., & Wilhelm, O. (2023). The road to hell is paved with good intentions: How common practices in scale construction hurt validity. *Assessment*, *30*, 1811–1824. https://doi.org/10.1177/10731911221124846
Thoemmes, F., Rosseel, Y., & Textor, J. (2018). Local fit evaluation of structural equation models using graphical criteria. *Psychological Methods*, *23*, 27–41. https://doi.org/10.1037/met0000147
Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *The Journal of Experimental Education*, *80*, 26–44. https://doi.org/10.1080/00220973.2010.531299
Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *76*, 913–934. https://doi.org/10.1177/0013164413495237
Wulff, D. U., & Mata, R. (2025). Semantic embeddings reveal and address taxonomic incommensurability in psychological measurement. *Nature Human Behaviour*, *9*, 944–954. https://doi.org/10.1038/s41562-024-02089-y
Zhang, X., & Wu, H. (2024). Investigating structural model fit evaluation. *Structural Equation Modeling*, *31*, 863–881. https://doi.org/10.1080/10705511.2024.2350023
Zimny, L., Schroeders, U., & Wilhelm, O. (2024). Ant colony optimization for parallel test assembly. *Behavior Research Methods*, *56*, 5834–5848. https://doi.org/10.3758/s13428-023-02319-7

Appendix

Optimization Criteria

The optimization criteria for the BSO were standardized across all simulation conditions. A logit transformation was applied to enhance differentiation near cutoff values and to scale to the range between 0 and 1 (e.g., Schroeders et al., 2016). The inverse logit-transformed criteria were labeled as nectar values (ν). Slopes were set within a commonly used range of 15–100 (e.g., Olaru et al., 2019; Schroeders et al., 2016, 2024) and tested through pilot trials, as empirical estimations (e.g., Zimny et al., 2024) were impractical given the high number of datasets and potential model configurations. A weighted approach (with weights specified in the weighting function, Equation A8) was adopted for the following 5 criteria: (a) global model fit, (b) local model fit, (c) maximum factor correlation, (d) number of retained items, and (e) a minimum threshold for item retention. It should be noted that although multi-objective optimization is beneficial, poor weighting during the aggregation of criteria into a single objective can introduce bias (Ngatchou et al., 2005).

Global model fit was assessed using the Comparative Fit Index (CFI), and Root Mean Square Error of Approximation (RMSEA), with thresholds of CFI $\geq$ 0.95 and RMSEA $\leq$ 0.06 indicating good fit, as defined by the following formulas:

$$\nu_{CFI} = \frac{1}{1 + \exp\left(30 \cdot (.95 - CFI)\right)} \tag{A1}$$

$$\nu_{RMSEA} = \frac{1}{1 + \exp\left(-50 \cdot (.06 - RMSEA)\right)} \tag{A2}$$

Local model fit was evaluated by penalizing the highest 10% of values in the standardized residual variance-covariance matrix. The function is defined as:

$$\nu_{localfit} = \frac{1}{1 + \exp\left(0.2 \cdot (2.58 \cdot n_{pen} - sum_{res\_max})\right)} \tag{A3}$$

with

$$n_{pen} = 0.1 \cdot \frac{n_{items} \cdot (n_{items} - 1)}{2} \tag{A4}$$

The components of (A3) are: n_{pen} is the number of residuals entering the penalty. It is defined as 10% of all unique pairwise values in the standardized residual variance-covariance matrix. This focuses the criterion on penalizing only clear local misfit, while keeping it robust to a few random large residuals, as only the worst values are considered. sum_{res_max} is the sum of the n_{pen} largest standardized residuals. Large values of sum_{res_max} indicate large local misspecifications. 2.58 is the critical value of the standard normal distribution for $\alpha = 1\%$ (two-tailed). The product $2.58 \cdot n_{pen}$ therefore corresponds to the sum we would expect if all penalized residuals were just at the 1% cutoff. Thus, sum_{res_max} values clearly above this threshold are treated as evidence of poor local fit. In other words, if $\nu_{localfit} \approx 1$, the largest residuals are on average below what is expected under the 1% cutoff, indicating good local fit. Values approaching 0 indicate many large values in the standardized residual variance-covariance matrix (poor local fit)

The maximum factor correlation was restricted with a threshold of 0.80 to prevent multicollinearity and to promote model parsimony by merging factors exceeding the threshold:

$$\nu_{r_{max}} = \frac{1}{1 + \exp\left(-30 \cdot (.80 - r_{max})\right)} \tag{A5}$$

The fourth criterion, item retention, favored models preserving at least 75% of items, discouraging excessive item removal for fit improvement, assuming a later stage of test development:

$$\nu_{n_{items}} = \frac{1}{1 + \exp(25 \cdot (.75 - n_{items}))} \tag{A6}$$

Additionally, the minimum threshold criterion ensured sufficient item retention, setting a floor at y = 67% of the original items (or y = 50% for the massive noise condition). Models below this threshold received an (overall) nectar value of zero. Thus, the minimum scale length depends on the condition:

$$\nu_{n_{itemsmin}} = \begin{cases} 1 \text{ if } n_{items} \geq n_{max\_items} \cdot y \\ 0 \text{ if } n_{items} < n_{max\_items} \cdot y \end{cases} \tag{A7}$$

The overall optimization criteria, $\nu_{overall}$ was defined as a weighted sum of these criteria, prioritizing r_{max} and global fit indices:

$$\nu_{overall} = 2 \cdot \nu_{CFI} + 2 \cdot \nu_{RMSEA} + 3 \cdot \nu_{r_{max}} + \nu_{n_{items}} + \nu_{residuals} \tag{A8}$$

Before the simulation, pilot runs were conducted to validate the value ranges of the different criteria and the appropriateness of the weighting, both building the basis for the optimization process and the convergence behavior.