

---

# Specification Search in Structural Equation Modeling with SEM Forests

---



Dissertation zum Erwerb des Doctor of Philosophy (Ph.D.)  
am Munich Center of the Learning Sciences  
der Ludwig-Maximilians- Universität  
München

Vorgelegt von  
John Alexander Silva Díaz  
2025

Referent/in: Prof. Dr. Moritz Heene  
Korreferent/in: Prof. Dr. Stefan Ufer  
Tag der mündlichen Prüfung: 18.07.2025

## **Acknowledgements**

First and foremost, I am deeply grateful to my supervisors, Prof. Dr. Moritz Heene, Prof. Dr. Stefan Ufer, and Prof. Dr. Ronny Scherer, for their invaluable guidance, constructive criticism, and unwavering support throughout my doctoral journey. I am also profoundly thankful to Prof. Dr. Andreas Brandmaier, whose invaluable expertise, wisdom, and the tremendous privilege of having him as a coauthor made this dissertation possible. I am especially grateful to Moritz for his patience, understanding, kindness, and unconditional support throughout this long and challenging journey, one that involved navigating the complexities of a global pandemic and the unexpected challenges it brought.

To my beloved wife, whose unwavering support, understanding, and love have been my anchor throughout this demanding journey. To my daughter, whose laughter and curiosity have been a constant reminder of the joy and wonder in life. I dedicate this work to you, with hopes that it inspires you to pursue your own dreams with passion and determination. To my parents, thank you for your unconditional love, encouragement, and sacrifices. Your trust and confidence in me have been a constant source of strength and motivation.

I also want to express my deepest gratitude to Prof. Dr. Javier Alejandro Corredor Aristizábal, whom I consider a true game-changer in my academic journey and in the careers of many Colombian psychologists. His passion for science and the academic spaces he created in our Universidad Nacional de Colombia provided the ideal environment to launch our careers. His mentorship and dedication have profoundly shaped my approach to research and learning.

Last, the research presented in this work was supported by the Elite Network of Bavaria [Project number: K-GS-2012-209]. I would like to extend my sincere gratitude to the opportunities (conference attendance, incubator stay, international knowledge exchange) made possible by the ENB.

## **Abstract**

Model misspecification is a prevalent challenge in applied SEM, often requiring specification search to improve model fit. Traditional approaches, such as modification indices, are limited to variables already included in the model and are therefore ineffective at detecting omitted influential variables and their interaction effects. To address these limitations, the two studies presented in this dissertation introduce SEM forests as a novel and robust technique for specification search in SEM. The first study evaluates the performance of SEM forests to identify unique, mixed, and interaction covariate paths across different factor loading magnitudes, covariate path magnitudes, and sample sizes. The results indicate that SEM forests consistently do not incorrectly identify noninfluential omitted covariate paths under all examined conditions and accurately identify influential omitted covariate paths in multiple condition combinations explored, particularly when covariate-latent variable regression coefficients are strong and sample sizes are large. The second study provides a step-by-step guide for using SEM forests with the *semTree* R package, covering data preparation, model specification, forest generation, results interpretation, and model respecification. This practical guide equips researchers with the tools to apply SEM forests for specification search in SEM, addressing the limitations of traditional methods regarding omitted variables. Together, these studies demonstrate SEM forests as a robust alternative for specification search, enabling the identification of omitted influential covariates and interactions that traditional methods may overlook, ultimately enhancing the validity and reliability of SEM models.

## Table of Contents

<b>1. Preamble .....</b>	<b>1</b>
<b>2. General Introduction .....</b>	<b>2</b>
2.1. Structural Equation Modeling .....	3
2.1.1. Key Concepts in SEM .....	4
2.1.2. SEM Model Fit .....	6
2.1.3. Specification Search in SEM .....	8
2.2. SEM Trees and Forests .....	11
2.2.1. Ensemble SEM Trees: The SEM Forests Technique .....	13
<b>3. Study 1: Evaluation of Structural Equation Model Forests' Performance to Identify Omitted Influential Covariates .....</b>	<b>16</b>
3.1. Abstract .....	16
3.2. Introduction .....	16
3.3. Decision Trees and SEM Trees .....	18
3.4. Ensemble Methods and SEM Forests .....	20
3.4.1. SEM Forests Variable Importance Output .....	22
3.5. Questions and Hypotheses .....	23
3.6. Procedure.....	24
3.6.1. Population Structural Equation Model .....	24
3.6.2. Experimental Design .....	26
3.6.3. Data Generation.....	26
3.6.4. SEM Forests Growth .....	27

3.7. Results .....	28
3.8. Discussion .....	35
3.9. Data Availability Statement .....	40
3.10. References Study 1 .....	41
3.11. Appendices .....	47
3.11.1. Appendix A – Study 1 .....	47
3.11.2. Appendix B – Study 1 .....	478
<b>4. Study 2: A Practical Guide to Use SEM Forests for Specification Search in Structural Equation Modeling.....</b>	<b>49</b>
4.1. Abstract .....	49
4.2. Introduction .....	49
4.3. SEM Forests as a Tool for SEM Specification Search.....	51
4.4. A Running Example .....	52
4.5. A Practical Guide for Model Specification Search with SEM Forests .....	53
4.5.1. Model Specification.....	54
4.5.2. SEM Forests Growing .....	56
4.5.3. Importance Analysis .....	59
4.5.4. Model Respecification.....	62
4.5.5. Evaluation of Model Fit.....	66
4.6. Concluding Remarks .....	68
4.7. Compliance with Ethical Standards .....	71
4.8. References Study 2.....	72

4.9. Appendix – Study 2.....	76
<b>5. General Conclusions .....</b>	<b>79</b>
<b>6. References for the General Introduction and Conclusions.....</b>	<b>82</b>
<b>7. List of Figures.....</b>	<b>87</b>
<b>8. List of Tables .....</b>	<b>88</b>

## **1. Preamble**

This dissertation consists of two academic articles. Study 1 was published as: Silva Díaz, J. A., Heene, M., & Brandmaier, A. M. (2024). Evaluation of Structural Equation Model Forests' Performance to Identify Omitted Influential Covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–13.  
<https://doi.org/10.1080/10705511.2024.2417866>.

Study 2 was submitted to “Structural Equation Modeling: A Multidisciplinary Journal” as an article entitled “A Practical Guide to Use SEM Forests for Specification Search in Structural Equation Modeling” and is currently under review. The two articles are first authored by John Alexander Silva Díaz.



## 2. General Introduction

Structural Equation Modeling (SEM) is one of the most widely applied statistical techniques in the social sciences, offering a powerful framework for modeling complex relationships between observed and latent variables. A fundamental aspect of SEM is evaluating how well a specified model fits the observed data. However, in practice, it is quite common that initial SEM models do not achieve an adequate fit, requiring further model refinement. When a model fails to adequately represent the data, researchers typically conduct a specification search, a process aimed at identifying alternative variable relationships that may improve model fit (Kline, 2016; Marcoulides & Falk, 2018).

Traditional specification search methods often suffer from several limitations, including overfitting and an increased risk of capitalizing on chance characteristics of the data, which can compromise the validity and generalizability of SEM models (MacCallum et al., 1992). Given these challenges, there is a clear need for more systematic, data-driven approaches for specification search. Such approaches should not only mitigate these issues but also ensure that model adjustments maintain the conceptual integrity of SEM while improving its empirical accuracy.

This dissertation proposes the use of SEM forests as a novel approach to conducting specification searches in SEM. The first study that this dissertation comprises evaluates the performance of SEM forests as a specification search tool under varying data conditions, assessing their accuracy, robustness, and applicability. The second study provides a comprehensive step-by-step guideline for researchers interested in implementing SEM forests in their own specification search processes, ensuring accessibility and practical usability of this method within the SEM framework. Through these contributions, this research aims to advance methodological approaches in SEM, enhancing both the reliability and efficiency of model specification procedures. Before presenting the two primary studies that constitute this

dissertation, it is essential to provide an overview of SEM, specification search, and SEM forests. These three areas serve as the theoretical foundation of this work, offering the necessary framework for understanding the methodologies and analyses employed throughout the dissertation.

## **2.1. Structural Equation Modeling**

SEM is a statistical tool based on factor analysis, path analysis, and simultaneous equation modeling, that allows for establishing linear, non-linear or interaction effects between latent and observed variables (Hancock & Mueller, 2013; Holbert & Stephenson, 2002; Mulaik, 2009). These variables can include different measurement scales, such as dichotomous, ordinal, or continuous data. SEM assesses the fit of the hypothesized structure of variable relations—the proposed model—with empirical data (Holbert & Stephenson, 2002; Mulaik, 2009). Kline (2016) characterizes SEM as a disconfirmatory approach, capable of rejecting false models but not proving their validity.

The origins of SEM can be traced to Sewall Wright's development of path analysis and Charles Spearman's introduction of factor analysis, both formulated in the early 20th century, as well as to the statistical advancements pioneered by Karl Jöreskog in the 1970s (Hancock & Mueller, 2013). Jöreskog's key contributions include the description of the mathematical solution for maximum-likelihood estimation of model parameters using the variance-covariance matrix, and the development of a formal chi-square test that enabled model fit evaluation by comparing observed and hypothesized relationships, allowing statistical disconfirmation of proposed models (Jöreskog, 1970; Mulaik, 2009). Furthermore, Jöreskog (1970) marked one of the earliest applications of LISREL (linear structure relations), the pioneering software developed for SEM analysis. Notably, Jöreskog (1970) illustrated his proposed methodology using an application in educational research, developing

a model to predict achievement in mathematics and science based on the verbal and quantitative scales of the Scholastic Aptitude Test (SCAT).

Since its inception, SEM has evolved into one of the most widely employed statistical techniques in the social and behavioral sciences. For instance, between 1994 and 2001, over 60% of articles published by the American Psychological Association (APA) used SEM (Hershberger, 2003). More recently, Bollen et al. (2022) reviewed the application of SEM in articles published in the journal *Social Science Research* between 2011 and 2022, identifying its use in measuring abstract constructs, assessing measurement invariance, validating new scales and indicators, modeling various relationships between latent variables (e.g., mediation and moderation effects), and predicting membership in latent classes, among other applications. The versatility of SEM for this wide range of analytical purposes has been instrumental for its popularization over the past decades. The increasing prevalence of SEM has also been largely driven by methodological and computational advancements that have enhanced its accessibility. Early SEM software, for instance, required specialized expertise, such as proficiency in matrix algebra, limiting its use to researchers with advanced methodological training. However, subsequent developments have facilitated broader adoption across various disciplines with user-friendly software, such as the free, open-source *R* package *lavaan* (Rosseel, 2012), which is employed throughout this dissertation.

### ***2.1.1. Key Concepts in SEM***

SEM allows for representing relationships among variables through a series of regression equations, estimating and testing relationships between exogenous and endogenous latent variables and their observed indicators. Specifically, SEM estimates parameters associated with the measurement model, which specifies the relationships between latent variables and their indicators, and the structural model, which defines the relationships among latent variables (Holbert & Stephenson, 2002; Kline, 2016; Mulaik,

2009). Key parameters in SEM include factor loadings ( $\lambda$ ), which indicate the magnitude of the relation between latent variables and indicators, and regression coefficients ( $\beta$ ), which quantify the strength and the direction of the relationship between exogenous and endogenous variables. Among other parameters, SEM also estimates covariances and correlations between latent variables, as well as error terms for both latent variables and indicators, which account for the unexplained variance in the model. The estimation of these parameters is typically performed using Maximum Likelihood (ML), provided that certain distributional assumptions hold (e.g., multivariate normality). Alternative estimation methods, including Generalized Least Squares (GLS) and Bayesian approaches, are also commonly employed (Mulaik, 2009).

The complete process to use SEM is described in detail in the second study that this dissertation encompasses (p. 37). As an introduction, users must first specify a model that represents the hypothesized relationships between exogenous and endogenous latent variables, as well as between latent variables and their indicators. To enable parameter estimation, the proposed model must be identified, meaning that unique parameter values can be determined from the observed data and the model's constraints (Kline, 2016; Mulaik, 2009). An identified model requires more unique pieces of information provided by the data (e.g., variances and covariances of observed variables) than unknown parameters to be estimated (Lei & Wu, 2007). If the model is underidentified, that is, unique parameter estimates cannot be obtained, it must be modified to achieve identification. Once model identification is established, practitioners can proceed with the parameter estimation.

In the case of ML estimation, initial values are heuristically assigned to the free parameters of the model, generating a model-implied covariance matrix based on these initial values. The final objective of ML estimation is to minimize the differences between the model-implied covariance matrix and the covariance matrix derived from the collected data.

Through an iterative optimization process, ML adjusts parameter estimates to reduce this discrepancy function. The process continues until the algorithm identifies an optimal set of parameter values, achieving model convergence (Hancock & Mueller, 2013; Mulaik, 2009). However, some models may fail to converge due to multiple causes (e.g., underidentified models, small sample size, multicollinearity, poor or extreme parameter starting values, etc.). When a model converges the discrepancy function stops changing significantly with additional parameter adjustments (Hancock & Mueller, 2013; Kline, 2016). Mathematically, this criterion is met when the difference between the fitting function values across successive iterations falls below a predefined threshold (e.g.,  $10^{-6}$ ).

### **2.1.2. SEM Model Fit**

If the model converges, practitioners must assess its goodness-of-fit. Model fit determines how well the model explains the observed data and can be interpreted as a statistical hypothesis test, assuming the null hypothesis is that the model fits the data (Kline, 2016; Lei & Wu, 2007). A model is considered to have an acceptable goodness-of-fit when the differences between the implied and observed covariance matrices are minimal, with the understanding that these differences arise from sampling errors and model-imposed constraints (Tomarken & Waller, 2003). There is an ongoing debate about the best strategies for assessing model fit in SEM, and no universally accepted rule of thumb exists regarding which statistics and thresholds should be used. Fit statistics can be categorized into model test statistics, usually chi-square, and a huge variety of approximate fit indices (Kline, 2016). The chi-square statistic assesses whether the magnitude of the difference between the observed covariance matrix and the implied covariance matrix can be explained by sampling error (Mulaik, 2009). However, this statistic is highly sensitive to data conditions such as sample size and parameter magnitudes, and do not provide useful information regarding the degree of the misfit (Kline, 2016; Tomarken & Waller, 2003; Van Voorhis & Morgan, 2007).

Approximate fit indices are not significance tests but continuous measures that indicate how well the model fits the data, based on different criteria depending on the index used. Although fit indices are expected to be sensitive to the magnitude of misspecification while remaining robust to other data-related factors (e.g., sample size, number of indicators per construct, location of misspecified parameters), this ideal is far from reality (Tomarken & Waller, 2003). The two main studies presented in this dissertation employed the Bentler Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR), along with the chi-square test.

CFI is an incremental fit index that compares the proposed model against a null model, which assumes zero covariances among the observed variables (Bentler, 1990). CFI values range from zero to one, with one representing the best possible fit. However, CFI performance is affected by model characteristics, such as the number of variables, having a worse performance in models with a large number of variables (Kenny & McCoach, 2003). RMSEA is an absolute fit index that measures how well the proposed model fits the population covariance matrix, considering the number of parameters in the model to avoid penalizing parsimonious models (Browne & Cudeck, 1992). RMSEA values range from zero to one, with zero indicating the best possible fit. However, RMSEA tends to overestimate misfit in small samples (Kline, 2016), and is negatively affected by an increase in the degrees of freedom (Heene et al., 2011). As for SRMR, it is also an absolute fit index that measures the average standardized difference between the observed correlation matrix and the model-implied correlation matrix (Bentler, 1995). SRMR values range from zero to one, with zero indicating the best possible fit. Hu and Bentler (1999) note that SRMR is influenced by model complexity, having lower SRMR values in more complex models. Heene et al. (2011) also reported that the magnitude of the factor loadings influences the power of the SRMR, RMSEA, CFI, and chi-square test to detect model misfit.

In the second main study presented in this dissertation, the Akaike Information Criterion (AIC; Akaike, 1974), and the Bayesian Information Criterion (BIC, Schwarz, 1978) were also employed as model selection criteria. Both AIC and BIC are predictive fit indices that prioritize model parsimony, with BIC imposing a stronger penalty on model complexity than AIC, particularly when sample sizes are large (Kline, 2016). For both indices, the model with the smallest AIC or BIC value is preferred. However, Preacher and Merkle (2012) have highlighted that BIC can be sensitive to sampling error, leading to substantial variability in BIC values even in large samples. Furthermore, Sugiura (1978) provides a mathematical proof demonstrating that AIC tends to favor more complex models.

Regarding the thresholds for interpreting the fit indices reported in the main studies of this dissertation (i.e., CFI, RMSEA, and SRMR), Hu and Bentler (1999) recommend interpreting these indices collectively, with acceptable fit indicated by  $CFI \geq .95$ ,  $RMSEA \leq .06$ , and  $SRMR \leq .08$ . It is crucial to emphasize that these thresholds should not be viewed as rule of thumbs but rather as guidelines proposed by the authors based on simulation studies. Whether following these criteria or alternative ones, practitioners must determine an appropriate strategy for assessing the adequacy of their models' goodness-of-fit. If the model does not fit the data, which is a usual scenario in SEM, a specification search is necessary to continue with the SEM analysis. Specification search is a fundamental component of this dissertation, as its primary objective is to introduce SEM forests as a novel tool for conducting specification search in SEM. A more detailed discussion of specification search is provided in the following section.

### ***2.1.3. Specification Search in SEM***

Specification search, also known as model modification, is a process to detect and correct model specification errors, aiming to enhance model fit or achieve greater parsimony by freeing or imposing model constraints (MacCallum, 1986; MacCallum et al., 1992;

Marcoulides & Falk, 2018). Hancock and Thompson (2010) propose to distinguish between the use of specification search with confirmatory purposes, requiring stricter psychometric standards, and its application for hypothesis generation, which may accept more flexible standards to foster the creation of new models for future validation. Specification search methods can be broadly categorized into sequential and non-sequential approaches.

Sequential methods modify one model constraint at a time, whereas non-sequential methods explore combinations of parameters to maximize model fit (Green & Thompson, 2010).

Specification search methods may employ various statistics, such as comparing different SEM models using the chi-square difference test or applying information criterion indices (i.e., AIC, BIC), among other indices, to find a model that minimize specification errors. Specification errors occur when the proposed model fails to adequately represent the true model, defined as the correct network of relationships among the variables under investigation (MacCallum, 1986). Specification search is inherently an exploratory, data-driven process that requires cross-validation to mitigate its intrinsic risk of capitalization on chance (MacCallum et al., 1992; Marcoulides & Falk, 2018). Capitalization on chance refers to drawing conclusions based on random sample variations rather than true underlying effects, leading to results that lack generalizability (MacCallum, 1986; MacCallum et al., 1992).

The most commonly employed strategy for conducting a specification search in SEM is Modification Indices (MIs). MIs is a special case of the Lagrange multiplier test, that estimates the minimum reduction in the overall chi-square value that would result from freeing a single constraint at a time. (Green & Thompson, 2010; MacCallum, 1986).

Parameters with the largest MIs are those that, when freed, contribute most significantly to improving the chi-square goodness-of-fit index. Then, fixed parameters with the highest MIs are sequentially added to the model to improve its fit. However, since MIs are chi-square-



based, they share the biases associated with the chi-square statistic when measuring model fit, such as sensitivity to sample size and parameter magnitudes. In this sense, MacCallum et al. (1992) report that MI cross-validation results were unstable with small and medium samples, and that MIs values were highly unstable for small samples across repeated sampling. Furthermore, MIs are impractical for complex models with numerous potential modifications, as they do not comprehensively explore the entire model parameter space, and exploring all theoretically relevant modifications would require estimating a substantial number of refitted models (Marcoulides & Falk, 2018).

Given the limitations of MIs, multiple alternative techniques for specification search in SEM have been proposed. Many of them are based on Wald tests, where all potential parameters are initially included in the model, and then each parameter is sequentially removed to minimize the decrease in model fit (Green & Thompson, 2010; Marcoulides & Falk, 2018). Wald-based tests are also part of the chi-square family tests. Marcoulides et al. (1998) introduced a technique based on the Tabu search approach, a nonsequential method that sets a model selection criterion (e.g., chi-square statistic, AIC, BIC) and iteratively explores neighboring models, which are models that differ by only one free term, by making small modifications, such as freeing or fixing a parameter. Marcoulides and Falk (2018) developed an R-based environment code for conducting specification search utilizing the Tabu search procedure with the BIC as the optimization criterion. The authors highlighted that this code is also compatible with alternative search methods, including genetic algorithms and ant colony optimization, among others.

The specification search techniques outlined are certainly not comprehensive, since the aim of this overview was to illustrate alternatives techniques to the traditional MIs, rather than provide a comprehensive list of available methods. However, to our knowledge, no specification search technique has yet specifically addressed the issue of identifying

misspecifications related to omitted covariates. The primary objective of this dissertation is to propose the use of SEM forests as an innovative technique for conducting specification search in SEM, particularly associated with the identification of omitted influential covariates. The following section introduces the principles of SEM forests and elaborates on the rationale supporting their potential application as a specification search tool.

## 2.2. SEM Trees and Forests

SEM forests, proposed by Brandmaier et al. (2016), are ensembles of decision trees applied in the SEM context, known as SEM trees (Brandmaier et al., 2013). A decision tree is a non-parametric decision-making analysis that recursively splits a data set by maximizing an information criterion or by applying statistical tests to evaluate the significance of the splits. SEM trees, originally introduced by Brandmaier et al. (2013) and further developed in the R package *semtree* (Brandmaier et al., 2023), identify subgroups within a sample that exhibit similar response patterns based on a set of covariate predictors and their interactions. This approach maximizes differences across subgroups and similarities within subgroups (Brandmaier et al., 2016). Moreover, SEM trees enable the detection of covariate interactions that predict linear and non-linear relationships among structural parameters, allowing for the identification of covariate-specific subgroups that account for unexplained heterogeneity.

Given, for example, an observed outcome  $y$  and a vector of covariates  $x$ , a decision tree recursively partitions the covariates space in  $x$  based on significant differences in  $y$ . Postsplit models are obtained by freeing all parameters of the original SEM, however, researchers can also restrict the statistical criterion to test only specific parameters of interest (Brandmaier et al., 2013). The *semtree* package allows different methods for tree growing, including naïve, fair and score-guided approaches. Arnold et al. (2020) developed score-guided SEM trees. These trees evaluate the heterogeneity of model parameters through scores indicating how well the model parameters represent individual data points. Higher scores

reflect a larger misfit between a given model parameter and individual data points. The scores are sorted with respect to the covariates under scrutiny and aggregated into a test statistic that evaluates the null hypothesis of homogeneous parameters.

SEM trees are particularly useful when dealing with a large number of covariates and unknown interactions. The focus of decision tree techniques on interaction effects is clear when we examine one of their precursors, automatic interaction detection (AID) (Morgan & Sonquist, 1963) AID was developed to identify the predictors that have the greatest impact on reducing predictive error by growing branches of binary exclusive groups, with the selection criterion being the largest unexplained sum of squares. The resulting branches reveal which variables are most important and their interactions. Moreover, although tree-based methods were developed for dichotomous variables, SEM trees can also be applied to categorical, ordinal or continuous covariates. This is achieved through a dichotomization process, which can use methods such as the one-against-the-rest scheme for categorical data or an exhaustive split search for ordinal or continuous data (Brandmaier et al., 2013).

The criteria for deciding when to stop new splits in decision trees is a central aspect of their generation. Various evaluation functions, such as information gain and GINI impurity, have been proposed to assess the need for further splits (Brandmaier et al., 2013, 2016). In contrast, SEM trees initially employed the likelihood ratio as the criterion to evaluate whether a split was necessary (Brandmaier et al., 2013), However, this approach was later replaced by more efficient score-based tests (Arnold et al., 2020). Postsplit models are obtained by allowing all parameters of the original SEM to vary. However, it is also possible to restrict the score-based test to examine only differences in selected parameters that are of particular interest to the researcher. The search across multiple predictors generates a multiple testing problem, requiring an adjustment to the  $p$ -value to account for this issue (e.g., Bonferroni correction) (Brandmaier et al., 2013, 2016).

Despite their advantages, SEM tree outputs can be unstable, raising concerns about their generalizability. As Brandmaier et al. (2016) note: “In each inner node of a tree, both the associated predictor variable and its split point are chosen to be locally optimal and, thus, can easily be influenced by small perturbations of the sample at hand. A slightly different choice of a split point may lead to a different choice of the subsequent split in the children of a node; in this way, small perturbations are typically magnified down the tree” (p. 568).

### ***2.2.1. Ensemble SEM Trees: The SEM Forests Technique***

Ensemble methods enhance the robustness and accuracy of individual models. SEM forests (Brandmeier et al., 2016) are ensembles of SEM trees that allow practitioners to identify relevant covariates from a set of potential predictors, among other applications detailed in Study 1. SEM forest generates a set of trees, each one based on a random sample of the original data set. The combination scheme of the random forests aggregates predictions of individual trees, for example, predicting a continuous outcome as the average of individual tree predictions, or a dichotomous outcome as the majority category over the individual tree predictions. Ordinal, continuous, and categorical variables are dichotomized using an exhaustive split search. Additionally, SEM forests randomly select a subset of predictors at each node, reducing the number of tests and potential splits compared to SEM trees, where all potential predictors are tested at each level. With a large number of predictors, SEM forest analysis can run faster than a single SEM tree analysis (Brandmeier et al., 2016).

The SEM forests generation process is described by Brandmaier et al. (2016). Initially, a set of models is fitted to the data and then split with respect to each covariate into submodels. The submodel with the best increases in fitness is compared against a predetermined threshold. If the split is statistically significant, the process continues recursively, generating a set of decision trees. A resampled training data set,  $D_i^{train}$ , is then generated for each tree using bootstrapping or subsampling, where  $i = 1, \dots, t$ , with  $t$

representing the total number of trees in a forest. The remaining cases are assembled into an out-of-bag sample,  $D_i^{OOB}$ , for each  $i$ . Based on each  $D_i^{train}$ , an SEM tree is grown. At each node of the trees comprising the random forest, a subset of candidate covariates is tested. The size of the set of candidate covariates at each node,  $c$ , can be determined using  $c = \log_2(m)$ ,  $c = \sqrt{m}$ , or  $c = m/3$ , where  $m$  is the total number of potential covariates (Brandmaier et al., 2016).

SEM forests include two types of aggregate measures that allow the identification of sources of variability in the model: variable importance and case proximity. Variable importance is a nonparametric estimator of the relative importance of a set of covariates and their interactions, based on the information they provide about the model-predicted distribution. This measure, based on permutation accuracy importance, evaluates the decrease in fitness due to the random permutation of a predictor. Specifically, variable importance quantifies the impact that a potentially relevant covariate has on the covariance structure of a model (Brandmaier et al., 2016).

Operationalizing, variable importance in SEM forests involves calculating the average decrease in log-likelihood importance across all trees generated from  $D_i^{OOB}$ . Specifically, variable importance is the average of the difference between the log-likelihood of the *OOB* samples for each tree and the log-likelihood of the scrambled *OOB* samples for each tree. Scrambled *OOB* samples are obtained by randomly permuting the column values corresponding to the subset of candidate covariates. To compute this, the log-likelihood for each tree of  $D_i^{OOB}$  is calculated as  $LL(D_i^{OOB} | T_i)$ . Then, the potential covariates in  $D_i^{OOB}$  are randomly permuted, obtaining a scrambled sample  $\tilde{D}_i^{OOB}$ , and its log-likelihood  $LL(\tilde{D}_i^{OOB} | T_i)$  is calculated. Finally, the following likelihood average is calculated:

$$\frac{1}{t} \sum_{i=1}^t LL(D_i^{OOB} | T_i) - LL(\tilde{D}_i^{OOB} | T_i),$$

repeating this procedure for each potentially influential covariate (Brandmaier et al., 2016).

The second aggregate measure of SEM forests, case proximity, assesses the internal structure of the data. Case proximity allows the identification of hidden patterns by performing case-based clustering, which measures similarities in the covariate space. Case-based clustering is a clustering technique based on proximity matrices and multidimensional scaling that evaluates the internal structure of the data (Brandmaier et al., 2016). Case proximity allows to detect outliers, identify prototypes, uncover heterogeneity data heterogeneity, and recover information lost in the aggregation of individual tree structures within the forest. The studies presented in this dissertation centers on SEM forest variable importance and does not incorporate SEM forest case proximity.

We propose the use of SEM forest variable importance to conduct specification search related to omitted covariate paths, as SEM forests is a robust technique capable of detecting homogeneous subgroups within datasets, such as potentially influential predictors. Additionally, SEM trees and forests reduce the risk of overfitting, finding potential generalizable features in the data (Arnold et al., 2020; Brandmaier et al., 2013). The first study presented below uses data simulation to examine the performance of SEM forests in identifying omitted influential covariates, while the second study provides a more pedagogical guide aimed at a broader audience interested in using SEM forests with the `semTree` package (Brandmaier et al., 2023) for specification search related to omitted covariates.

### **3. Study 1: Evaluation of Structural Equation Model Forests' Performance to Identify Omitted Influential Covariates**

#### **3.1. Abstract**

Model misspecification is typical in applied structural equation modeling (SEM). Traditional specification search methods, such as modification indices, search for misspecifications within the model's variables but overlook influential variables not initially included and fail to detect interactions. This study evaluates SEM forests as a complementary method to conduct SEM specification search related to omitted influential covariates. The omitted influential paths include unique, mixed, and interaction paths. SEM forests' performance is evaluated under different factor loading magnitudes, covariate path magnitudes, and sample sizes. Results show SEM forests accurately identify omitted influential covariates without falsely identifying non-influential covariates in large samples (1,000) with strong covariate-latent variable paths ( $\beta = .5$ ).

*Keywords:* Model fit; model specification search; SEM forests; structural equation modeling; variable importance.

#### **3.2. Introduction**

Structural equation modeling (SEM) is a widely used statistical tool in social sciences for establishing complex multivariate relations between latent variables while controlling for measurement error (Holbert & Stephenson, 2002; Mulaik, 2009). SEM is typically used as a confirmatory method where researchers, with a theory-driven approach, specify and test structural and measurement relations among variables. However, multivariate models in the social sciences often fail to account for all relevant variables and paths, leading to misspecified models, biased parameter estimates and incorrect interpretations (e.g., Kaplan, 1988; Mulaik, 2009; Tomarken & Waller, 2003).

Researchers often use a modification index-based approach for model specification search. This method estimates the increase in chi-square-based model fit when they sequentially add identified omitted relevant paths (e.g., regression weights) (MacCallum et al., 1992; Mulaik, 2009; Saris et al., 2009). However, modification indices inherit biases inherent to the chi-square test, such as the undesired influences of nuisance parameters like sample size and parameter magnitudes (Saris et al., 2009). They also assume the model is correctly specified in terms of included variables, failing to detect omitted variable bias. The latter occurs when omitted variables correlate with both included predictors and dependent variables, leading to biased coefficients of the included predictors and incorrect model fit indices (Kline, 2016; Tomarken & Waller, 2003; Wilms et al., 2021). Moreover, modification indices are not sensitive to misspecifications involving interaction effects (Brown & Templin, 2023; Mooijart & Satorra, 2009) and their heuristic use can lead to overfitting (MacCallum et al., 1992). This is especially problematic in social science, where interactions reflect the contextualized effect of variables regulated by others (Morgan & Sonquist, 1963; Tomarken & Waller, 2003). Despite these limitations, latent models that account for measurement error are powerful for handling interaction effects, as they manage the increased error resulting from multiplying individual measures (Cortina et al., 2021).

How can we move beyond such limitations of modification indices in SEM for detecting model misspecification? Various alternatives for specification search in SEM have been proposed, such as Marcoulides and Drezner (2001) genetic algorithm search, Scheines et al. (1998) vanishing tetrads, Saris et al. (1987) expected parameter change (EPC), Marcoulides and Drezner (2003) ant colony optimization (ACO) algorithm, or Marcoulides et al. (1998) tabu search variable selection. Brandmaier et al. (2013, 2016) proposed SEM trees and SEM forests, the fusion of exploratory decision trees and random forests with SEM. These techniques enable a data-driven and yet theory-guided exploration of models,



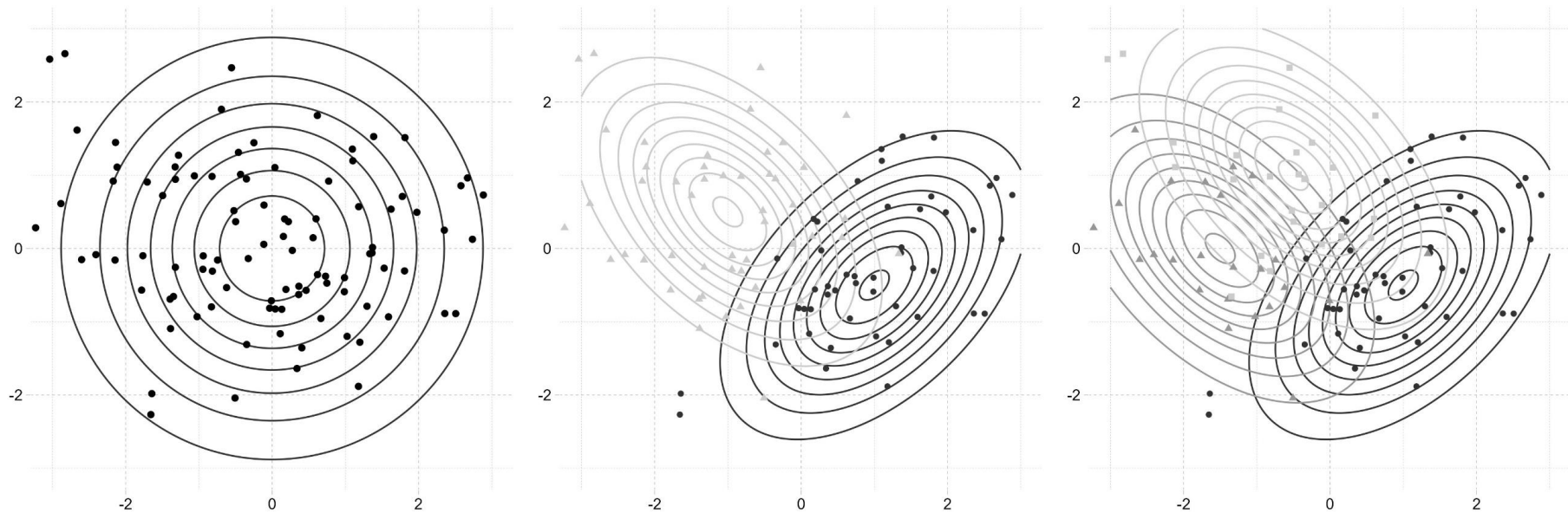
extending the confirmatory approach of SEM with the exploratory approach of decision trees. SEM forests allow the identification of possible covariates and covariate interactions that predict parameter heterogeneity in SEM, potentially informing about variables that provide additional predictive information. Moreover, tree and forest structures reduce the chance of overfitting, finding potential generalizable features in the data (Arnold et al., 2020; Brandmaier et al. 2013).

### **3.3. Decision Trees and SEM Trees**

A decision tree is a nonparametric predictive model that recursively splits a data set to create groups with similar observations within each group and dissimilar observations between groups. Decision trees find locally optimal splits by maximizing an information criterion or applying statistical tests to determine the significance of the splits. They can be described as a data-driven, yet theory-constrained search in model space, starting with a hypothesized model and expanding it if the trees identify further relevant variables (Breiman et al., 1984). SEM trees, proposed by Brandmaier et al. (2013) and implemented in the R package *semtree* (Brandmaier et al., 2023), combine the nonparametric nature of decision trees with parametric SEMs as outcomes. They assume that observed data are drawn from different underlying multivariate normal distributions (Figure 1). SEM trees account for heterogeneity not explained by an initial theory-based model by hierarchically splitting a data set with respect to a potential set of covariate predictors and their interactions, maximizing differences in parameter estimates across the resulting groups while finding similar response patterns within groups (Brandmaier et al., 2016). Each node of the tree represents simultaneously a data partition with respect to a covariate and a SEM that induces the resulting subgroups. The partitions remain as long as relevant differences in the parameter estimates are discovered.

## Figure 1 – Study 1

*Illustration of SEM tree splitting process*



*Note.* A SEM represents a multivariate normal fit (illustrated as contours) to some data (illustrated as points). Left: A single Gaussian model describes the entire set. Center: Potential split into two subgroups for better fit. Right: Further split in left subgroup for improved model fit.

Figure adapted from the Handbook of Structural Equation Modeling (2nd ed.), by A. M. Brandmaier and R. C. Jacobucci, 2023, Guilford Press.

Copyright 2023 by Guilford Press.

The R package *semtree* (Brandmaier et al., 2023) includes different criteria for variable and split-point selection, including the recently implemented score-guided test (Arnold et al., 2020). Score-guided tests use log-likelihood function derivatives giving scores sorted by the covariates under scrutiny and aggregated into a test statistic to evaluate SEM parameters homogeneity across all levels of a given covariate (Arnold et al., 2020). Larger deviations from the expected score (i.e., zero) indicate greater parameter instability. Arnold et al. (2020) reported that score-based tests are unbiased in covariates selection (i.e., there is no preference for any covariate when the null hypothesis is true) and have higher statistical power than original SEM tree methods (i.e., naïve and fair). Score-based tests are also computationally more efficient since only one model needs to be estimated, in contrast to the original likelihood ratio approach implemented in SEM trees, which needs to estimate models for each split candidate (Arnold et al., 2020).

SEM trees are especially suited for data with many covariates and unknown interactions (Brandmaier et al., 2013). However, SEM tree results may be unstable because each split may be affected by particularities of the sample at hand, and the effect of a small perturbation in a node would be magnified in a cascade effect down the tree, calling their generalizability into question (Brandmaier et al., 2016). As a more robust alternative to SEM trees, Brandmaier et al. (2016) proposed the use of ensembled SEM trees, called SEM forests.

### **3.4. Ensemble Methods and SEM Forests**

An ensemble is a set of models generated from random samples of an original data set that is often more robust and accurate than individual models. Ensemble methods define a sampling scheme to generate random data and a combination scheme to aggregate individual model predictions into a final outcome. Basu et al. (2018) mentioned a series of decision trees ensembles to identify interaction effects, such as Random Forests (Breiman, 2001), Iterative Random Forests (Basu et al., 2018), and Node Harvest (Meinshausen, 2010). Basu et al.

(2018) highlighted that random forests are the best option for detecting high-order interactions since other decision trees ensemble techniques grow shallow trees to prevent overfitting, but at the cost of predictive accuracy when detecting high-order interactions.

Random forests are an ensemble method for tree-based methods like SEM trees. SEM forests are ensembled SEM trees that can guide the search for potential explanatory variables, assess their influence, determine which variables to control for, and generate new hypotheses about structural relations that are difficult to generate from a pure theory-driven approach in complex data sets. A key aim of SEM forests is to determine the importance of covariates in predicting an outcome. If a covariate is important in predicting an outcome variable, it will be more consistently selected by the individual trees that included it in the random subset of covariates considered for splitting, indicating its stability as a predictor and helping to identify potential influential covariates.

To generate a SEM forest, resampled training data sets are created for each tree using either bootstrapping aggregating (bagging) or subsampling from the original data. Bagging generates new random samples with replacement, matching each new sample with the original sample size, while subsampling draws smaller new random samples without replacement. A SEM tree is grown from these resampled datasets keeping the observations that were not part of the training as out-of-bag samples for each tree. To enhance tree diversity, only a random subset of the total covariates is evaluated at each split, with subset size  $c$  defined heuristically (i.e.,  $c = \log_2(m)$ , being  $m$  the total amount of covariates). Smaller  $c$  increases tree variability and independence but may miss relevant predictors and high-order interactions (Brandmaier et al., 2016).

Oshiro et al. (2012) explored the ideal number of trees in a random forest, based on the area under the ROC curve, which illustrates power against type-I error rate. They suggest using between 64 and 128 trees per forest, as adding more trees beyond this range only

escalates computational cost without notable performance improvement. However, in the context of SEM forests, Brandmaier et al. (2016) noted that the optimal number of trees depends on various factors, including the number of predictors, their interactions, data heterogeneity, and model complexity. Although specific thresholds for the number of trees in SEM forests are yet to be established, Brandmaier et al. (2016) heuristically proposed setting the number of trees to a relatively large value, such as 2,000 trees.

### ***3.4.1. SEM Forests Variable Importance Output***

In SEM forests, a permutation-based variable importance is provided as a nonparametric estimator of the relative importance of a set of covariates and their interactions. Variable importance evaluates the decrease in fit due to the random permutation of a predictor, assuming that if an important predictor is randomly permuted its functional relation with the model-predicted distribution is broken (Brandmaier et al., 2016). Variable importance gives aggregate information about unmodeled variables that a single SEM tree cannot provide, quantifying the influence of potentially relevant covariates on the model's covariance structure (Brandmaier et al., 2016).

SEM forests variable importance is calculated as the average decrease in log-likelihood importance when a predictor is removed from the forest. This is estimated based on a simple resampling scheme, computing the likelihood of observing data for each tree in the forest, then randomly permuting the predictor of interest, removing all outcome information, and recalculating the likelihood for each tree. This process is repeated for each predictor (Brandmaier et al., 2016). The larger the drop in likelihood after the permutation, the more important the variable. Decisions regarding sampling parameters, such as resampling type, number of trees, and candidate predictors at each node, can impact variable importance. Additionally, missingness and categorical variable imbalance are other factors

that may affect variable importance (Brandmaier et al., 2016). Then, informed definitions of those parameters are a key step to obtain accurate variable importance outcomes.

The inherent random variation of observed samples that entails the SEM forests generation process enhances the accuracy of variable importance estimates compared to single decision trees. That is, the instability disadvantage of individual trees becomes an advantage of forests, since random fluctuations allow an ensemble to be a better representation of the true partition of a sample. Increased diversity improves the performance of ensemble methods, with random forests outperforming other classification approaches such as generalized linear models and support vector machines (Bühlmann & Yu, 2002; Fernández-Delgado et al., 2014). However, forests improvement in accuracy comes at the expense of losing the straightforward interpretability of individual trees. Moreover, SEM forests do not specify variable relations, leaving the integration of covariates paths into the model open-ended. As with any data-driven technique, SEM forests do not provide a shortcut from data to theories and their outcomes should be tested on independent samples to validate their generalizability (Brandmaier et al., 2016).

### **3.5. Questions and Hypotheses**

SEM forests have been successfully applied to explore heterogeneity in large-scale empirical data sets, allowing to identify influential predictors in diverse contexts: students' attitudes toward collaboration in PISA 2015 (Li et al., 2021), individual differences in episodic memory (Brandmaier et al., 2016), late-life well-being decline (Brandmaier et al., 2017), and early predictors of adolescent emotion regulation (Van Lissa et al., 2023). However, to our knowledge, there is a lack of studies that explored to what extent SEM forests' performance to identify relevant omitted predictors is influenced by either nuisance parameters such as factor loadings magnitude or sample size, or by data parameters related to the degree of misspecification, such as the magnitude of omitted covariate paths or

interaction parameters. Usami et al. (2017, 2019) reported that SEM trees' ability to identify true classes explaining population heterogeneity in longitudinal data was influenced by the covariate's agreement with its true latent profile, sample size, and negatively affected by the number of true classes. However, Usami et al. (2017, 2019) examined SEM trees performance, but not SEM forests, and only in the context of longitudinal analysis.

This study addressed the mentioned gaps in SEM forests performance research by examining their ability to consistently and accurately identify influential omitted covariate paths while avoiding false detection of non-influential ones in underspecified models under different data conditions. SEM forests' performance was evaluated under different standardized factor loadings, strength of covariate-latent variable relationships, and varying sample sizes. We hypothesized that larger factor loadings and stronger covariate effects would lead to more consistent detection of influential omitted covariate paths due to increased statistical power (Heene et al., 2011; Meyvis & Van Osselaer, 2018; Van Voorhis & Morgan, 2007). We also expected SEM forests to perform better with larger sample sizes, as they reduce sampling errors (Van Voorhis & Morgan, 2007), stabilize variable and split-point selection in individual trees, and trees in a forest can grow deeper. Additionally, we explored different covariate-latent variable relationships: covariates with no influential paths, mixed influential paths, unique influential paths, and a covariate interaction variable. We expected a particular high performance in detecting the omitted interaction path, as SEM trees and forests are especially suitable to detect interactions (Brandmaier et al., 2013; 2016).

### **3.6. Procedure**

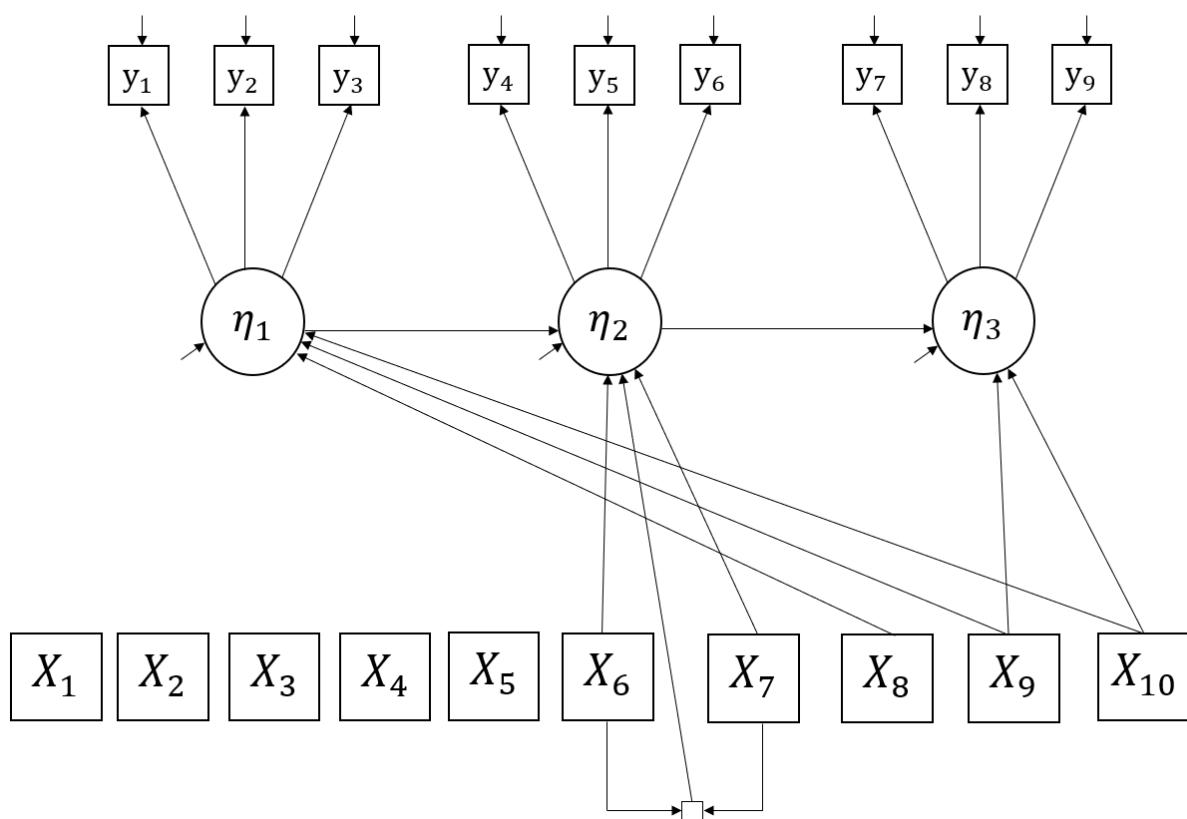
#### ***3.6.1. Population Structural Equation Model***

We specified a population latent model inspired by Paxton et al. (2001), who deduced their model structure from a literature review of SEM studies published in psychological and sociological key journals in a period of five years. The population model included nine

measured variables (indicators), three latent variables, and 10 exogenous variables (covariates) (Figure 2). Following Paxton et al. (2001), we defined a structural causality chain with mediator  $\eta_2$  regressed on predictor  $\eta_1$ , and outcome variable  $\eta_3$  regressed on mediator  $\eta_2$ . Additionally, each of the nine indicators loaded uniquely on one latent variable.

**Figure 2 – Study 1**

*Path Diagram of the Population Model*



*Note.* The population model has nine indicators  $y$  and three latent variables  $\eta_1$  to  $\eta_3$ , with 10 exogenous variables  $X_1$  to  $X_{10}$ . The box between  $X_6$  and  $X_7$  represents an interaction effect on  $\eta_2$ .

Our population model included 10 covariates, as SEM forests is a technique developed for larger sets of covariates. We regressed predictor  $\eta_1$  on covariates  $X_8, X_9$  and



$X_{10}$ , mediator  $\eta_2$  on covariates  $X_6$  and  $X_7$ , and outcome variable  $\eta_3$  on covariates  $X_9$  and  $X_{10}$ . The mediator  $\eta_2$  was also regressed on the interaction between  $X_6$  and  $X_7$ . With this set, the influential covariates  $X_6$ ,  $X_7$  and  $X_8$  had unique paths (each covariate relates to only one latent variable), while the influential covariates  $X_9$  and  $X_{10}$  had mixed paths (each covariate relates to two latent variables). Covariates  $X_1$  through  $X_5$  were defined as non-influential predictors with no paths to any latent variable. All paths between the latent variables  $\eta_1$ ,  $\eta_2$  and  $\eta_3$ , indicators  $y_1$  to  $y_9$ , and latent variables  $X_1$  to  $X_{10}$  were defined as linear relationships.

### 3.6.2. *Experimental Design*

We used a  $3 \times 4 \times 3$  fully crossed design, totaling 36 conditions, with factors including measurement quality, magnitude of covariate paths, and sample size. For measurement quality we defined three levels representing low, medium, and high standardized factor loadings:  $\lambda_1 = .4$ ,  $\lambda_2 = .6$ , and  $\lambda_3 = .8$ . For covariate path magnitudes, understood as the magnitude of standardized regression coefficients between covariates and latent variables, we defined four levels: null, low, medium, and high paths:  $\beta_0 = 0$ ,  $\beta_1 = .2$ ,  $\beta_2 = .35$ , and  $\beta_3 = .5$ . The interaction effect was fixed to .2 for all the conditions. For sample size, we used three levels corresponding to small, medium and large samples in the context of social sciences:  $N_1 = 200$ ,  $N_2 = 500$  and  $N_3 = 1,000$ .

### 3.6.3. *Data Generation*

We used the R package *simsem* (Pornprasertmanit et al., 2022) to generate sample data for each combination of our experimental conditions. For all the conditions, regression path coefficients between latent variables were set to .4, and measurement error correlations between indicators and the residual correlation among latent variables were both fixed to zero. To keep parameters in a standardized scale, indicator and exogenous latent variable variances were set to one. Noninfluential covariates  $X_1$  to  $X_5$  were randomly generated as continuous variables from a standard normal distribution. Influential covariates  $X_6$  to  $X_{10}$

were continuous variables generated from a multivariate normal distribution with mean vector  $\mu = 0$ , and off-diagonal covariance matrix elements uniformly distributed within the range  $[0, .1]$ . The interaction variable  $X_6X_7$  was computed as the product of  $X_6$  and  $X_7$ . For each of the 36 possible combinations of the experimental conditions, 500 data sets were generated. Out of a total of 18,000 cases, 98.6% (17,751 cases) were successfully generated, while 1.4% (249 cases) failed during data generation. All the cases not correctly generated correspond to the largest covariate path condition ( $\beta_3 = .5$ ), where 94.5% (4,251 of 4,500) were correctly generated, while 5.5% (249 of 4,500) were not generated.

### **3.6.4. SEM Forests Growth**

For each of the 17,751 cases generated, we generated one SEM forest composed of 500 trees using the `semforest` function of the R package `semtree` (Brandmaier et al., 2023). We employed score-guided SEM trees for split selection, since this method is faster and more powerful than other split methods available in the `semtree` package (i.e., fair, naïve; see Arnold et al., 2021). Subsampling was used as the resampling procedure, with three candidate covariates considered at each node, based on the heuristic of randomly draw  $\log_2(m)$  variables for every split evaluation (Brandmaier et al., 2016) with 10 being the total number of covariates ( $m$ ) for our data sets.

To promote robust forest growth and stable estimates, the minimum sample size per node was set to 100, ensuring no split attempts were made if a node had fewer than 100 observations. Additionally, the lower bound of the terminal nodes was set to 50, requiring at least 50 observations in a child node of a potential split to be valid. This decision came at the cost of having SEM forests with only two possible splits for our smallest sample size conditions ( $N_I = 200$ ). In random forests, it is often useful to let individual trees grow deep without applying a statistical cut-off. Thus, we kept the default package alpha level of 100% (i.e., never stop splitting based on the chi-square test statistic) as the most liberal statistical

criterion to get deep trees. This increased power for detecting interactions but at the cost of Type I error, meaning a possible increase of the incorrect detection of the non-influential covariates in any given tree of a forest. However, even if individual trees overfit, the variable importance resampling approach averages across many trees helping to prevent individual trees from overfitting the data.

To evaluate SEM forests' performance to correctly identify omitted influence covariate paths and not falsely identify no-influence covariate paths, we used the variable importance analysis integrated in the R package *semtree* (Brandmaier et al., 2023). Higher variable importance indicates greater model misfit when a covariate is omitted, meaning that including those variables improves model fit.

### 3.7. Results

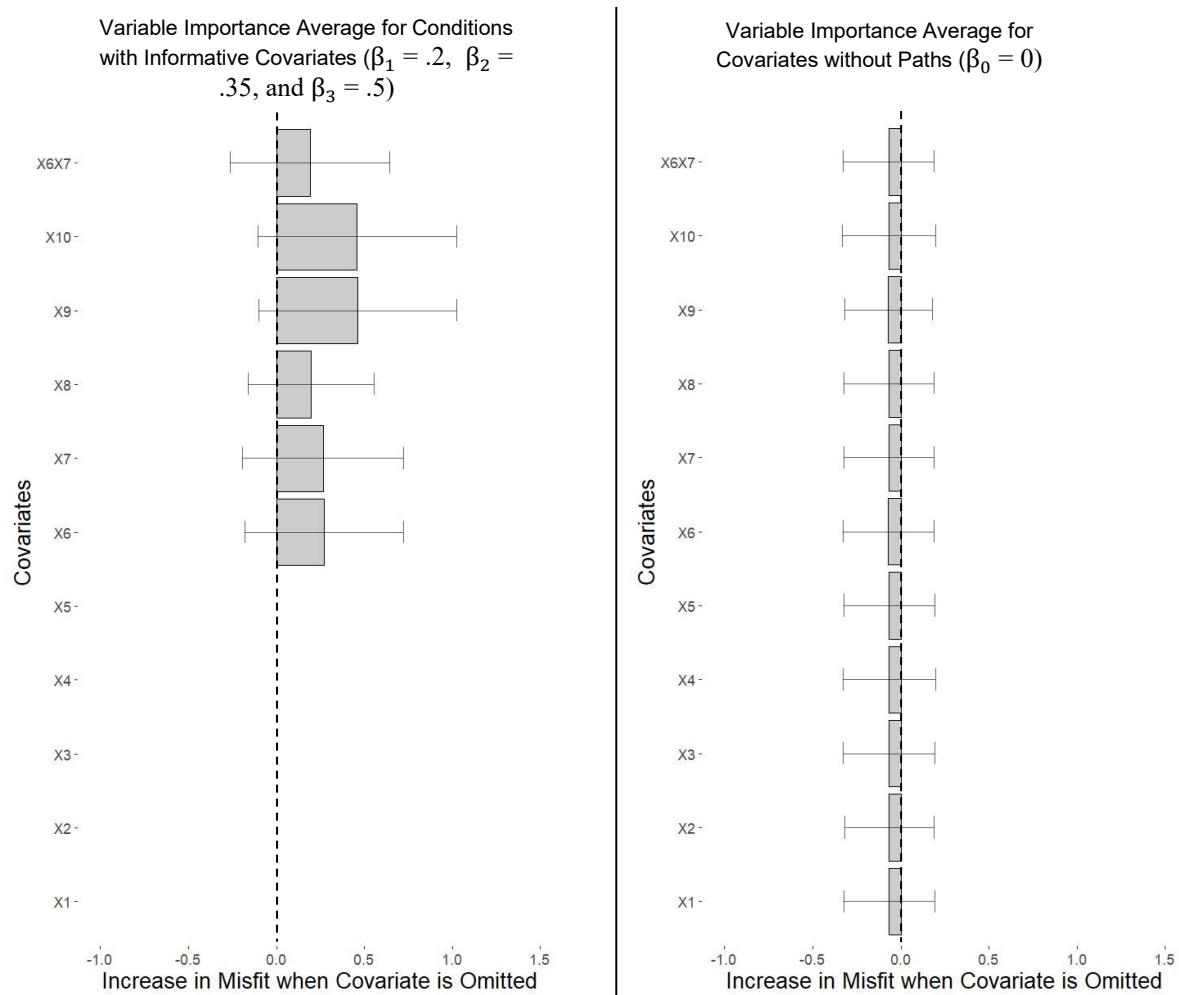
Figure 3 shows that influential covariates ( $X_6$  to  $X_{10}$  and the interaction variable  $X_6X_7$ ) had higher importance values than non-influential ones ( $X_1$  to  $X_5$ , and condition  $\beta_0 = 0$ ). Covariates with mixed paths ( $X_9$  and  $X_{10}$ ) also had higher importances than those with unique paths ( $X_6$ ,  $X_7$  and  $X_8$ ) and the interaction variable  $X_6X_7$ , which had similar importance scores. Non-influential covariates had variable importances slightly negative but still centered around zero, indicating unbiased estimates.

There was a notable dispersion in the importance values for both influential and non-influential covariates, indicating that SEM forests may sometimes misidentify covariates. To calculate SEM forests' probabilities of correctly selecting influential covariates and falsely selecting non-influential ones, we used the absolute value of the largest negative importance of each covariate as a threshold (Strobl et al., 2009). Importances above this threshold were classified as influential, while those below as noninfluential. The probability of correctly identifying influential covariates increased with larger covariate path magnitudes, factor loading magnitudes or sample sizes (Table 1). The probability of falsely identifying non-

influential covariates was close to zero across all covariate path magnitudes, factor loading magnitudes, and sample sizes.

### Figure 3 – Study 1

#### *Average Variable Importances: Influential vs. Non-Influential Covariates*



*Note.* On the left,  $X_6X_7$  is an interaction variable;  $X_6$ ,  $X_7$ , and  $X_8$  have unique influence paths (paths to only one latent variable);  $X_9$  and  $X_{10}$  have mixed influence paths (paths to two latent variables); and  $X_1$  to  $X_5$  are non-influential covariates. On the right, all covariates are non-influential.

**Table 1 – Study 1**

*SEM Forests' Probabilities of Correctly Selecting Influential Covariates and Falsely Selecting Non-Influential Covariates*

Probability of Correctly Selecting an Influential Covariate																		
	$\beta_1 = .2$									$\beta_2 = .35$								
	$N_1$			$N_2$			$N_3$			$N_1$			$N_2$			$N_3$		
	$F_1$	$F_2$	$F_3$	$F_1$	$F_2$	$F_3$	$F_1$	$F_2$	$F_3$	$F_1$	$F_2$	$F_3$	$F_1$	$F_2$	$F_3$	$F_1$	$F_2$	$F_3$
$X_6$	0	0	0	.02	.01	.23	.07	.44	.51	.01	.03	.04	.38	.9	.97	.73	1	1
$X_7$	0	0	0	.02	.06	.24	.27	.54	.74	0	.01	.02	.13	.88	.88	.93	1	1
$X_8$	0	0	0	.04	.07	.11	.13	.34	.29	0	0	.05	.09	.44	.79	.81	1	1
$X_9$	0	0	0	.06	.21	.46	.21	.77	.97	0	.01	.28	.62	1	1	1	1	1
$X_{10}$	.01	0	0	.03	.29	.13	.32	.73	.98	0	.31	.08	.82	.98	1	1	1	1
$X_6X_7$	0	0	0	.01	.07	.1	.1	.22	.29	0	.01	.02	.1	.27	.63	.26	.64	.99
	$\beta_3 = .5$																	
	$N_1$			$N_2$			$N_3$											
	$F_1$	$F_2$	$F_3$	$F_1$	$F_2$	$F_3$	$F_1$	$F_2$	$F_3$									
$X_6$	0	.41	.86	.72	1	1	1	1	1									
$X_7$	.04	.38	.96	.93	1	1	1	1	1									
$X_8$	.01	.07	.37	.94	1	1	1	1	1									
$X_9$	.3	.59	.94	1	1	1	1	1	1									
$X_{10}$	.3	.31	.97	1	1	1	1	1	1									
$X_6X_7$	.02	.12	.85	.25	.99	1	.87	1	1									

---

Probability of Falsely Selecting a Non-Influential Covariate

---

$\beta = 0$									
	$N_1$			$N_2$			$N_3$		
	$F_1$	$F_2$	$F_3$	$F_1$	$F_2$	$F_3$	$F_1$	$F_2$	$F_3$
$X_1$	0	0	0	0	.01	.01	.02	.01	.01
$X_2$	0	0	0	0	0	.01	.01	.01	.01
$X_3$	0	0	0	0	0	.01	.01	.01	0
$X_4$	0	0	0	0	0	0	.01	.01	.01
$X_5$	0	0	0	0	.01	.01	.01	.01	0
$X_6$	0	0	0	0	0	0	.01	.02	.02
$X_7$	0	0	0	0	.02	.01	.01	.02	.01
$X_8$	0	0	0	.01	.02	.01	.02	.01	.01
$X_9$	0	0	0	.01	.01	.01	.02	.02	.03
$X_{10}$	0	0	0	0	.01	.01	.04	.02	.01
$X_6X_7$	0	0	0	0	0	.01	.02	.02	.02

---

*Note.* Sample size levels:  $N_1 = 200$ ,  $N_2 = 500$  and  $N_3 = 1,000$ ; factor loading levels:  $F_1 = .4$ ,  $F_2 = .6$ , and  $F_3 = .8$ . For the probability of falsely selecting a non-influential covariate, covariates  $X_1$  to  $X_5$  are averaged across all experimental conditions since they were defined as non-influential, while covariates  $X_6$  to  $X_{10}$  and interaction variable  $X_6X_7$  correspond to the condition with no covariate effect ( $\beta_0 = 0$ ).

Specifically, probabilities to identify influential covariates were nearly 100% for covariates with medium or high path magnitudes ( $\beta_2 = .35$ ,  $\beta_3 = .5$ ) combined with medium or large samples ( $N_2 = 500$ ,  $N_3 = 1,000$ ), and medium or high factor loadings ( $\lambda_2 = .6$ ,  $\lambda_3 = .8$ ), with a few exceptions (Table 1). Under other conditions, the probabilities of identifying influential covariates were close to zero or varied widely for each covariate. Influential covariates with mixed paths ( $X_9$  and  $X_{10}$ ) typically had higher identification probabilities than those with unique paths ( $X_6$ ,  $X_7$  and  $X_8$ ) and the interaction variable ( $X_6X_7$ ). Although the interaction variable  $X_6X_7$  had its path fixed at .2 for every condition, its identification probability increased when other influential covariates had higher path magnitudes, with higher factor loading magnitudes, and with larger sample sizes (Table 1).

Table 2 shows how factor loading magnitude, covariate path magnitude, and sample size influenced the SEM forest variable importance averages. Larger covariate path and factor loading magnitudes led to higher variable importance scores. For sample size, small and medium samples ( $N_1 = 200$ ,  $N_2 = 500$ ) had similar importance scores, while the large sample ( $N_3 = 1,000$ ) had smaller importance scores. For non-influential covariates, variable importance averages were negative but close to zero across all conditions, except for the smallest sample size, which had a more negative variable importance average (-.182).

Figure 4 shows the influential and non-influential covariate mean differences per experimental condition. The largest mean differences between influential and non-influential covariates were from conditions with the large covariate path ( $\beta_3 = .5$ ), especially when combined with medium and large factor loadings ( $\lambda_2 = .6$  and  $\lambda_3 = .8$ ). Conditions with null ( $\beta_0 = 0$ ) and small covariate paths ( $\beta_1 = .2$ ) had negligible mean differences, regardless of the different factor loadings or sample sizes.

**Table 2 – Study 1***Average Variable Importance per Condition*

	Average importance scores of influential covariates			Average importance scores of non-influential covariates		
	Small	Medium	Large	Small	Medium	Large
Factor	.129	.279	.515	-.059	-.069	-.078
Loadings	(.285)	(.381)	(.609)	(.244)	(.251)	(.282)
Sample Size	.323	.321	.279	-.182	-.021	-.003
	(.687)	(.349)	(.287)	(.414)	(.101)	(0.42)
Covariate Paths	.031	.241	.671	Null ( $\beta_0 = 0$ )		
	(.264)	(.322)	(.541)	-.069 (.26)		

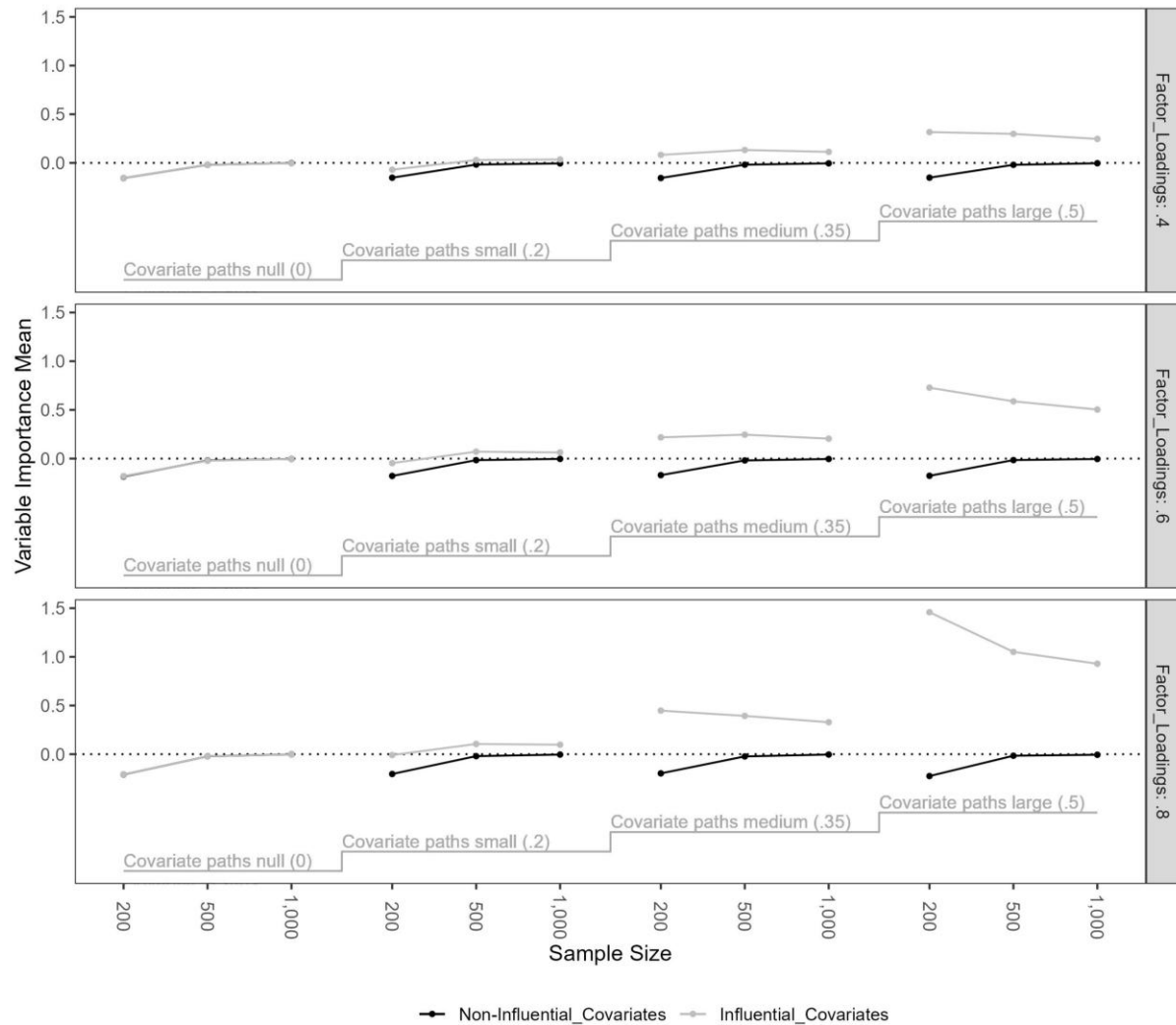
*Note.* Factor loading levels: small = .4, medium = .6, and large = .8; sample size levels: small = 200, medium = 500, and large = 1,000; covariate path levels: null = 0, small = .2, medium = .35, and large = .5. For non-influential covariates, importance scores are averaged from all covariates corresponding to the experimental condition with no covariate effect ( $\beta_0 = 0$ ).



**Figure 4 - Study 1**

*Variable Importance Mean Differences per Condition: Influential vs. Non-Influential*

*Covariates*



*Note.* On the x-axis, sample size levels:  $N_1 = 200$ ,  $N_2 = 500$  and  $N_3 = 1,000$ ; on the y- axis, factor loading levels:  $F_1 = .4$ ,  $F_2 = .6$ , and  $F_3 = .8$ ; and on the steps, covariate path levels:  $\beta_0 = 0$ ,  $\beta_1 = .2$ ,  $\beta_2 = .35$ , and  $\beta_3 = .5$ . Under null covariate paths, non-influential and influential covariates showed nearly identical means, visually indistinguishable on the graph.

Appendices A and B provide a detailed graphical comparison of how each experimental condition affected the SEM forests' performance for influential and non-influential covariates. Influential covariates with mixed paths ( $X_9$  and  $X_{10}$ ) were most frequently identified across all conditions, while the interaction variable  $X_6X_7$  had similar importance scores to those with unique paths ( $X_6$ ,  $X_7$  and  $X_8$ ) (see Appendix A). Among experimental conditions, covariate path magnitude had the greatest impact on influential variable importance scores, with importance scores near zero for small paths ( $\beta_1 = .2$ ), and scores between .5 and 1 for large paths ( $\beta_3 = .5$ ). Higher factor loadings also increased importance scores, though less strongly than covariate paths. Sample sizes showed similar average importance scores, but with reduced score dispersions in larger samples. Factor loading and covariate path magnitudes did not influence variable importance dispersion (see Appendix A). Variable importance scores to identify non-influential covariates were slightly negative and close to zero for all factor loadings and sample sizes, though the small sample had more negative and dispersed scores (see Appendix B).

### 3.8. Discussion

This study examined SEM forests' performance to estimate variable importance scores for influential covariates, covariate interactions, and non-influential covariates, considering data conditions such as factor loadings magnitudes, covariate path magnitudes, and sample size. Using the permutation-based variable importance measure from the *semtree* R package (Brandmaier et al., 2023), results showed that SEM forests are sensitive to omitted influential covariates and provide unbiased importance scores centered around zero for noninfluential covariates across the explored conditions.

SEM forests' probability of correctly selecting influential covariates increases with larger covariate paths, factor loading magnitudes, or sample sizes, while the probability of falsely selecting non-influential covariates is nearly zero across all conditions. Specifically,

the probability of identifying omitted influential covariates was nearly 100% for covariates with high path magnitudes combined medium or large samples, and for medium path magnitudes combined with large samples and medium or large factor loadings. SEM forests also perform better at detecting omitted influential covariates with mixed paths (covariates related to two latent variables) than those with unique paths (covariates related to one latent variable).

Larger covariate paths and factor loadings resulted in higher variable importance scores, while sample sizes showed similar scores but less dispersed for larger samples. SEM forests more frequently identified omitted influential covariates with higher regression paths since larger effect sizes increase statistical power (Meyvis & Van Osselaer, 2018; Van Voorhis & Morgan, 2007). The better and more consistent SEM forests' performance with larger factor loadings is also in line with previous results showing that larger factor loadings have smaller standard errors (Heene et al., 2011). Less dispersed variable importances closer to the true value for larger samples are in line with previous results showing that larger samples decrease sampling error, allowing more stable statistics (Van Voorhis and Morgan, 2007). Thus, the effect of both factor loading magnitude and sample size shows that such nuisance parameters can affect the performance of variable importance measures.

We expected SEM forests to detect the omitted interaction variable, as Brandmaier et al. (2013, 2016) reported that SEM trees and forests are especially suited to detect predictor interactions. While the influential interaction variable was not the most detected omitted path, SEM forests identify its omission even though it was fixed to a small path of .2, when used with a large sample (1,000) and when other omitted influential covariates have beta paths above .35. We hypothesized that larger beta interactions would be even more detectable for SEM forests, but further studies are needed. Detecting omitted interactions is crucial since they are rather common in social sciences, and traditional methods used for model

specification search, such as modification indices, are not sensitive to them (Mooijjaart & Satorra, 2009). Despite the prevalence of interaction effects in social sciences, Cortina et al. (2021) noted that researchers often avoid latent variable models when dealing with interactions, possibly due to unfamiliarity and methodological obstacles. In addition, the lack of a reliable method to detect influential omitted interactions in latent models, such as SEM, may deter their use. While indices for detecting interactions exist for diagnostic classification models (Brown & Templin, 2023), to our knowledge such indices for SEM are still needed. Thus, using SEM forests to detect omitted covariates and covariate interactions could encourage the inclusion of interaction effects in latent variable models.

A fundamental drawback of SEM forests as a modification search method is their inability to specify where and how (i.e., with which functional form) to include detected omitted influential covariates and interactions in the model. Thus, practitioners may know through SEM forests *that* a set of covariates is relevant and should be included in their models, but they still should decide *where* and *how* to include those covariates or interactions. However, SEM forests include partial dependence plots, allowing users to conduct an exploratory analysis of how a given important predictor or set of predictors influences a given model parameter. Moreover, Brandmaier et al. (2016) provided an example of model modification based on SEM forests variable importance in a single-factor analysis with the Wechsler Adult Intelligence–Revised. As a result of an SEM forest analysis with 1,000 trees on a single-factor model that hypothesizes one latent factor for verbal cognitive ability, authors included the effect of the most important predictor detected by the importance analysis (i.e., education) on the factor structure. Brandmaier et al. (2016) tested the hypothesis that education predicted differences in mean verbal performance, including education as an exogenous predictor and restricting its effect to zero. The model fit including the zero constraint was unacceptable, but when the zero constraint was freeing, model fit

improved, a result that was confirmed by a likelihood ratio test. Brandmaier et al. (2016) exemplify how to include influential covariates in simple models, but the challenge remains for complex models with multiple latent variables. Further tutorial papers with step-by-step guidance on using SEM forests and incorporating influential covariates in complex models are still needed.

The conclusions presented in this study are limited to the simulation parameters and should not be generalized beyond these conditions. Further studies are needed to explore the impact of model complexity and non-linear relationships on SEM forests' performance, and to determine how and where to include omitted influential covariates. Random forests are particularly suitable for non-linear relationships since they are ensembles of decision trees that split the data recursively based on predictor values, creating partitions without assuming linearity (Breiman, 2001). Moreover, the ensemble nature of SEM forests and the non-parametric nature of their model search enhance their ability to capture diverse patterns, including non-linear relationships (Brandmaier & Jacobucci, 2023). While random forests are ideal for non-linear relationships, SEM forests can also handle linear relationships, as the results of this study suggest. Potentially, practitioners might use SEM forests to detect both linear and non-linear omitted covariate-latent variable paths, however, more research is needed to confirm SEM forests' performance with non-linear omitted influential covariates. On the other hand, SEM forests' potential ability to detect both linear and non-linear relationships would pose a challenge in determining the functional form to include a covariate in a model, as covariates may relate linearly or non-linearly to one or more latent variables.

Practitioners need to consider that while SEM forests are able to identify measured omitted influential covariates, they cannot detect unmeasured influential ones. Mixture multigroup factor analysis (MMG-FA) (De Roover, 2021; De Roover et al., 2022) offers an

alternative by identifying latent clusters that correspond to unmeasured covariates.

Specifically, MMG-FA clusters model groups according to a specific level of measurement invariance (e.g., equal factor loadings, equal intercepts) and allows exploring potential measured and unmeasured covariates that explain cluster memberships (De Roover, 2021; De Roover et al., 2022). Tree-based analysis can partially capture the effect of unmeasured influential covariates if they are highly correlated with measured covariates selected for splitting, with those correlated measured covariates serving as proxies and capturing some effect of the unmeasured ones, as suggested by Strobl et al. (2015) with Rasch trees. In these cases, predictive accuracy might hold, but interpretability could suffer, since those measured covariates selected for splitting could simply reflect the true effect of a correlated unmeasured influential covariate, and interpreting those proxies as having a direct causal impact on the model would be misleading (Strobl et al., 2015).

Users of SEM forests should also consider that the *semtree* R package (Brandmaier et al., 2023) calculates marginal variable importance, which can be biased when predictors are correlated (Strobl et al., 2008). Strobl et al. (2008) recommend using conditional variable importance for random forests, which are not yet included in the *semtree* R package (Brandmaier et al., 2023). This study set predictor correlations between zero and .1 to avoid convergence issues and making marginal variable importance applicable. Moreover, covariate correlations were set low considering that random forests marginal variable importance tends to overselect correlated predictors (Strobl et al., 2008), and that with highly correlated predictors, importance is often spread across them rather than attributed to a single predictor, leading to potential misinterpretation of which variables are most important (Breiman, 2001). Although predictors in psychological research tend to show low to moderate correlations, ranging from .20 to .50 (Cohen, 1988), the covariate correlations specified in this study are smaller than the typical predictor correlations found in social

sciences. Thus, practitioners need to consider predictor correlations when interpreting SEM forests variable importance results, and further studies that explore the impact of higher covariate correlations on SEM forests variable importance performance are needed. Finally, practitioners should note that since SEM forests use likelihood ratio tests for split selection, they may inherit biases proper of the chi-square test, such as undesired influences of sample size and parameter magnitudes (cf. Saris et al., 2009). However, as it is common practice with random forests, we allow individual trees to grow deep without applying a statistical cut-off, setting the alpha level to 100% (i.e., never stop splitting based on the chi-square test statistic).

In summary, practitioners can benefit from using SEM forests to address model misspecification due to omitted influential covariates, particularly with samples above 1,000, or around 500 if the omitted covariate-latent variable paths are high (around .5). SEM forests are also particularly effective in detecting omitted influential covariates with paths to multiple latent variables. However, given the lack of clear guidance on where and how to include influential covariates detected by SEM forests, we suggest that their inclusion should be guided by causal inquiries. An interpretation guided by theoretical frameworks and that considers different levels of causation (e.g., counterfactual, intervention, association) (Pearl, 2009) can help users to understand the model structure, hypothesize the direct or indirect effects of predictors, and determine what type of formal relationships are expected in the model.

### **3.9. Data Availability Statement**

The data that support the findings of this study are openly available in the Open Science Framework (OSF) at <http://doi.org/10.17605/OSF.IO/EAPHC>.

### 3.10. References Study 1

- Arnold, M., Voelkle, M. C., & Brandmaier, A. M. (2020). Score-guided structural equation model trees. *Frontiers in Psychology*, 11, 564403.  
<https://doi.org/10.3389/fpsyg.2020.564403>
- Basu, S., Kumbier, K., Brown, J. B., & Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 115, 1943–1948.  
<https://doi.org/10.1073/pnas.171123611>
- Brandmaier, A. M., Prindle, J. J., Arnold, M., & Van Lissa, C. J. (2023). Semtree: Recursive Partitioning for Structural Equation Models. R package version 0.9.19.  
<https://github.com/brandmaier/semtree>
- Brandmaier, A. M., & Jacobucci, R. C. (2023). Machine learning approaches to structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling*. (2nd ed.), (pp. 722–739). Guilford.
- Brandmaier, A. M., Prindle, J. J., McArdle, J. J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. *Psychological Methods*, 21, 566–582. <https://doi.org/10.1037/met0000090>
- Brandmaier, A. M., Ram, N., Wagner, G. G., & Gerstorf, D. (2017). Terminal decline in well-being: The role of multi-indicator constellations of physical health and psychosocial correlates. *Developmental Psychology*, 53, 996–1012.  
<https://doi.org/10.1037/dev0000274>
- Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, 18, 71–86.  
<https://doi.org/10.1037/a0030001>



Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.

<https://doi.org/10.1023/A:1010933404324>

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Chapman and Hall/CRC, <https://doi.org/10.1201/9781315139470>

Brown, C., & Templin, J. (2023). Modification indices for diagnostic classification models. *Multivariate Behavioral Research*, 58, 580–597.

<https://doi.org/10.1080/00273171.2022.2049672>

Bühlmann, P., & Yu, B. (2002). Analyzing bagging. *The Annals of Statistics*, 30, 927–961.

<https://doi.org/10.1214/aos/1031689014>

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd ed.). Routledge.

<https://doi.org/10.4324/9780203771587>

Cortina, J. M., Markell-Goldstein, H. M., Green, J. P., & Chang, Y. (2021). How are we testing interactions in latent variable models? Surging forward or fighting shy? *Organizational Research Methods*, 24, 26–54.

<https://doi.org/10.1177/1094428119872531>

De Roover, K. (2021). Finding clusters of groups with measurement invariance: Unraveling intercept non-invariance with mixture multigroup factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 28, 663–683. <https://doi.org/10.1080/10705511.2020.1866577>

De Roover, K., Vermunt, J. K., & Ceulemans, E. (2022). Mixture multigroup factor analysis for unraveling factor loading noninvariance across many groups. *Psychological Methods*, 27, 281–306. <https://doi.org/10.1037/met0000355>

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems. *The Journal of*

Machine Learning Research, 15, 3133– 3181.

<https://doi.org/10.5555/2627435.2697065>

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Böhner, 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>

Holbert, R. L., & Stephenson, M. T. (2002). Structural equation modeling in the communication sciences, 1995–2000. *Human Communication Research*, 28, 531–551.

<https://doi.org/10.1111/j.1468-2958.2002.tb00822.x>

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research*, 23, 69–86. [https://doi.org/10.1207/s15327906mbr2301\\_4](https://doi.org/10.1207/s15327906mbr2301_4)

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. (4th ed.). Guilford.

Li, J., Zhang, M., Li, Y., Huang, F., & Shao, W. (2021). Predicting students' attitudes toward collaboration: Evidence from structural equation model trees and forests. *Frontiers in Psychology*, 12, 604291. <https://doi.org/10.3389/fpsyg.2021.604291>

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, 111, 490–504. <https://doi.org/10.1037/0033-2909.111.3.490>

Marcoulides, G. A., & Drezner, Z. (2001). Specification searches in structural equation modeling with a genetic algorithm. In Marcoulides, G. A. & Schumacker, R. E. (Eds.), *New developments and techniques in structural equation modeling*. (pp. 247–268). Psychology Press. <https://doi.org/10.4324/9781410601858>

- Marcoulides, G. A., & Drezner, Z. (2003). Model specification searchers using ant colony optimization algorithms. *Structural Equation Modeling: A Multidisciplinary Journal*, 10, 154–164. [https://doi.org/10.1207/S15328007SEM1001\\_8](https://doi.org/10.1207/S15328007SEM1001_8)
- Marcoulides, G. A., Drezner, Z., & Schumacker, R. E. (1998). Model specification searches in structural equation modeling using tabu search. *Structural Equation Modeling: A Multidisciplinary Journal*, 5, 365–376. <https://doi.org/10.1080/10705519809540112>
- Meinshausen, N. (2010). Node harvest. *The Annals of Applied Statistics*, 4, 2049–2072. <https://doi.org/10.1214/10-AOAS367>
- Meyvis, T., & Van Osselaer, S. M. (2018). Increasing the power of your study by increasing the effect size. *Journal of Consumer Research*, 44, 1157–1173. <https://doi.org/10.1093/jcr/ucx110>
- Mooijjaart, A., & Satorra, A. (2009). On insensitivity of the chi-square model test to nonlinear misspecification in structural equation models. *Psychometrika*, 74, 443–455. <https://doi.org/10.1007/s11336-009-9112-5>
- Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58, 415–434. <https://doi.org/10.2307/2283276>
- Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. Chapman and Hall/CRC. <https://doi.org/10.1201/9781439800393>
- Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How many trees in a random forest?. In Perner, P. (Ed.), *Machine learning and data mining in pattern recognition. MLDM 2012. Proceedings*. 8 (pp. 154–168). Springer. [https://doi.org/10.1007/978-3-642-31537-4\\_13](https://doi.org/10.1007/978-3-642-31537-4_13)
- Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling: A*

- Multidisciplinary Journal, 8, 287–312. [https://doi.org/10.1207/S15328007SEM0802\\_7](https://doi.org/10.1207/S15328007SEM0802_7)
- Pearl, J. (2009). *Causality: models, reasoning, and inference*. (2nd ed.). Cambridge University. <https://doi.org/10.1017/CBO9780511803161>
- Pornprasertmanit, S., Miller, P., Schoemann, A. M., & Jorgensen, T. D. (2022). Simsem: SIMulated structural equation modeling. R package version 0.5-16.909. Retrieved from <https://CRAN.R-project.org/package=simsem>
- Saris, W. E., Satorra, A., & Sorbom, D. (1987). The detection and correction of specification errors in structural equation models. *Sociological Methodology*, 17, 105–129. <https://doi.org/10.2307/271030>
- Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal*, 16, 561–582. <https://doi.org/10.1080/10705510903203433>
- Scheines, R., Spirtes, P., Glymour, C., Meek, C., & Richardson, T. (1998). The tetrad project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33, 65–117. [https://doi.org/10.1207/s15327906mbr3301\\_3](https://doi.org/10.1207/s15327906mbr3301_3)
- Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9, 307. <https://doi.org/10.1186/1471-2105-9-307>
- Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, 80, 289–316. <https://doi.org/10.1007/s11336-013-9388-3>
- Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and

random forests. *Psychological Methods*, 14, 323–348.

<https://doi.org/10.1037/a0016973>

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with “well fitting” models.

*Journal of Abnormal Psychology*, 112, 578–598. <https://doi.org/10.1037/0021-843X.112.4.578>

Usami, S., Hayes, T., & McArdle, J. (2017). Fitting structural equation model trees and latent growth curve mixture models in longitudinal designs: The influence of model misspecification. *Structural Equation Modeling: A Multidisciplinary Journal*, 24, 585–598. <https://doi.org/10.1080/10705511.2016.1266267>

Usami, S., Jacobucci, R., & Hayes, T. (2019). The performance of latent growth curve model-based structural equation model trees to uncover population heterogeneity in growth trajectories. *Computational Statistics*, 34, 1–22.

<https://doi.org/10.1007/s00180-018-0815-x>

Van Lissa, C. J., Beinhauer, L., Branje, S., & Meeus, W. H. J. (2023). Using machine learning to identify early predictors of adolescent emotion regulation development. *Journal of Research on Adolescence: The Official Journal of the Society for Research on Adolescence*, 33, 870–889. <https://doi.org/10.1111/jora.12845>

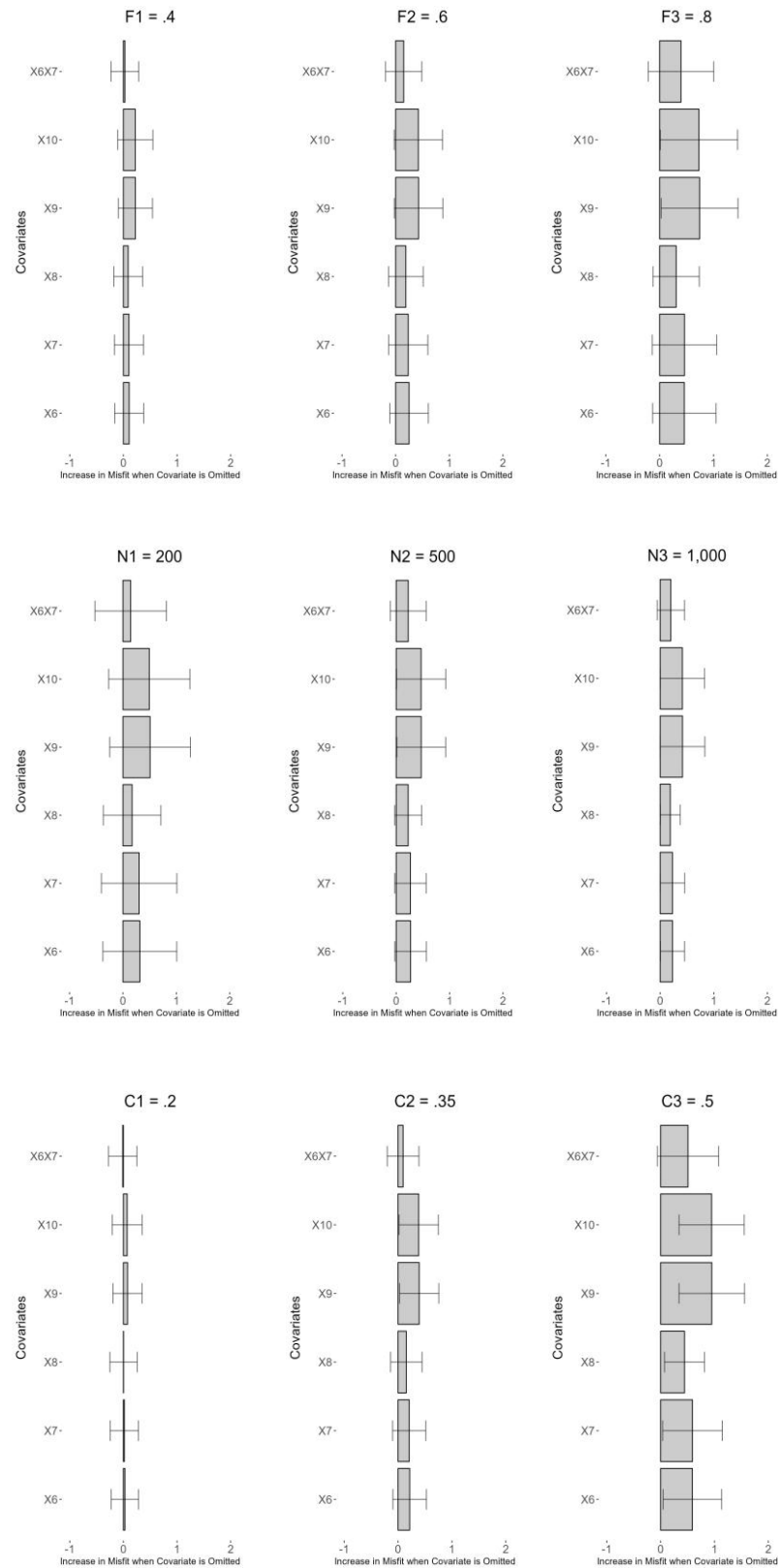
Van Voorhis, C. W., & Morgan, B. L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology*, 3, 43–50. <https://doi.org/10.20982/tqmp.03.2.p043>

Wilms, R., McEathner, E., Winnen, L., & Lanwehr, R. (2021). Omitted variable bias: A threat to estimating causal relationships. *Methods in Psychology*, 5, 100075. <https://doi.org/10.1016/j.metip.2021.100075>

### 3.11. Appendices

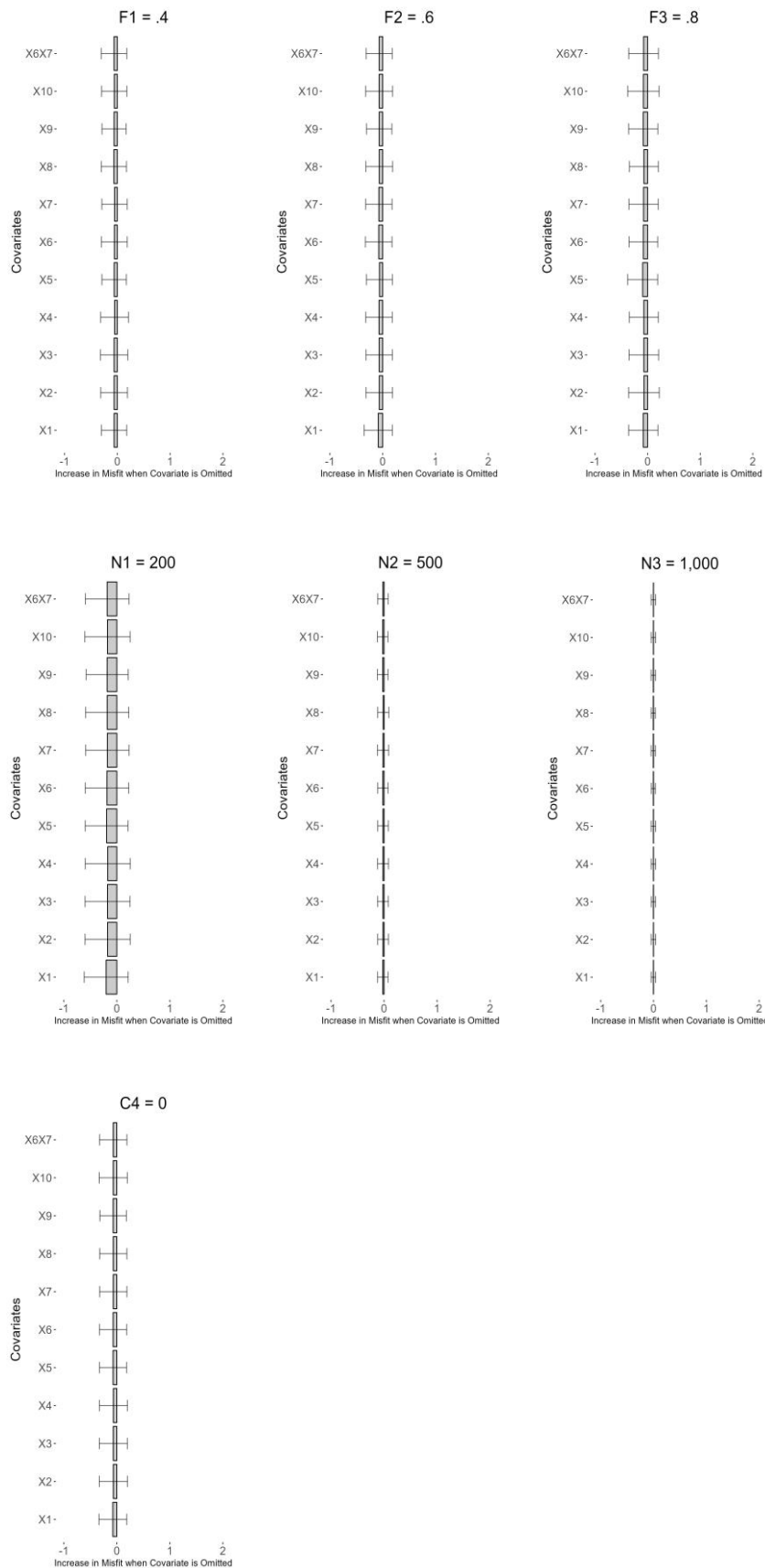
#### 3.11.1. Appendix A – Study 1

##### *Average Variable Importance per Experimental Condition for Influential Covariates*



### 3.11.2. Appendix B – Study 1

#### Average Variable Importance per Experimental Condition for Non-Influential Covariates



## **4. Study 2: A Practical Guide to Use SEM Forests for Specification Search in Structural Equation Modeling**

### **4.1. Abstract**

Structural equation modeling (SEM) is widely used in social sciences and educational research. Since model misspecification is quite prevalent in applied SEM, researchers frequently conduct specification searches to refine model structure and improve fit. Although modification indices are the most common approach for specification search, they have several flaws, such as overlooking influential variables not initially included in the model and failing to detect interactions. Consequently, multiple alternatives for conducting specification search have been proposed in the literature. In this paper, we provide a step-by-step guide to use SEM forests as an effective tool for identifying omitted influential covariates in SEM, using the free R package *semTree*.

*Keywords:* SEM forests, structural equation modeling, guideline specification search, omitted variables

### **4.2. Introduction**

Structural equation modeling (SEM) is a statistical technique widely used in psychological and educational research since it allows to establish relationships among observed and unobserved variables, while accounting for measurement error. Kline (2016) outlines a standard SEM analysis consisting of six basic steps. First, practitioners should propose a theory-based hypothetical model that defines the expected relationships among latent and observed variables (step 1) and assess whether unique model parameters can be estimated from the initial model, ensuring that the model is identified (step 2). If the model is identified, practitioners proceed to select appropriate measures and collect data (step 3). Then, practitioners calculate parameter estimates (e.g., factor loadings) and evaluate model fit (step 4). If the model does not adequately explain the data (indicating misspecification), a



condition known as model misfit, practitioners may conduct a specification search and a model respecification (step 5). If the respecified model fits the data, users can then proceed to interpret and report the results (step 6). In this study, we provide a guideline for using SEM forests as a novel tool for specification search, capable of identifying misspecifications related to influential covariates that were not initially included in the model.

Model misspecification (addressed in step 5) is more the rule than the exception in applied SEM, as multivariate models in the social sciences often fail to account for all relevant variables and relationships. Relying on misspecified models may lead to biased parameters estimates and incorrect interpretations (e.g., Kaplan, 1988; Mulaik, 2009). Even well-fitting models should be interpreted with caution, as various ambiguities can affect their validity, such as equivalent models, omitted variables, and potential biases of fit measures (Tomarken & Waller, 2003).

This article serves as a guide on how to conduct specification search in SEM using SEM forests. The objective of specification search in SEM is to identify missing paths and improve model fit, aiming for a parsimonious and substantively meaningful model that fits the data (Kline, 2016; MacCallum, 1986). Specification search in SEM is often conducted through a modification index-based approach (MacCallum et al., 1992; Mulaik, 2009; Saris et al., 2009), which several software packages offer for analysis (e.g., R package lavaan, Mplus, R package OpenMx, AMOS in SPSS). Modification indices are used to identify potentially omitted paths (e.g., regression weights) in the initial hypothesized model, which are then added in a sequential model modification process to significantly improve model fit (Kline, 2016; MacCallum, 1986). Specifically, modification indices estimate the increase in chi-square-based model fit resulting from the removal of a parameter restriction, suggesting a sequence of progressively more complex models. However, modification indices are not sensitive to interaction effects (Mooijart & Satorra, 2009) and assume the initial model is

correctly specified in terms of included variables, overlooking potential influential omitted variables.

To address these limitations, this article provides a step-by-step practical guide for using SEM forests (Brandmaier et al., 2016) as a complementary approach for specification search associated with omitted covariates in SEM, allowing more accurate models and improved models interpretations. The next section gives an introduction to how SEM forests work, followed by a description of the data set used in this guide to exemplify how to conduct an SEM specification search with SEM forests. Then we proceed with an actual SEM specification search using the R package *semtree* (Brandmaier et al., 2024), providing the relevant code lines and outcomes, and explaining how to use and interpret SEM forests.

#### **4.3. SEM Forests as a Tool for SEM Specification Search**

SEM forests, developed by Brandmaier et al. (2016), are a decision-tree-based method to explore heterogeneity in a data set. SEM forests are random forests, a tree-based nonparametric statistical approach for classification analysis by recursive partitioning (Brandmaier et al., 2016; Breiman, 2001), applied to SEM. Heterogeneity implies that subgroups in a data set may have differences in parameter estimates (e.g., path coefficients, residuals, etc.) due to group membership. SEM forests emphasize the identification of omitted influential covariates that may explain group membership. The inclusion of the identified covariates is expected to reduce heterogeneity and enhance the model's predictive power. Thus, SEM forests serve as a tool for SEM specification search related to omitted influential covariates, since the inclusion of omitted covariates identified by the SEM forests should increase the fit of the model with the data. Beyond this use, SEM forests can also help to identify explanatory variables, assess their influence, determine which variables to control for, and generate hypotheses about structural relations in large data sets (Brandmaier et al., 2016).

SEM forests were developed based on SEM trees (Brandmaier et al., 2013). SEM trees, a tool to account for heterogeneity in a data set, assume that observed data are drawn from different underlying multivariate normal distributions. SEM trees recursively split the initial data set into subgroups based on covariates not included in the model, identifying those that explain the largest differences in model parameters (Brandmaier et al., 2013). However, SEM tree results can be unstable, as splits may be influenced by sample-specific factors, with small node perturbations potentially cascading and undermining generalizability. To address this limitation, Brandmaier et al. (2016) introduced SEM forests, which aggregate results of individual trees to form an ensemble. A covariate consistently selected by individual trees indicates its stability as a predictor in the aggregated forest results. This aggregation mitigates the instability of individual trees, providing more robust and reliable outcomes. For more technical details, we refer interested readers to Arnold et al. (2020), Brandmaier et al. (2013, 2016), and Silva Díaz et al. (2024).

#### **4.4. A Running Example**

This guide describes how to conduct a misspecification search using SEM forests with data from the Self-Concept and Interest when Studying Mathematics (SISMa) project, which explores the role of affective variables during the university transition phase in mathematic-related university programs (Kosiol et al., 2019; Rach & Ufer, 2020; Ufer et al., 2017). Within the SISMa project, a series of studies were conducted to examine mathematical self-concept and mathematical self-interest during the study entry phase in mathematic-related university study programs in Germany. From an extensive set of variables collected by the SISMa project, we defined a longitudinal model with three self-concept facets collected during the start of the first semester (t1) and about eight weeks into the first semester (t2) as predictors. As outcome we defined three facets of study satisfaction collected at t2.

Additionally, the data set includes eight potentially informative covariates, such as demographics, grades, and interest in mathematics.

The model proposed for this SEM specification search example includes general mathematical self-concept measured at t1 and t2 as potential predictors of tendency to change study program. A set of eight variables are explored as potential informative covariates for this longitudinal model: general interest in mathematics, school-related interest in mathematics, university-related interest in mathematics, mathematical performance (measured via a short-written test at t1), general school qualification, study program (Bachelor of Mathematics or Bachelor in Teacher Education), age and gender. Items for all questionnaire scales were measured using four-point Likert scales.

#### **4.5. A Practical Guide for Model Specification Search with SEM Forests**

This section presents a step-by-step procedure for using SEM forests to identify potential informative covariates in SEM. To illustrate this process, a subset of data collected by the SISMa project is described and analyzed, providing the R code using the *semtree* package (Brandmaier et al., 2024). The steps needed to conduct a specification search with SEM forests are: 1. model specification, 2. SEM forests growing, 3. SEM forests variable importance analysis, 4. Model respecification, and 5. Evaluation of model fit.

The data analyzed in this study is composed of 202 observations and 18 variables. Ten of these variables are indicators of three latent variables, while the remaining eight are potentially informative covariates. The latent variable *general mathematical self-concept* was measured by four indicators at t1 and also by four indicators at t2, while the outcome variable *tendency to change study program* was measured by two indicators at t2. Regarding the potential informative covariates, *general interest in mathematics* was measured on a continuous scale ranging from zero to three, while *school-related interest in mathematics* and *university-related interest in mathematics* were measured on a continuous scale ranging from

zero to three. *Mathematical performance* was measured on a continue scale ranging from zero to one, *general school qualification* was on a continue scale ranging from one to four with one indicating the best grade since the German system uses inverted school grades, *study program* was a dichotomous variable depending on the study program (i.e., Bachelor of Mathematics or Mathematics Teacher Education Program), *gender* was also measured dichotomously (i.e., female, male), and *age* was measured in a continuous scale.

#### **4.5.1. Model Specification**

Regarding model specification, the first step to conduct a specification search with SEM forests, the *semtree* package supports models specified with the R libraries *Open MX* (Boker et al., 2023) and *lavaan* (Rosseel, 2012). *OpenMx* provides greater flexibility and control over more details of the model specification, making it suitable for advanced users, but it requires more coding expertise. *Lavaan* is more intuitive and user-friendly, making it accessible to users with less coding expertise, though it offers less customization for complex models. For this study, *lavaan* was chosen for model specification since the initial model used to exemplify the use of SEM forests for specification search is a simple longitudinal model with only three latent variables and 10 indicators.

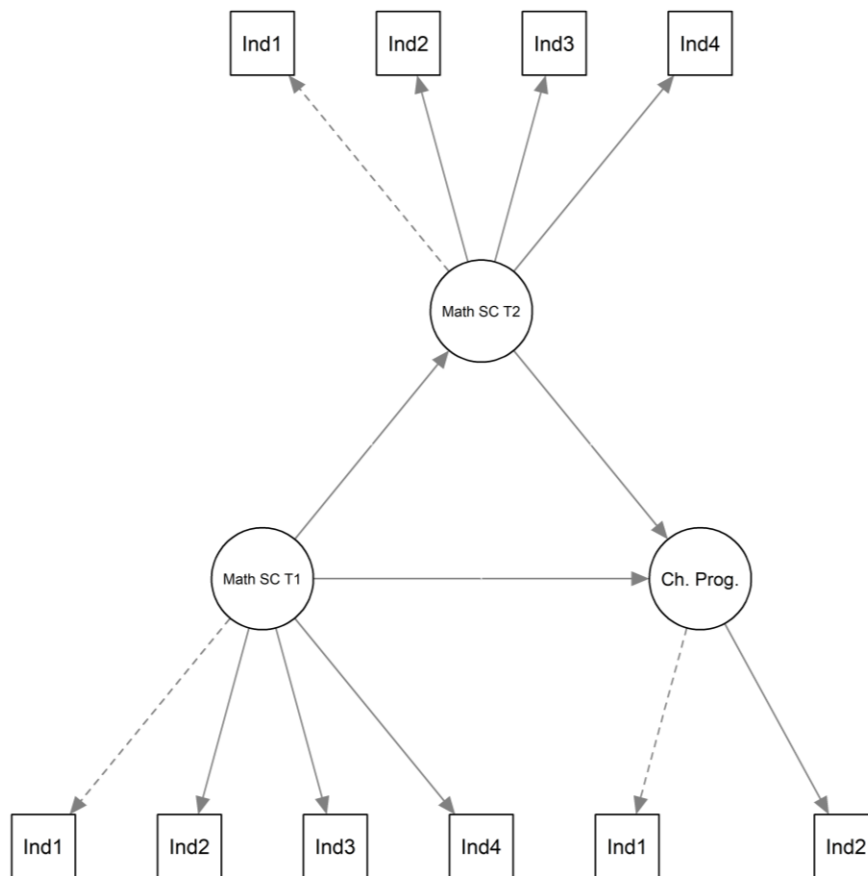
Practitioners should tailor their data preparation process according to the specific characteristics of each data set. For the data analyzed in this study, variables with few unique values (i.e., gender and study program) were defined as factor variables, while all other variables were defined as numeric. Missing values were imputed using the R package *mice* (Buuren & Groothuis-Oudshoorn, 2011), which allows imputing incomplete multivariate data by chained equations. Predictive mean matching (Little, 1988) was used to impute numeric variables, while polytomous regression was applied to categorical variables. Missing values were not imputed for gender, age and general school qualification since these variables have an extensive contextual meaning that cannot be accounted for by imputation methods. The

proportion of missing values across variables ranged from 0 to 16.83% ( $M = 1.87\%$ ,  $SD = 3.85\%$ ), with all variables having percentage of missing data below 3.5%, except for the covariate *school-related interest in mathematics* which had a missing value rate of 16.83%. After data imputation, the final sample, called *ds\_imp1*, is composed of 202 observations for the SEM forests analysis.

Using the lavaan function *sem()*, an initial model was generated and saved in the object *sem\_initial\_model*:

```
sem_initial_model <- sem(model = initial_model, data = ds_imp1, do.fit = TRUE).
```

Within the *sem()* function, first we call the initial model specified with lavaan (i.e., *initial\_model*), then we call the data set prepared for analysis after data imputation (i.e., *ds\_imp1*), and finally with *do.fit = TRUE* the model is fitted to the data estimating the model parameters and computing model fit. Notice that the initial model does not include any covariate (see Figure 1). A lavaan object with the fitted sem model (in this example *sem\_initial\_model*) is required to subsequently generate the SEM forests.

**Figure 1 – Study 2***Initial Model with No Covariates*

*Note.* Initial longitudinal model with Math SC T1 = general mathematical self-concept at t1, Math SC T2 = general mathematical self-concept at t2, and outcome variable Ch. Prog. = tendency to change the study program.

#### **4.5.2. SEM Forests Growing**

The initial model proposed for this example specified only relationships between the indicators and latent variables. Therefore, the next step is to use SEM forests to explore whether any of the 8 covariates included in the data set are informative for the initial model. To generate the SEM forests using the R package *semtree*, the parameters for tree and forest growth must be specified using the functions *semforest.control()* and *semtree.control()*. These

functions allow for the customization of various parameters. At a minimum, we recommend practitioners to specify the following: sampling method, split selection method, number of subsampled covariates, Bonferroni correction, significance level for splitting, minimum node sample size, lower bound for potential terminal nodes, and number of trees.

The sampling method in SEM forests involves creating random subsets of observations from the original sample to grow SEM trees. Two sampling methods are available in the `semtree` package: bootstrapping aggregating (bagging) and subsampling. Bagging generates random samples with replacement, matching each random sample with the size of the original sample, while subsampling draws smaller random samples without replacement. We strongly suggest the use of subsampling since the bootstrapping method is biased in favor of covariates with more categories and those with stronger correlations to the outcome variable, leading to biased variable importance measures (Strobl et al., 2007). Regarding the split method selection, we suggest the use of the recently implemented score-guided tests, which are unbiased in covariate selection and have higher statistical power compared to the original SEM tree methods (i.e., naïve and fair) (Arnold et al., 2020). Additionally, score-guided tests are computationally more efficient because they require the estimation of only one model, in contrast to the original likelihood ratio approach implemented in SEM trees, which needs to estimate models for each split candidate (Arnold et al., 2020).

To enhance tree diversity, SEM forests evaluate only a random subset of covariates at each split. Users must specify the size of this subset. A smaller subset increases tree variability and independence but may overlook relevant predictors and high-order interactions (Brandmaier et al., 2016). The covariates subset size can be determined heuristically using  $\log_2(m) = c$ , where  $m$  is the total number of covariates and  $c$  is the subset size. For this example, with  $m = 8$ , the subset  $c$  was set to three covariates. Regarding the



Bonferroni correction, it addresses the multiple-testing problem that arises when evaluating multiple covariates for node splitting at the same time, artificially inflating type-I error rates. However, Bonferroni correction is overly conservative, generating sparse trees and hugely reducing the power of the SEM forests to identify potential informative covariates when having a large set of covariates (Arnold et al., 2020; Brandmaier et al., 2013). In a simulation study, Silva Díaz et al. (2024) report type-I errors rate close to zero, even without Bonferroni correction, for samples between 200 and 1,000. Consequently, Bonferroni correction was not applied in this study, and we suggest that practitioners avoid using it unless they have serious concerns about type-I error rates.

As for the alpha level for splitting at a given node, it was set to 100% in this analysis due to the relatively small sample size ( $N = 202$ ). This liberal criterion allows deeper trees and increases the power for detecting interactions, but at the cost of Type-I error, meaning a possible increase of the incorrect detection of non-influential covariates (Silva Díaz et al., 2024). Users with large samples may set smaller alpha levels to reduce the risk of Type I error. However, the results by Silva Díaz et al. (2024) indicate minimal type I error rates for small samples, despite having high alpha levels.

As for the lower bound of a potential terminal node and the minimum node sample size, users should be aware that too small nodes may affect the stability and robustness of the estimates. Based on SEM forests' power and type-I error rates for a sample of 200 (Silva Díaz et al., 2024), the lower bound for terminal nodes was set to 50, requiring at least 50 observations in any child node of a potential split to be valid, while the minimum sample size per node was set to 100, preventing splits attempts in nodes with fewer observations. These decisions imply for our data set ( $N = 202$ ) having SEM forests with only two possible splits per tree. To mitigate the limitations of shallow tree depths, generating SEM forests with a large amount of trees is essential. Determining the optimal number of trees in a forest

depends on factors such as the number of predictors and their interactions, data heterogeneity, and model complexity. While specific thresholds for the number of trees in SEM forests are yet to be established, Brandmaier et al. (2016) heuristically suggest using a relatively large number, such as 2,000 trees. Following this recommendation, we used 2,000 trees for our SEM forest. The R code implementing all the parameters described is presented in Figure 2.

## Figure 2 – Study 2

*Code to Generate the SEM Forests using the R package semtree*

```
# Control object to specify forests growing
control <- semforest.control()

control$num.trees <- (num.trees <- 2000)

control$sampling <- "subsample" # More accurate than bootstrapping
control$semtree.control$method <- "score" # Other methods: naive, fair
control$mtry <- 3 # log2(8), being 8 the total number of covariates
control$semtree.control$min.N <- 100
control$semtree.control$min.bucket <- 50
control$semtree.control$bonferroni <- F
control$semtree.control$alpha <- 1
print(control)

# Grow the forest
forest_dsreal <- semforest(sem_initial_model, ds_imp1, control)
```

3. SEM Forests Variable

### 4.5.3. Importance Analysis

After generating the SEM forest, users can calculate variable importance, a metric that quantifies the relative importance of each explored covariate. SEM forests variable importance is calculated by averaging, across all trees, the difference in prediction accuracy

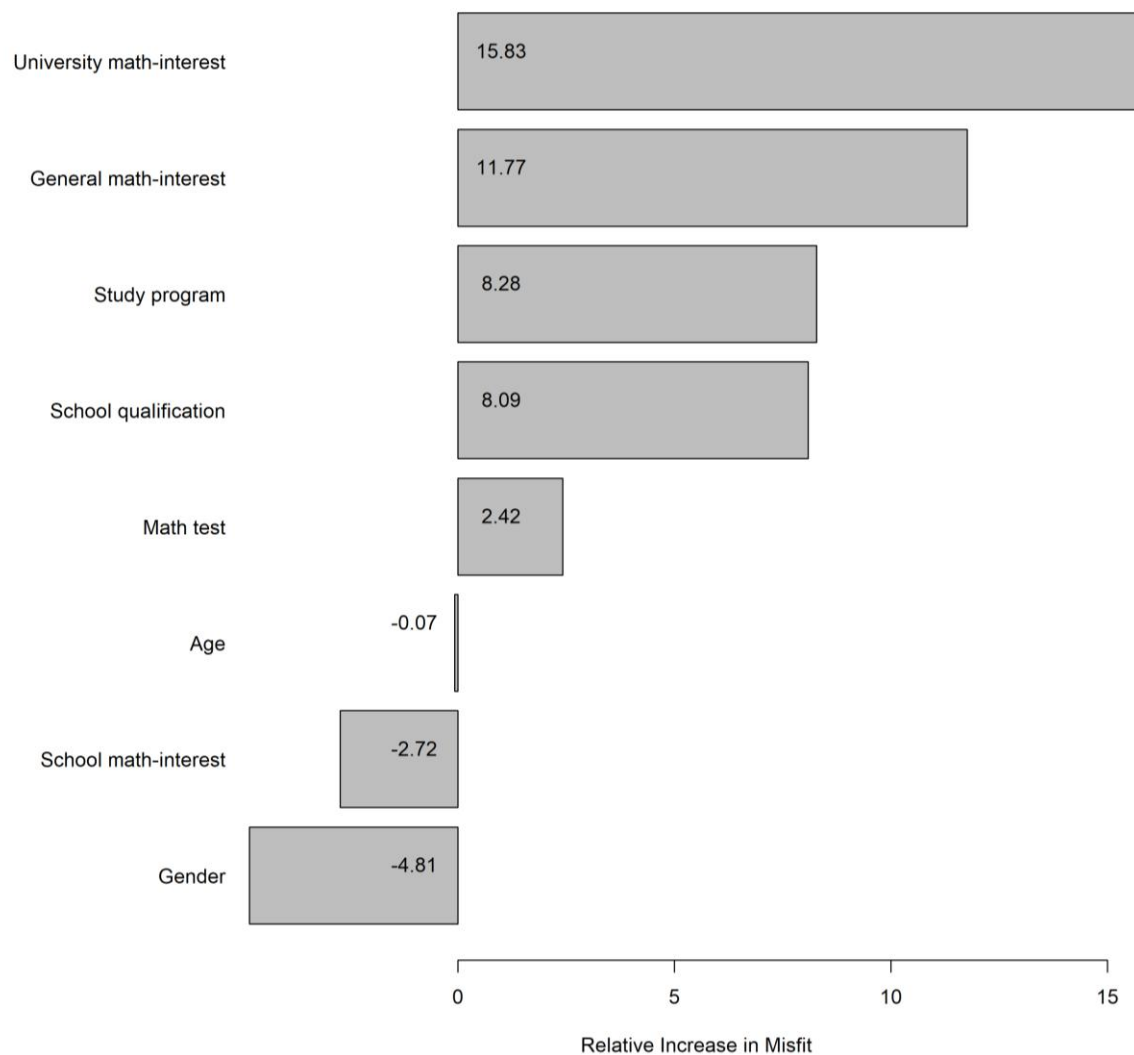
before and after permuting the values of each covariate explored (Brandmaier et al., 2016; Strobl et al., 2009). Specifically, SEM forest variable importance implements permutation accuracy importance, which measures the impact of randomly permuting a covariate's values on the goodness-of-fit of each estimated model. Permutation serves as a proxy of removing a covariate's effect on the model, so the larger the drop of the goodness-of-fit after permutation the greater importance of that covariate for the model (Brandmaier et al., 2016).

Variable importance values can be positive or negative. Values close to zero implies the covariate has little predictive value, indicating inconsistent inclusion in the individual trees or inclusion by chance. Negative values indicate that randomly permuting the covariate's values increased just by chance its predictive value. Conversely, positive values imply that the permutation decreased the predictive value of the covariate, entailing the covariate without permutation is relevant for model goodness-of-fit (Strobl et al., 2009). Using variable importance *z*-scores is discouraged because they depend on the forest's tuning parameters and are therefore not comparable across studies. Similarly, variable importance significance tests are not recommended because their power also depends on arbitrary forest tuning parameter choices, such as the number of trees (Strobl et al., 2009).

SEM forests variable importance is calculated using the *semtree* package function *varimp()* on the SEM forests object generated in step 2 (for this example *forest\_dsreal*). To visualize the results as in Figure 3, use the *plot()* function on the variable importance object (for this example *forest\_dsreal*). To display the specific variable importance values, use *print(varimp\_obj, sort.values = TRUE)*, where *varimp\_obj* is the variable importance object, and *sort.values = TRUE* sorts the variable importance values from the covariate with the smallest importance to the covariate with the largest importance.

### Figure 3 – Study 2

#### *SEM Forests Variable Importance*



*Note.* Relative variable importance of each of the eight potentially influential covariates examined.

Since there are no specific thresholds for interpreting variable importance values, a descriptive ranking of covariates is recommended instead of interpreting or comparing absolute variable importance values. Strobl et al. (2009) suggest excluding for further exploration variables with negative, zero or positive importance scores within the same range as the negative values. This interpretation intends to remove values likely attributable to

randomness, as negative importance scores indicated improved prediction solely due to random value permutations. Since applying the rule suggested by Strobl et al. (2009) in the data set here analyzed would imply including five of the eight covariates explored (see Figure 3), we used a more conservative approach. To avoid selecting covariates with negligible contributions, covariates within at least the top 30% of ranked variable importances were selected as suitable for further exploration. Users should be aware that this 30% heuristic is not suitable in all cases; for instance, if all covariates have very similar importance scores, it may not be informative. Based on our results (see Figure 3), *general interest in mathematics* and *university-related interest in mathematics* were the covariates selected to be included in the initial model.

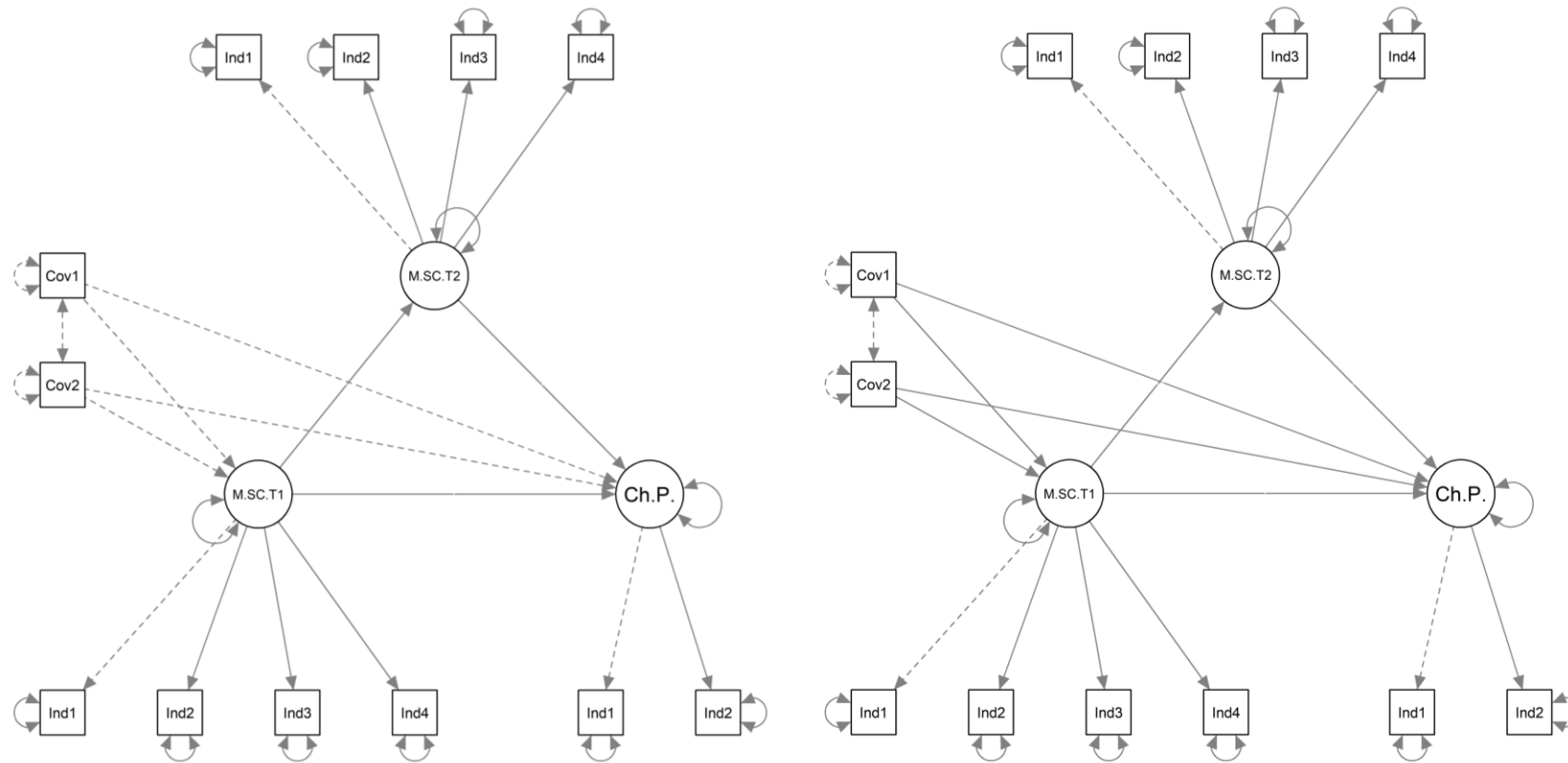
#### **4.5.4. Model Respecification**

If the SEM forests detect any covariate as potentially influential for the initial model, users should decide how to include them. SEM forests variable importance helps to identify a set of influential covariates to be included in the model, but does not specify where and how (i.e., with which functional form) to include them. As with other specification search methods, these decisions should be theory-based to avoid modified models without conceptual meaning, and data-driven overfitting that jeopardizes the performance of the model with new samples.

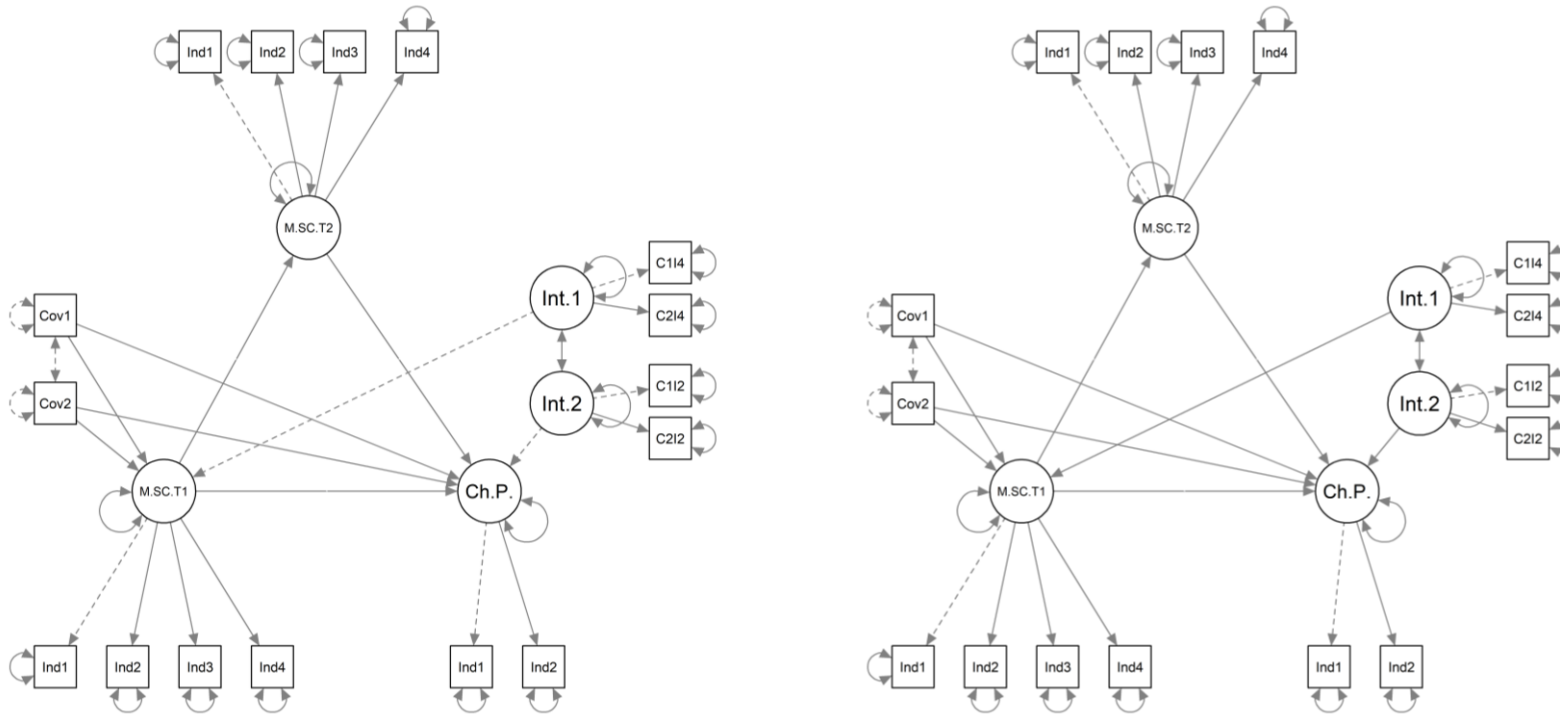
For this example, *general interest in mathematics* and *university-related interest in mathematics* were included in four different ways. The first and second modified models include the two covariates as direct predictors of *general mathematical self-concept at t1* and the outcome variable *tendency to change study program* (see Figure 4). For the model represented at the left of Figure 4, the influences of the two covariates were restricted to zero, while for the model at the right of Figure 4 the influence of the covariates were freed. If the covariates *general interest in mathematics* and *university-related interest in mathematics*

were not related to the latent variables *general mathematical self-concept* at t1 and *tendency to change study program*, the model with the covariate influences restricted to zero should have a reasonable fitness, and the model with the free effects should not have a better goodness-of-fit. Model fit results are presented in the next section.

As an alternative specification, the covariates were also included in a moderation model following the rationale described above, with the covariate interaction effects restricted to zero (see model at left of Figure 5) and the covariate interaction effects freed (see model at right of Figure 5). Moderation occurs when a third variable influences the effect of an independent variable on a dependent variable. This is typically specified using interaction terms. For this example, latent interaction effects were calculated as the product of the two identified covariates and one indicator of their related latent variables (i.e., *general mathematical self-concept at t1* and *tendency to change study program*) (see Figure 5). Since the primary objective of this model was not its theoretical interpretation but rather to demonstrate how covariates identified by SEM forests can be incorporated into entirely different model structures— and how goodness-of-fit indices may be equally good in those different structures—we included the identified covariates in two interaction terms without relying on theoretical justification or empirical guidelines. However, we strongly emphasize that practitioners should always incorporate identified covariates based on sound theoretical or statistical foundations.

**Figure 4 – Study 2***Model with Identified Influential Covariate Direct Effects*

*Note.* At left model with covariate effects fixed to zero, at right model with covariate effects free. M.SC.T1 = general mathematical self-concept at t1, M.SC.T2 = general mathematical self-concept at t2, Ch.P. = tendency to change the study program, Cov1 = university-related interest in mathematics, Cov2 = general interest in mathematics.

**Figure 5 – Study 2***Moderation Model with Identified Influential Covariates*

*Note.* At left model with covariate interaction effects fixed to zero, at right model with covariate interaction effects free. M.SC.T1 = general mathematical self-concept at t1, M.SC.T2 = general mathematical self-concept at t2, Ch.P. = tendency to change the study program, Cov1 = university-related interest in mathematics, Cov2 = general interest in mathematics, Int.1 = interaction covariates and indicator 4 of general mathematical self-concept at t1, Int.2 = interaction covariates and indicator 2 of tendency to change the study program.



The resulting products of the interactions were double-mean centered, meaning mean-centering the variables before computing the interaction terms and then mean-centering the interaction term itself. Double-mean centering prevents multicollinearity between the interaction terms and low-order terms (e.g., main effects) and improves the interpretability of the interaction term (Lin et al., 2010). The interaction effects and the double-mean centering of the resultant products were calculated using the function *indProd()* of the R package SemTools (Jorgensen et al., 2022).

#### 4.5.5. Evaluation of Model Fit

Once the users have defined the modified models, they should compare the goodness-of-fit of the models to evaluate if the introduced covariates improved the model. Table 1 shows the CFI, RMSEA, and SRMR of the five models specified, named, initial model, modified model with covariates fixed, modified model with covariates free, moderation model with covariate interactions fixed, and moderation model with covariate interactions free. The model fit including the zero constraint for the covariate direct effects model was unacceptable (RMSEA = .09, CFI = .89, SRMR = .15). For the model with the covariate direct effect, when freeing the regression of *general interest in mathematics* and *university-related interest in mathematics* on *general mathematical self-concept at t1* and *tendency to change study program* model fit was acceptable (RMSEA = .07, CFI = 0.94, SRMR = .06). The direct effect model with the zero constraint also had higher BIC (3453.7) and AIC (3377.7) than the direct effect model without constraints (BIC = 3428.3, AIC = 3339.0), indicating that the direct effect model without constraints is preferred (see Table 2). The inclusion of the covariate regressions in the direct effect models was formally tested with a likelihood ratio test that significantly rejected the removal of the regression path ( $\chi^2 = 46.68$ ,  $df = 4$ ,  $p < .001$ ).

**Table 1 – Study 2***Model Fit Comparison*

	Initial Model	Model with Covariate Direct Effects Fixed	Model with Covariate Direct Effects Free	Moderation Model with Covariates Fixed	Moderation Model with Covariates Free
CFI	.94	.89	.94	.93	.93
RMSEA	.09	.09	.07	.06	.06
SRMR	.06	.15	.06	.06	.06

**Table 2 – Study 2***Comparison of Models with Covariates Free and Fixed: AIC, BIC and Chi-Square*

	AIC	BIC	$\chi^2$	DF	$\chi^2$ Difference	DF Difference	p-Value
Modified Model with Covariates Free	3339.0	3428.3	96.291	48			
Modified Model with Covariates Fixed	3377.7	3453.7	142.973	52	46.682	4	< .001

*Note.* *DF* = degrees of freedom

Regarding the moderation models, the goodness-of-fit indices were acceptable for both the moderation model with the covariate interaction effect free (i.e., RMSEA = .06, CFI = .93, SRMR = .06), and the moderation model with the covariate interaction effects fixed (i.e., RMSEA = .06, CFI = .93, SRMR = .06) (see Table 1). However, the likelihood ratio test indicated that the removal of the interaction effects in the moderation model was not statistically rejected ( $\chi^2 = .86$ ,  $df = 2$ ,  $p = .65$ ) (see Table 3). These results suggest that the acceptable goodness-of-fit indices were attributable to the direct effects of the identified covariates rather than the interaction effects proposed in the moderation model. Consequently, we conclude that incorporating the identified covariates as interaction effects in the proposed moderation model was not appropriate for the data analyzed. The models

with direct covariate effects (Figure 4) were not formally compared with the moderation models (Figure 5) using the likelihood ratio test since the moderation model introduced new latent variables (i.e., interaction effects) making them not nested models. In cases where practitioners identify two non-nested models with equally good goodness-of-fit indices and likelihood ratio tests supporting the inclusion of covariates with different structural specifications, the decision regarding which model to adopt should be guided by theoretical plausibility.

**Table 3 – Study 2**

*Comparison of Moderation Models with Covariates Free and Fixed: AIC, BIC and Chi-Square*

	AIC	BIC	$\chi^2$	DF	$\chi^2$ Difference	DF Difference	p-Value
Moderation Model with Covariates Free	3791.8	3917.5	173.83	95			
Moderation Model with Covariates Fixed	3788.7	3907.8	174.69	97	.856	2	.65

*Note.* DF = degrees of freedom

#### 4.6. Concluding Remarks

This study provides a step-by-step guide for SEM users exploring alternative specification search methods to identify covariates not included in the model but considered as potential influential predictors. In the specification search example presented here, SEM forests from the R package semtree (Brandmaier et al., 2024) identified two covariates not initially included in the model as potentially influential. Adding these covariates resulted in a more complete model with an acceptable goodness-of-fit, whether as direct effects on two latent variables or within a moderation model. SEM forests are thus particularly useful for

specification search, as traditional modification indices are not able to identify specific influential covariates not included in the initial model. While modification indices may indicate unmodeled variance due to missing observed variables through high values for error term correlations, they cannot identify which specific omitted observed variables should be included. SEM forests, on the other hand, are able to identify specific covariates not included in the initial model that may improve it. The lack of modification indices to detect influential variables not included in the model may lead to omitted variable bias. This implies that model parameters and model fit indices may be biased due to omitted variables correlated with both included predictors and dependent variables (Kline, 2016; Tomarken & Waller, 2003; Wilms et al., 2021).

Practitioners should, however, bear in mind that the accuracy of the SEM forests depends on factors such as the adequate preparation of the data set (e.g., handling missing values), and the specification of the parameters for tree and forest growth described in this guide, named: sampling method, split selection method, number of subsampled covariates, Bonferroni correction, significance level for splitting, minimum node sample size, lower bound for potential terminal nodes, and number of trees. In addition to those parameters, the *semtree* package provides a wide range of customization options to meet more specific user needs. Regarding the lack of SEM forests to specify where to include the identified omitted covariates within the model, users may use the partial dependence plots included in the *semtree* package to explore possible options. Partial dependence plots allow to visualize the influence of identified influential covariates on specific model parameters. A flat partial dependence plot indicates that the covariate has minimal to no effect, whereas a steep slope suggests an effect. Users may then identify the influential covariate and include it in the model. However, we suggest the use of partial dependence plots with continuous variables or at least with one with more categories, since with discrete variables, as was the case of the

data analyzed in this example (i.e., almost all variables were 4-point Likert scale), partial dependence plots will have only a few points to plot, making it less informative than with continuous variables. Beyond statistical considerations, decisions on covariate inclusion should be conceptually guided, as identified covariates can be incorporated into different models with equally good fit, as shown in this example, where both the direct effects model and the moderation model had similar goodness-of-fit.

While specification search is generally required in SEM since initial hypothesized models rarely fit the data, users should consider that modifications should have theoretical coherence and that excessive post-hoc changes may lead to overfitting. Strategies to evaluate the generalizability of the modified model include cross-validation, fitting the modified model with two separate independent sample, split-sample validation, dividing the data set into specification search and validation subsets, or bootstrapping. This involves estimating parameter stability through bootstrap samples generated by randomly resampling the original data set. Practitioners must also be aware of the risk of relying merely on empirical evidence when respecifying a misfitting model.

Relying only on empirical evidence increases the risk of capitalization on chance, leading to results that cannot be generalized, and interpretations and conclusions based on random variations proper of a particular sample more than on true underlying effects (MacCallum et al., 1992). The issue of capitalization on chance is a latent threat for the traditional modification indices but also for the here proposed SEM forests. Therefore, practitioners must have a strong theoretical grounding that supports the modification of the initial proposed model, especially when dealing with small samples since they have a greater risk of capitalization on chance (Kline, 2016). We expect this guide promotes the use of SEM forests as a specification search tool in SEM, helping to mitigate omitted variable bias, and helping to identify influential covariates that were initially deemed irrelevant. Or, in the case

of a large set of covariates, providing a tool to determine which ones are relevant to their models.

#### **4.7. Compliance with Ethical Standards**

Informed consent was gathered from all participants. At the time of the data collection study (2015), ethical approval was not required for survey-based research and was not a standard practice in the field of university mathematics education.

#### 4.8. References Study 2

- Arnold, M., Voelkle, M. C., & Brandmaier, A. M. (2020). Score-guided structural equation model trees. *Frontiers in Psychology, 11*, 564403.  
<https://doi.org/10.3389/fpsyg.2020.564403>
- Boker, S.M., Neale, M.C., Maes, H.H., Spiegel, M., Brick, T.R., Estabrook, R., Bates, T.C., Gore, R.J., Hunter, M.D., Pritikin, J.N., Zahery, M., & Kirkpatrick, R.M. (2023). *OpenMx: Extended Structural Equation Modelling*. (Version 2.21.11) [R].  
<https://CRAN.R-project.org/package=OpenMx>
- Brandmaier, A. M., Prindle, J. J., Arnold, M., & Van Lissa, C. J. (2024). *Semtree: Recursive Partitioning for Structural Equation Models*. (Version 0.9.20) [R].  
<https://github.com/brandmaier/semtree>
- Brandmaier, A. M., Prindle, J. J., McArdle, J. J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. *Psychological Methods, 21*(4), 566–582. <https://doi.org/10.1037/met0000090>
- Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods, 18*(1), 71–86.  
<https://doi.org/10.1037/a0030001>
- Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5–32.  
<https://doi.org/10.1023/A:1010933404324>
- Buuren, S. V., & Groothuis-Oudshoorn, K. (2011). mice: multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67.  
<https://doi.org/10.18637/jss.v045.i03>
- Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). *semTools: Useful tools for structural equation modeling*. (Version 0.5-6) [R]. Retrieved from <https://CRAN.R-project.org/package=semTools>

- Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research*, 23(1), 69–86. [https://doi.org/10.1207/s15327906mbr2301\\_4](https://doi.org/10.1207/s15327906mbr2301_4)
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford.
- Kosiol, T., Rach, S., & Ufer, S. (2019). (Which) Mathematics interest is important for a successful transition to a university study program? *International Journal of Science and Mathematics Education*, 17(7), 1359–1380. <https://doi.org/10.1007/s10763-018-9925-8>
- Lin, G. C., Wen, Z., Marsh, H. W., & Lin, H. S. (2010). Structural equation models of latent interactions: Clarification of orthogonalizing and double-mean-centering strategies. *Structural Equation Modeling*, 17(3), 374–391. <https://doi.org/10.1080/10705511.2010.488999>
- Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3), 287–296. <https://doi.org/10.1080/07350015.1988.10509663>
- MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, 100(1), 107–120. <https://doi.org/10.1037/0033-2909.100.1.107>
- MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, 111, 490–504. <https://doi.org/10.1037/0033-2909.111.3.490>
- Mooijart, A., & Satorra, A. (2009). On insensitivity of the chi-square model test to nonlinear misspecification in structural equation models. *Psychometrika*, 74(3), 443–455. <https://doi.org/10.1007/s11336-009-9112-5>



- Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. CRC Press/Taylor & Francis. <https://doi.org/10.1201/9781439800393>
- Rach, S., & Ufer, S. (2020). Which prior mathematical knowledge is necessary for study success in the university study entrance phase? Results on a new model of knowledge levels based on a reanalysis of data from existing studies. *International Journal of Research in Undergraduate Mathematics Education*, 6(3), 375–403. <https://doi.org/10.1007/s40753-020-00112-x>
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal*, 16(4), 561–582. <https://doi.org/10.1080/10705510903203433>
- Silva Díaz, J. A., Heene, M., & Brandmaier, A. M. (2024). Evaluation of structural equation model forests' performance to identify omitted influential covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, 32(2), 319–331. <https://doi.org/10.1080/10705511.2024.2417866>
- Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1), 1–21. <https://doi.org/10.1186/1471-2105-8-25>
- Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14(4), 323–348. <https://doi.org/10.1037/a0016973>

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with “well fitting” models.

*Journal of Abnormal Psychology*, 112(4), 578–598. <https://doi.org/10.1037/0021-843X.112.4.578>

Ufer, S., Rach, S., & Kosiol, T. (2017). Interest in mathematics = interest in mathematics?

What general measures of interest reflect when the object of interest changes. *ZDM*, 49(3), 397–409. <https://doi.org/10.1007/s11858-016-0828-2>

Wilms, R., Mäthner, E., Winnen, L., & Lanwehr, R. (2021). Omitted variable bias: A threat to estimating causal relationships. *Methods in Psychology*, 5, 100075.

<https://doi.org/10.1016/j.metip.2021.100075>

#### 4.9. Appendix – Study 2

##### *Model Specification Templates*

# Initial model template (model without covariates).

```
initial_model <- " sc_all1 =~ M1SA111 + M1SA112 + M1SA113 + M1SA114
                  sc_all2 =~ M2SA111 + M2SA112 + M2SA113 + M2SA114
                  m2_sw =~ M2SW1 + M2SW2
                  sc_all2 ~ a*sc_all1 # direct effect
                  m2_sw ~ b*sc_all2 + c*sc_all1
                  indirect := a*b
                  total := c + a*b"
```

# Model with identified influential covariates (in\_all1 and in\_hs1) fixed to 0.

```
model_cov_fixed <- " sc_all1 =~ M1SA111 + M1SA112 + M1SA113 + M1SA114
                    sc_all2 =~ M2SA111 + M2SA112 + M2SA113 + M2SA114
                    m2_sw =~ M2SW1 + M2SW2
                    sc_all1 ~ 0*in_hs1 + 0*in_all1
                    sc_all2 ~ a*sc_all1 # direct effect
                    m2_sw ~ b*sc_all2 + c*sc_all1 + 0*in_hs1 + 0*in_all1
                    indirect := a*b
                    total := c + a*b"
```

# Model with identified influential covariates (in\_all1 and in\_hs1) free.

```
model_cov_free <- " sc_all1 =~ M1SA111 + M1SA112 + M1SA113 + M1SA114
                   sc_all2 =~ M2SA111 + M2SA112 + M2SA113 + M2SA114
                   m2_sw =~ M2SW1 + M2SW2
```

```

sc_all1 ~ NA*in_hs1 + NA*in_all1

sc_all2 ~ a*sc_all1 # direct effect

m2_sw ~ b*sc_all2 + c*sc_all1 + Na*in_hs1 + NA*in_all1

indirect := a*b

total := c + a*b"

```

# Moderation model with covariate effects fixed to 0.

```

interaction <- indProd(ds_imp1, var1 = c("in_hs1", "in_all1"), var2 = c("M1SA114"), match =
F, meanC = T, residualC = F, doubleMC = T)

interaction2 <- indProd(interaction, var1 = c("in_hs1", "in_all1"), var2 = c("M2SW2"), match
= F, meanC = T, residualC = F, doubleMC = T)

moderation_model_fixed <- " # interaction latent factor - DOUBLE MEAN CENTERED

```

```

int1    =~ in_hs1.M1SA114 + in_all1.M1SA114

int2    =~ in_hs1.M2SW2 + in_all1.M2SW2

sc_all1 =~ M1SA111 + M1SA112 + M1SA113 + M1SA114

sc_all2 =~ M2SA111 + M2SA112 + M2SA113 + M2SA114

m2_sw =~ M2SW1 + M2SW2

sc_all1 ~ NA*in_hs1 + NA*in_all1 + 0*int1

sc_all2 ~ a*sc_all1 # direct effect

m2_sw ~ b*sc_all2 + c*sc_all1 + NA*in_hs1 + NA*in_all1 + 0*int2

indirect := a*b

total := c + a*b"

```

# Moderation model with covariate effects free.

```

interaction <- indProd(ds_imp1, var1 = c("in_hs1", "in_all1"), var2 = c("M1SA114"), match =
F, meanC = T, residualC = F, doubleMC = T)

interaction2 <- indProd(interaction, var1 = c("in_hs1", "in_all1"), var2 = c("M2SW2"), match
= F, meanC = T, residualC = F, doubleMC = T)

moderation_model <- " # interaction latent factor - DOUBLE MEAN CENTERED

int1    =~ in_hs1.M1SA114 + in_all1.M1SA114

int2    =~ in_hs1.M2SW2 + in_all1.M2SW2

sc_all1 =~ M1SA111 + M1SA112 + M1SA113 + M1SA114

sc_all2 =~ M2SA111 + M2SA112 + M2SA113 + M2SA114

m2_sw   =~ M2SW1 + M2SW2

sc_all1 ~ NA*in_hs1 + NA*in_all1 + NA*int1

sc_all2 ~ a*sc_all1 # direct effect

m2_sw   ~ b*sc_all2 + c*sc_all1 + NA*in_hs1 + NA*in_all1 + NA*int2

indirect := a*b

total := c + a*b"

```

## 5. General Conclusions

This dissertation introduces SEM forests (Brandmaier et al., 2016) as a novel technique for conducting specification search in SEM related to omitted influential covariates. The first study employs data simulation to evaluate the performance of SEM forests in accurately identifying influential omitted covariates while minimizing the misidentification of noninfluential covariates. The second study included in this dissertation presents a practical guide for utilizing the R package *semtree* to perform specification searches with SEM forests, complemented by a real-world data example.

From the first study we conclude that SEM forests are sensitive to omitted influential covariates and consistently yield unbiased importance scores close to zero for noninfluential covariates. The performance of SEM forests to correctly identify influential covariates improves with larger covariate paths, stronger factor loadings, and larger sample sizes, while the probability of incorrectly identifying noninfluential covariate paths remains nearly zero across all conditions. Consequently, practitioners can reasonably assume that covariates identified by SEM forests are unlikely to be false positives when their sample sizes, factor loadings, or covariate-latent variable paths are comparable to those examined in this study ( $N = 200, 500, 1000$ ;  $\lambda = .4, .6, .8$ ;  $\beta = 0, .2, .35, .5$ ). Moreover, SEM forests are more effective at identifying omitted influential covariates related to multiple latent variables, compared to covariates with unique paths. The primary limitation of SEM forests lies in their potential lack of power to detect genuinely influential covariates in models with small sample sizes. Moreover, the detection of interaction effects, although feasible, remains challenging for SEM forests under the conditions examined, highlighting the need for further investigation. This is a critical area of inquiry, as interaction effects are prevalent in the social sciences but are often overlooked in specification search due to methodological complexities (Cortina et al., 2021).

The second study provides comprehensive guidelines on best practices and recommended SEM forests parameters to achieve reliable results with power and type I error rates comparable to those reported in Study 1. This involves a careful data preparation, including handling missing values, and appropriately configuring the SEM forests' sampling method, split selection, number of trees, and Bonferroni correction, among others. The practical example used to illustrate the specification search demonstrates how SEM forests successfully identified two influential covariates initially omitted from the model, despite the relatively small sample size of the analyzed dataset. Including these covariates improved the model goodness-of-fit, underscoring SEM forests' advantage over traditional modification indices, which can indicate unmodeled variance in variables already included in the model, but cannot identify influential parameters of omitted variables. Failure to account for such variables can result in biased model parameters and fit indices due to unaccounted variable correlations (Kline, 2016; Wilms et al., 2021).

SEM forests cannot specify where or how to include detected omitted covariates or interactions. Partial dependence plots are suggested to guide the inclusion of identified covariates, exploring the impact of identified covariates in specific model parameters, although they are primarily informative with continuous variables. Additionally, SEM forests cannot identify unmeasured influential covariates. Alternatives like mixture multigroup factor analysis (MMG-FA) can be more effective in this regard (De Roover, 2021; De Roover et al., 2022). Practitioners should also consider that SEM forests calculate marginal importance, which can be biased with correlated predictors (Strobl et al., 2008). Future studies should also examine the impact of higher predictor correlations and nonlinear relationships in SEM forests performance.

The performance of the respecified models was measured using a combination of approximate fit indices (i.e., CFI, RMSEA, SRMR), chi-square test, and BIC and AIC. At

present, no single model fit evaluation strategy is considered optimal, as each one has limitations. The selected strategy model fit evaluation strategy may indicate acceptable model fit while still being vulnerable to issues such as model equivalence or sensitivity to omitted variables (Kline, 2016). Tomarken and Waller (2003) observe that, although fit indices are somewhat responsive to the presence of omitted variables, they do not capture all types of omitted variable structures. Thus, a model can exhibit good fit even when important predictors are missing. This is especially relevant to SEM forests specification search, since its main advantage is precisely to identify omitted influential predictors. That is, a modified model that includes omitted influential covariates identified by SEM forests and that shows acceptable goodness-of-fit may still have omitted covariates that could enhance the model, whose effects are not detected by fit indices. Residual variances and covariances can provide clues for detecting omitted variables or paths, and sensitivity analysis has been proposed to assess possible bias due to omitted variables (Tomarken & Waller, 2003).

Overall, SEM forests provide a robust methodological approach for identifying omitted influential covariates and addressing omitted variable bias, particularly in models with numerous potentially influential covariates. It is crucial to emphasize that any specification search strategy, including modification indices and SEM forests, must ensure a strong theoretical foundation and implement rigorous validation procedures to mitigate the risks of overfitting and unreliable conclusions. In this sense, validation techniques such as cross-validation, split-sample validation, and bootstrapping are recommended to assess the stability of parameter estimates and minimize the risk of findings driven by sample-specific variations rather than underlying structural relationships.



## 6. References for the General Introduction and Conclusions

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- Arnold, M., Voelkle, M. C., & Brandmaier, A. M. (2020). Score-guided structural equation model trees. *Frontiers in Psychology*, 11, 564403. <https://doi.org/10.3389/fpsyg.2020.564403>
- 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>
- Bentler, P. M. (1995). *EQS structural equations program manual*. Multivariate Software.
- Bollen, K. A., Fisher, Z., Lilly, A., Brehm, C., Luo, L., Martinez, A., & Ye, A. (2022). Fifty years of structural equation modeling: A history of generalization, unification, and diffusion. *Social Science Research*, 107, 102769. <https://doi.org/10.1016/j.ssresearch.2022.102769>
- Brandmaier, A. M., Prindle, J. J., Arnold, M., & Van Lissa, C. J. (2023). Semtree: Recursive Partitioning for Structural Equation Models. R package version 0.9.19. <https://github.com/brandmaier/semtree>
- Brandmaier, A. M., Prindle, J. J., McArdle, J. J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. *Psychological Methods*, 21, 566–582. <https://doi.org/10.1037/met0000090>
- Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, 18, 71–86. <https://doi.org/10.1037/a0030001>
- 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>

- Cortina, J. M., Markell-Goldstein, H. M., Green, J. P., & Chang, Y. (2021). How are we testing interactions in latent variable models? Surging forward or fighting shy? *Organizational Research Methods*, 24, 26–54.  
<https://doi.org/10.1177/1094428119872531>
- De Roover, K. (2021). Finding clusters of groups with measurement invariance: Unraveling intercept non-invariance with mixture multigroup factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 28, 663–683. <https://doi.org/10.1080/10705511.2020.1866577>
- De Roover, K., Vermunt, J. K., & Ceulemans, E. (2022). Mixture multigroup factor analysis for unraveling factor loading noninvariance across many groups. *Psychological Methods*, 27, 281–306. <https://doi.org/10.1037/met0000355>
- Green, S. B., & Thompson, M. S. (2010). Can specification searches be useful for hypothesis generation? *Journal of Modern Applied Statistical Methods*, 9(1), 160–171.  
<https://doi.org/10.22237/jmasm/1272687300>
- Hancock, G. R., & Mueller, R. O. (Eds.). (2013). *Structural equation modeling: A second course*. (2nd ed.). IAP.
- 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.  
<https://doi.org/10.1037/a0024917>
- Hershberger, S. L. (2003). The growth of structural equation modeling: 1994-2001. *Structural Equation Modeling*, 10(1), 35–46. [https://doi.org/10.1207/S15328007SEM1001\\_2](https://doi.org/10.1207/S15328007SEM1001_2)
- Holbert, R. L., & Stephenson, M. T. (2002). Structural equation modeling in the communication sciences, 1995–2000. *Human Communication Research*, 28(4), 531–551. <https://doi.org/10.1111/j.1468-2958.2002.tb00822.x>

- 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>
- Jöreskog, K. G. (1970). A general method for estimating a linear structural equation system. *Biometrika*, 57(2), 239–251. <https://doi.org/10.1002/j.2333-8504.1970.tb00783.x>
- 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](https://doi.org/10.1207/S15328007SEM1003_1)
- Kline, R. B. (2016). *Principles and practice of structural equation modeling*. (4th ed.). Guilford.
- Lei, P. W., & Wu, Q. (2007). Introduction to structural equation modeling: Issues and practical considerations. *Educational Measurement: Issues and Practice*, 26(3), 33–43. <https://doi.org/10.1111/j.1745-3992.2007.00099.x>
- MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological bulletin*, 100(1), 107–120. <https://doi.org/10.1037/0033-2909.100.1.107>
- MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, 111(3), 490–504. <https://doi.org/10.1037/0033-2909.111.3.490>
- Marcoulides, G. A., Drezner, Z., & Schumacker, R. E. (1998). Model specification searches in structural equation modeling using tabu search. *Structural Equation Modeling: A Multidisciplinary Journal*, 5(4), 365–376. <https://doi.org/10.1080/10705519809540112>
- Marcoulides, K. M., & Falk, C. F. (2018). Model specification searches in structural equation modeling with R. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(3), 484–491. <https://doi.org/10.1080/10705511.2017.1409074>

- Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58, 415–434.  
<https://doi.org/10.2307/2283276>
- Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. Chapman and Hall/CRC. <https://doi.org/10.1201/9781439800393>
- Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods*, 17(1), 1–14.  
<https://doi.org/10.1037/a0026804>
- 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>
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464. <https://doi.org/10.1214/aos/1176344136>
- Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections: Further analysis of the data by Akaike's. *Communications in Statistics-theory and Methods*, 7(1), 13–26.  
<https://doi.org/10.1080/03610927808827599>
- Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9, 307.  
<https://doi.org/10.1186/1471-2105-9-307>
- Tomarken, A. J., & Waller, N. G. (2003). Potential problems with “well fitting” models. *Journal of Abnormal Psychology*, 112, 578–598. <https://doi.org/10.1037/0021-843X.112.4.578>
- Van Voorhis, C. W., & Morgan, B. L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology*, 3, 43–50. <https://doi.org/10.20982/tqmp.03.2.p043>

Wilms, R., Mäthner, E., Winnen, L., & Lanwehr, R. (2021). Omitted variable bias: A threat to estimating causal relationships. *Methods in Psychology*, 5, 100075.

<https://doi.org/10.1016/j.metip.2021.100075>

## 7. List of Figures

Figure 1 – Study 1 .....	19
Figure 2 – Study 1 .....	25
Figure 3 – Study 1 .....	29
Figure 4 - Study 1 .....	34
Appendix A – Study 1 .....	47
Appendix B – Study 1 .....	478
Figure 1 – Study 2 .....	56
Figure 2 – Study 2 .....	59
Figure 3 – Study 2 .....	61
Figure 4 – Study 2 .....	64
Figure 5 – Study 2 .....	65
Appendix – Study 2 .....	76

**8. List of Tables**

Table 1 – Study 1 .....	30
Table 2 – Study 1 .....	33
Table 1 – Study 2 .....	67
Table 2 – Study 2 .....	67
Table 3 – Study 2 .....	68