Modifying the Wide Format Approach To Multilevel Structural Equation Modeling To Mitigate Estimation Problems

Dissertation

zur Erlangung des Doktorgrades
der Wirtschafts- und Sozialwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen

vorgelegt von
M.Sc. Julia-Kim Walther

Tübingen
2025

1. Betreuer: Prof. Dr. Steffen Zitzmann

2. Betreuer: Prof Dr. Martin Hecht

3. Betreuer: Prof. Dr. Benjamin Nagengast

Tag der mündlichen Prüfung: 26.05.2025

Dekan:in: Prof. Dr. Taiga Brahm, Prof Dr. Dominik Papies

1. Gutachter: Prof. Dr. Steffen Zitzmann

2. Gutachter: Prof. Dr. Martin Biewen

The fool
who persists
in [their] folly
will become wise.
— William Blake

To whom it may concern

Acknowledgements

To thank selected people would reflect nothing but my subjective theory about who has made an impact on the thesis at hand and Bayesian statistics was merely a side note in here. Thus, I simply thank anyone who considers themselves to have had an impact, and similarly anyone who does not think so, in the style of how programming tutorials start – "Thank you, world!".

Berlin, June 23, 2025 JK

Abstract

In psychology and the education sciences, multilevel data is omnipresent. In this data structure observational units are nested within higher level units, such as students within classes. A powerful tool for estimating parameters across these different levels is multilevel structural equation modeling (SEM). Multilevel data can be arranged in two data formats, long (LF) and wide (WF) format, and for both, multilevel SEM approaches are available. While both approaches are applicable within the established *lavaan* package in the free and open-source software *R*, the major advantage of the WF approach is its modifiability. Within my thesis, I first (1) redefine „small samples“ for multilevel data and compare the performance of both approaches here. I then show that the WF approach can be modified to mitigate estimation problems such as non-convergence and inaccuracy under conditions of (2) small samples and (3) heterogeneous variances. Findings suggest that both approaches exhibit comparable performance in small samples and that the proposed extensions of the WF approach offer an accessible avenue for researchers dealing with multilevel analysis, particularly when data acquisition is limited or heterogeneous populations are under investigation.

Zusammenfassung

Forschungsfelder wie Psychologie und Erziehungswissenschaften widmen sich der Beobachtung, Erklärung und Vorhersage menschlichen Verhaltens. Das erworbene Wissen fließt in politische Entscheidungsprozesse ein, beispielsweise in die Frage, wie Schulen organisiert werden sollten, damit Schüler:innen ihr Potenzial optimal entfalten können. Ein solch zentrales Ziel mit weitreichenden Konsequenzen benötigt solide Methodik um Entscheidungen auf Basis einer starken empirischen Grundlage treffen zu können.

Die erhobenen Daten in diesen Feldern sind oft Mehrebenendaten. Das heißt, sie weisen eine „hierarchische“ Struktur auf, in der niedrigere Ebenen in höhere Ebenen eingebettet sind, wie zum Beispiel Schüler:innen in Schulen. Wird die Abhängigkeit in diesen Mehrebenendaten nicht angemessen berücksichtigt, können Effekte zwischen den Ebenen nicht isoliert werden und Parameter werden verzerrt geschätzt (z.B., Clarke, 2008; Julian, 2001). Ein mächtiges Analyse-Tool welches diese Abhängigkeit berücksichtigt ist Mehrebenen-Strukturgleichungsmodellierung. Diese ermöglicht die separate Schätzung von Effekten auf den verschiedenen Ebenen sowie die Modellierung von Messfehlern und komplexen Beziehungen zwischen latenten und beobachteten Variablen. Mehrebenendaten können entweder im „langen“ oder „weiten“ Format angeordnet werden, und für beide Formate existieren Ansätze für Mehrebenen-Strukturgleichungsmodellierung. Letzterer lässt sich jedoch in der kostenlosen und quelloffenen Software *R* besser modifizieren. Daher beschäftigt sich die vorliegende Arbeit mit der Validierung und Erweiterung des Ansatzes im langen Format zur Mehrebenen-Strukturgleichungsmodellierung im *lavaan* Packet in *R* im Sinne der Prinzipien der Open Science Bewegung.

In meinen drei Dissertationsprojekten definiere ich zunächst (1) den Begriff „kleine Stichproben“ in Mehrebenendaten genauer und zeige, dass beide Ansätze zur Mehrebenen-Strukturgleichungsmodellierung auch unter diesen Bedingungen äquivalent sind. Anschließend erweitere ich den Ansatz im weiten Format um zwei regelmäßig auftretende Schätzprobleme, nicht konvergierende und ungenaue Modelle, die häufig unter Bedingungen von (2) kleinen Stichproben und (3) heterogenen Varianzen auftreten, abzumildern. Die Ergebnisse legen nahe, dass die vorgeschlagenen Erweiterungen eine zugängliche Möglichkeit für Forscher:innen bietet die Mehrebenenanalysen durchführen, insbesondere wenn die Datenerhebung begrenzt ist oder heterogene Populationen untersucht werden.

Contents

Introduction

Professions such as psychology and the education sciences are dedicated to observing, explaining and predicting human behaviour. The acquired knowledge informs decision making in politics such as how schools should be organized to help students reach their potential. For such a pivotal purpose with lasting impact, sound methodology is a prerequisite for basing decisions on strong empirical foundations.

Observational data in these fields often has a "hierarchical" structure in which lower level units are nested within higher level units such as students within classrooms. If the dependence in this multilevel data is not adequately accounted for, conflation of effects across levels and biased parameter estimates might be the result (e.g. Clarke, 2008; Julian, 2001). A powerful tool to account for the dependence is multilevel structural equation modeling (SEM): it allows to estimate effects at different levels separately, account for measurement error, and model complex relationships among latent and observed variables. Multilevel data can be arranged in either long (LF) and wide (WF) format, and for both, multilevel SEM approaches exist.

A recent empirical investigation on common data analysis practices revealed that only 21% of studies published in APA journals used analysis approaches that accounted for the multilevel nature of the data, and of these, only 4% used multilevel SEM (Blanca et al., 2018). Even more so, only a marginal fraction indicated having used the free and open-source software *R* (3%) whereas a large majority (61%) used the proprietary software M*plus* (Blanca et al., 2018). Although the latter is undeniably the more versatile and powerful SEM software, the accessibility and transparency of *R* is superior. In its established SEM package *lavaan* (Rosseel, 2012), both multilevel SEM approaches can be utilised but the WF approach has more options to be modified. Thus, the present thesis revolves around validating and extending the WF approach to multilevel SEM in the *lavaan* package in *R* in the spirit of Open Science principles.

Within the three research projects within my thesis, I first (1) define "small samples" in multilevel data more precisely and validate that both the LF and WF approaches are equivalent in these. I then extend the WF approach to mitigate two common estimation problems, namely non-convergence and inaccuracy, under conditions of (2) small samples and (3) heterogeneous variances. In the subsequent chapters, I will elaborate on multilevel modeling, small samples in multilevel data, the application of regularization in these settings, and heterogeneous variances and how my research endeavours are embedded within these.

1 Multilevel Modeling

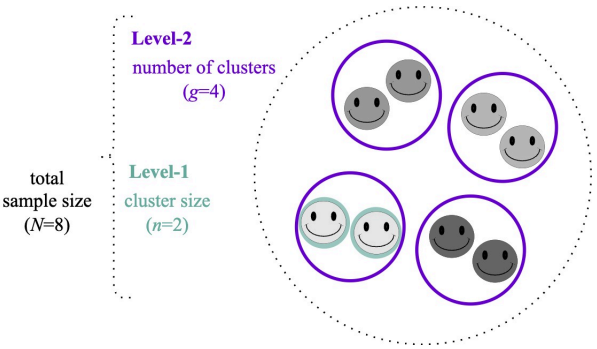
A large body of observational data in psychology and the educational sciences has a so-called "hierarchical" structure. Other terms frequently found in the literature are "clustered", "nested", or "multilevel". As the term hierarchy implies, the data has different levels. Lower level units (e.g., level-1) are nested within higher level units (e.g., level-2), such as students nested within classrooms or clients nested within therapists. These are examples of naturally occurring clusters. However, clustering may also be introduced by research designs such as interventions with experimental and control groups and a pre-post design. Here, time points are nested within participants which are nested within experimental and control groups.

The existence of such data structures is neither accidental nor ignorable. Clustering induces unobserved heterogeneity, which signifies that units across clusters are less similar to each other than units within clusters. For example, students within a classroom share the same influences, such as teachers, their teaching practices, and the students interact with each other. Clients of a therapist share the practice of the therapist but they usually cannot interact with each other unless it is a group therapy setting. Subjects in experimental or control groups may have little in common other than receiving the same intervention (if the randomisation has been done thoughtfully). Thus, while the level of clustering might differ, clusters tend to become differentiated, and the resulting dependencies ought to be considered in analysis.

Traditional single-level methodology, such as ordinary least squares (OLS) regression, does not account for the dependencies in the data. This dependency gives rise to two core problems which B. O. Muthén (1994) classifies into sampling and varying parameters. Hierarchical data is obtained by cluster sampling such as selecting classrooms and then students within these classrooms. Thus, the assumption of independent and identically distributed observations does not hold. Ignoring the clustering and analysing the level-1 units leads to conflated effects across levels and biased parameter estimates (e.g. Clarke, 2008; Julian, 2001). Evidence suggests that this is even the case when there are as few as two observations in each cluster (Clarke, 2008). In addition, measurement models might suffer from upward biased (co)variance estimates, such as factor (co)variances and residual variances (Julian, 2001). The overestimation of level-1 (co)variances is plausible when considering its conflation with the unaccounted

level-2 (co)variances. The standard errors of these level-1 coefficients, estimated by dividing their variability through the sample size, are then underestimated. Appropriate analysis needs to specify stochastic variation that accounts for the cluster sampling, such as formulating a model that decomposes the variation of an outcome into variation within and between clusters, commonly referred to as within-between decomposition. Combining the evidence, when researchers are merely interested in level-1 (e.g. students) effects, the number of clusters is sufficiently large and the cluster size is balanced, then correcting standard errors may suffice to deal with the sampling problem (Angrist & Pischke, 2009; Kreft & Yoon, 1994; McNabb & Murayama, 2021). Nevertheless, this cannot account for the second drawback, namely that parameter variability across clusters (e.g., classrooms), or put differently, heterogeneity, remains unmodeled. In certain research fields, however, acknowledging varying parameters is fundamental. For instance, in educational effectiveness research, individual student-level variables (e.g., motivation, cognitive ability) and contextual classroom-level variables (e.g., teaching practices, school resources) are considered when modeling outcomes of interest (e.g., academic success). It takes into account that "how much a student learns depends on the identity of the classroom to which that student is assigned" (p. 320, Monk, 1992). To address both problems simultaneously, multilevel modeling (MLM), also known as hierarchical linear modeling (HLM), linear mixed-effect modeling (LMM) or random effects modeling, has been developed in the context of educational research (see e.g., Aitkin & Longford, 1986; Goldstein & McDonald, 1988; Goldstein et al., 1993).

Figure 1.1: Two-Level Data



Although manifold levels of data are conceivable, the focus within the present thesis is on two levels, because two-level models are the most frequently used multilevel models in practice (Dedrick et al., 2009). In Figure 1.1, a conceptual example of two-level data is depicted. At level-2, four units (e.g., classrooms) are sampled, $g = 4$. The units at this level are called cluster. At level-1, two units (e.g., students) are observed within each level-2 unit, $n = 2$. This is often referred to as the cluster size. In total, eight level-1 units are observed, $N = 8$. Next, we will consider how relations in this data can be modelled within the general multilevel modeling framework.

1.1 The Multilevel Modeling Framework

To begin with, let us revisit a simple single-level linear regression model in which only level-1 is modelled. It reads:

$$Y_i = \beta_0 + \beta_1 X_i + e_i \tag{1.1}$$

where the outcome Y_i of individual i is modelled as a function of the predictor X_i, for which β_0 is the intercept and β_1 is the slope. The intercept is defined as the conditional mean of the outcome when the value of the predictor is 0. The slope is the change in outcome depending on the value of the predictor. e_i denotes the residual error of individual i which is the deviation between values in the outcome that are observed and those that are predicted by the model. Residual errors are assumed to be normally distributed with a mean of zero and variance σ_{2e}: $e \sim N(0, \sigma_{2e})$.

When taking into account that individuals are nested within clusters, then regression coefficients, such as intercept and slope, are allowed to vary by cluster. The single-level regression model above becomes a random intercept and slope model with one level-1 predictor. The equation at level-1 reads:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij} \tag{1.2}$$

where Y_{ij} is the outcome for individual i in cluster j, β_{0j} and β_{1j} are the intercept and slope of predictor X_{ij} in cluster j, respectively, and e_{ij} is the residual error which is (still) assumed to be normally distributed with a mean of zero and variance σ_{2e}: $e \sim N(0, \sigma_{2e})$. As the intercept and slope of the predictor vary by cluster at level-2, their equations are as follows:

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{1.3}$$

where γ_{00} is the average intercept across clusters (i.e., the grand mean of the scores of all individuals across all clusters when the predictor equals 0) and u_{0j} is the deviation from cluster j to the average intercept. The latter is assumed to be normally distributed with a mean of μ_{u_0} and variance σ_{2u_0}: $u_{0j} \sim N(\mu_{u_0}, \sigma_{2u_0})$. Similarly:

$$\beta_{1j} = \gamma_{10} + u_{1j} \tag{1.4}$$

where where γ_{10} is the average slope across clusters (i.e., the grand mean of the scores of all individuals across all clusters when the predictor equals 0) and u_{1j} is the deviation from cluster j to the average slope. The deviation is assumed to be normally distributed with a

mean of μ_{u_1} and variance σ_{2u_1}: $u_{1j} \sim N(\mu_{u_1}, \sigma_{2u_1})$.

In many textbooks, the presented effects are differentiated as fixed and random effects. Both γ_{00} and γ_{10} can be thought of as fixed effects because they remain constant across all clusters. In contrast, u_{0j} and u_{1j} vary from cluster to cluster. They are thus considered random effects. A model which incorporates both types of effects is often called random coefficient or mixed effects model. Note, however, that the terms fixed and random effects are neither solely tied to MLMs nor unambiguous in definition. For instance, they also denote two classes of modeling frameworks that encompass several approaches with different assumptions and methods, respectively (for an overview see e.g., Townsend et al., 2013). I will use these terms throughout the present work but only in relation to MLMs.

Whether is is necessary to use the MLM framework for analysis depends on the empirical level of clustering. It is suggested to estimate a random intercept-only model (i.e., an unconditional model without a predictor) in order to decompose the variance of the outcome into variance "within" (i.e., level-1) and "between" (i.e., level-2) clusters[1]. By means of this within-between framework, the level of clustering can be assessed empirically by the Intraclass Correlation (ICC). The ICC is defined as the between variance (σ_{2u_0}) divided through the total variance (i.e., sum of the between and within variance, $\sigma_{2u_0} + \sigma_{2e}$ J. J. Hox et al., 2017). ICC values range from zero to one in which small values indicate that units within clusters are not very similar to each other and large values claim that they are similar. In psychology and the education sciences, ICCs between 0.05 and 0.25 are frequently observed (Adams et al., 2004; Gulliford et al., 1999). Some researchers consider a non-zero ICC as sufficient evidence for a substantial level of clustering which justifies the application of the MLM framework for analysis. Other researchers argue that not only the ICC but the design effect, which is influenced by the ICC and the cluster size and which quantifies the effect of independence violations on standard error estimates, should be considered when deciding whether do use MLM (see Peugh, 2010). Regardless of which guidelines are used to justify using or not using a MLM framework, the level of clustering in hierarchical data ought always be examined empirically.

The general MLM framework of a two-level random intercept and slope model as presented above (see Equation 1.2, Equation 1.3 and Equation 1.4) might be extended in several ways. To name a few, firstly, cross-level interactions (see e.g., Snijders & Bosker, 2012) can be included. The idea behind these is that the level-2 variation of intercept and slope of level-1 predictors might be explained by level-2 variables. For instance, differences across classrooms in the effect of motivation on academic success might be related to teacher characteristics. Relating level-1 and level-2 parameters, instead of just modeling both simultaneously, is a core strength of MLM. Secondly, more than two levels can be modelled. For instance, when students are nested in classrooms nested within schools, then a three-level analysis might be considered (see e.g., Raudenbush & Bryk, 2002). However, when the number of units or the variance at

[1] For clarification, note that when referring to levels of *parameters*, the terms within- and between-cluster variables may be used but not when levels of *sample sizes* are concerned. Here, the terms level-1 and level-2 units, or cluster size and number of clusters, are prevalent.

the higher level(s) (including level-2) is small, then convergence and accuracy of higher level parameter estimates of the model are threatened (D. Hox & McNeish, 2020; J. J. Hox & Maas, 2001; Lüdtke et al., 2008, 2011; Muthen & Satorra, 1995; Stegmueller, 2013; Zitzmann, 2018; Zitzmann et al., 2015). Alternatively, level-3 (e.g., schools) can to be modelled as categorical predictor in two-level model. P. W. Hill and Rowe (1996) provides an empirical review and guidelines for more than two-level MLMs in the educational sciences. Thirdly, the application of MLMs can been extended to model non-linear relationships between variables such as log linear models for discrete outcomes (see e.g., Goldstein, 1991; Rabe-Hesketh & Skrondal, 2008). For instance, one might be interested in whether students pass a test (or not) depending on the motivation of the students and whether there are differences between classrooms. More extensions, such as Bayesian estimation and latent class models, can be found, for instance in Snijders and Bosker (2012).

Over the course of the last two decades, MLM has become a standard approach in the analysis of hierarchical data (Goldstein, 2011). However, certain limitations of the general MLM framework remain. Firstly, most MLMs are restricted to model univariate outcomes. Thus, complex relations between variables, such as mediator and moderator relations, cannot be modelled. Secondly, modeling of measurement error is confined to a subset of MLMs, particularly Bayesian ones (for an overview see Goldstein et al., 2008). Within psychological research, where most phenomena of interest are non-observable, latent variables that are tied by complex relations, this, however, restricts the general application of MLMs and the reliability of the results obtained. Fortunately, structural equation models (SEM) incorporate structural and measurement models, and their combination, multilevel SEMs (see e.g. the work of Bryk & Raudenbush, 1987; McDonald, 1993; McDonald & Goldstein, 1989; B. O. Muthén, 1989, 1990, 1994), represent a considerable advance over conventional MLMs.

1.2 Multilevel Structural Equation Modeling

Let us consider the measurement models first. The outcome Y_{ij} is modelled by latent variables at the between- and within-cluster levels. A general formulation of a latent factor model for two-level data reads:

$$Y_{ij} = \underbrace{\upsilon + \Delta_B \eta_{Bj} + \epsilon_{Bj}}_{\text{Between-Cluster}} + \underbrace{\Delta_W \eta_{Wij} + \epsilon_{Wij}}_{\text{Within-Cluster}} \tag{1.5}$$

where a conventional factor analysis, $\Delta\eta + \epsilon$, is extended to both the between-cluster (B) and within-cluster (W) level, respectively, where Δ denotes the factor loadings, η represents the latent factors and ϵ the measurement errors. The measurement intercepts υ belong only to the between part as these are modeled by the cluster means.

In the structural models relations can be modelled at the between- and within-cluster level similarly, see Equation 1.6 and Equation 1.7:

$$\eta_{Bj} = \alpha + B_B \eta_{Bj} + \zeta_{Bj} \tag{1.6}$$

$$\eta_{Wij} = B_W \eta_{Wij} + \zeta_{Wij} \tag{1.7}$$

where B are regression weights (with zero diagonals elements) and ζ are the residual errors. The factor means α, analogously to the measurement intercepts, are only present at the between-cluster level. Extensions may include adding contextual variables z and allowing for varying factor means α_j and measurement intercepts ν_j across clusters (see e.g., B. O. Muthén, 1990; B. O. Muthén & Satorra, 1989). Moreover, complex relations, such as mediator and moderator relationships, even across levels, may be modelled (see e.g., Preacher et al., 2016) and models may be extended to three levels (see e.g., B. O. Muthén, 1994; Preacher, 2011).

To estimate the afore mentioned model parameters, the total variance of the outcome Y_{ij} is decomposed into between- and within-cluster components:

$$V(Y_{ij}) = \Sigma_T = \Sigma_B + \Sigma_W \tag{1.8}$$

Traditionally, a multilevel SEM is analyzed in the 'long format' (LF) approach as described by several authors (see e.g., Bryk & Raudenbush, 1987; McDonald, 1993; McDonald & Goldstein, 1989; B. O. Muthén, 1989, 1990, 1994). In LF, level-1 units are represented by separate rows. Thus, it is sometimes called 'univariate' approach. However, a multilevel SEM in the LF approach can be estimated as a single-level restricted confirmatory factor analysis (CFA; see Bryk & Raudenbush, 1987; Chou et al., 1998; MacCallum et al., 1997; McArdle & Epstein, 1987; Mehta & Neale, 2005; Meredith & Tisak, 1990). In the 'wide format' (WF) or 'multivariate' approach, level-1 units are represented by separate columns. Each row is independent and corresponds to level-2 unit (i.e., classroom). Each observed variable is split times the cluster size, or as Mehta and Neale (2005) put it "people are variables, too". This way, univariate multilevel models (MLM) become multivariate single–level models (CFA). Relations between these two classes of models have also been discussed in univariate and multivariate SEM formulations of latent growth curve (MacCallum et al., 1997; Willett & Sayer, 1994), explicit MLM models (Rovine & Molenaar, 2000), and even in more complex MLMs such as hierarchical factor models (Bauer, 2003; Curran, 2003) and the non-linear MLMs mentioned earlier which have been shown to be formally equivalent with item response theory models (Rijmen et al., 2003). Barendse and Rosseel (2020) described a general SEM framework for using the WF approach for both continuous and discrete data and gave an in-depth tutorial on how to apply the approach in *lavaan*. Past research demonstrated analytical and empirical equivalence of the LF and WF approaches with regard to convergence and estimation accuracy (Barendse & Rosseel, 2020; Mehta & Neale, 2005). Moreover, the WF approach has been praised for its versatility. For instance, it has been used to integrate out nuisance parameters, more

specifically, level-1 parameters when level-2 parameters were of primary interest, in order to save run time (Hecht et al., 2020). More recently, Barendse and Rosseel (2023) proposed to fit multilevel models with many latent variables, whose estimation often fails with most frequentist estimation methods, in the WF approach with pairwise maximum likelihood estimation.

In the following, the differences between the LF and the WF approach with regard to data matrix, covariance matrices, and model specification will be illustrated. For this, a small example data set which consists of five clusters ($g = 5$), with two units in each cluster ($n = 2$) is presented. For every unit two continuous variables ($p = 2$), x_1 and x_2[2], were observed, which were aggregated in order to obtain between-cluster variables. These kind of variables are often referred to as "contextual" variables (see e.g., Boyd & Iversen, 1979; Raudenbush & Bryk, 2002).

1.2.1 The Long and Wide Format Approaches

Data Matrix

When arranging the data set in long format (LF), the total data matrix (LF-T) contains p columns and $N = g \cdot n$ rows. For the decomposition of the total variance of each contextual variable into within and between parts, the total data matrix (LF-T) is split into the between-cluster data matrix (LF-B) and the within-cluster data matrix (LF-W). To achieve this, the cluster means are estimated and subtracted from the observed values. These cluster means constitute the between-cluster data matrix LF-B with p columns and g rows. The deviations from these cluster means constitute the within-cluster data matrix LF-W with p columns and $g \cdot n = N$ rows. All three LF data matrices can be found on the left side in Figure 1.2.

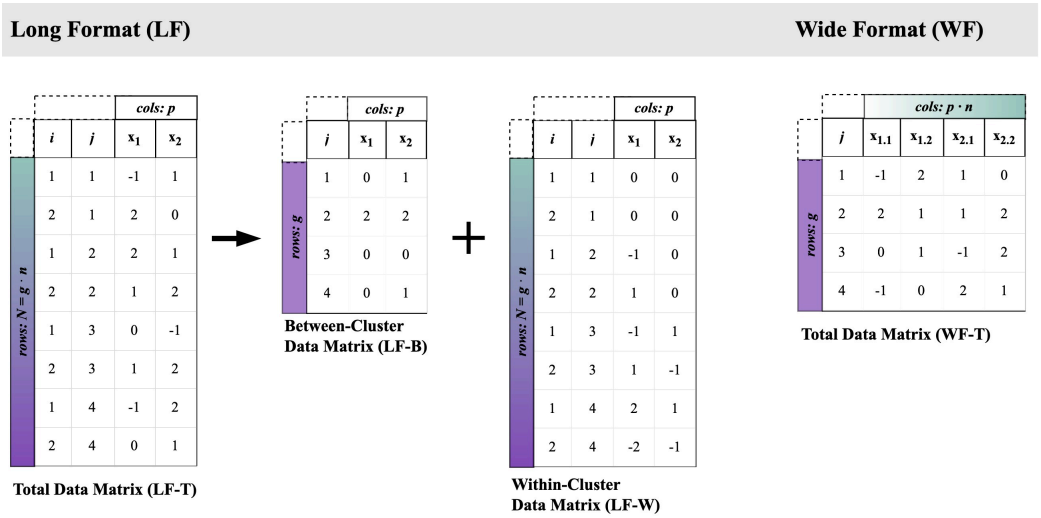
In contrast, when arranging the data set in the wide format (WF), the total data matrix (WF-T) consists of $p \cdot n$ columns and g rows. No decomposition of the within and between parts of the total variance of contextual variables is taking place. However, each observed variable p is split into n new variables. In this respect, recall how Mehta and Neale (p.1, 2005) put it: "people [n] are variables too". Thus, WF-T is a single-level represented two-level data matrix, which can be found to the right in Figure 1.2.

When contrasting the LF and WF data matrices further, one important distinction has to be emphasized. Whereas every observed variable p in both LF-W and LF-B includes information on *all* N units, the $p \cdot n$ "observed" variables in WF-T only contain information on *specific* g units. Put differently, for instance, $x_{1.1}$ is the observed variable x_1 for every 1st unit in the cluster ($i = 1$) for which one unit from every cluster, and five units in total ($g = 5$), have been observed. This is why, I refer to the former as *all-units* and to the latter as *specific-units* variable. Note that Barendse and Rosseel (2020) used the term *unit-specific* with respect to the WF variables, but to avoid confusing the term with person-specific effects, I do not use their

[2] Note that in the following, a random intercept-only model for x_1 and x_2 is discussed. The random intercept and slope model for y and x in this section has been introduced for reasons of comprehensiveness.

terminology.

Figure 1.2: Data Matrices in Long Format (LF) and Wide Format (WF)



g = number of clusters (where $j = 1, ..., g$), n = cluster size (where $i = 1, ..., n$), N = total sample size, p = number of observed variables. Example data set with $g = 4$, $n = 2$, $N = 8$, and $p = 2$.

There is one important difference in the applicability of the approaches in *lavaan* that has to be noted. The implementation of traditional maximum likelihood estimation (MLE) requires that the sample covariance matrix is positive definite (Hamaker et al., 2003; H. Singer, 2010; Van Montfort et al., 2018; Voelkle et al., 2012). This, amongst other things, necessitates that the number of columns is less than or equal to the number of rows in the supplied data matrix (LF-T or WF-T, respectively). Otherwise, at least one sample eigenvalue would become zero and the sample covariance matrix would turn non-positive definite (e.g., Duncan et al., 1997; Gorsuch, 1983b; Wothke, 1993). For instance, the example data set used in the figures throughout this chapter contains the minimal number of clusters required in WF-T given the number of observed variables and the cluster size (or, in other words, the number of columns equals the number of rows). Alternatively, the raw data formulation of MLE, full information maximum likelihood (FIML), may be used, which circumvents the problem (Hamaker et al., 2003; Trendafilov & Unkel, 2011; Unkel & Trendafilov, 2010; Voelkle et al., 2012). As the LF data matrices are longer than the WF data matrix this implies that there are samples which can only be used in the LF approach. As the number of columns increases substantially with large cluster sizes in the WF approach (i.e., number of columns equals $p \cdot n$), it is rather suited for data sets with smaller cluster sizes. However, as we will later see, one advantage of the WF approach is that in single-level SEM, a sample covariance matrix instead of a data matrix might be supplied. This cannot be done in multilevel SEM (i.e., the LF approach) so far in *lavaan* version 0.6-15.

Sample Covariance Matrix

In LF, one sample covariance matrix is estimated for each level off LF-W and LF-B. The MLEs for the sample covariance matrices of both levels, the between-cluster estimator in Equation 1.9 and the pooled within-cluster estimator in Equation 1.10, read:

$$\mathbf{S}_B = \frac{n}{g-1} \sum_{j=1}^{g} (\mathbf{X}_j - \overline{\mathbf{X}})(\mathbf{X}_j - \overline{\mathbf{X}})^T, \tag{1.9}$$

$$\mathbf{S}_{pw} = \frac{1}{N-g} \sum_{j=1}^{g} \sum_{i=1}^{n} (\mathbf{X}_{ij} - \overline{\mathbf{X}}_j)(\mathbf{X}_{ij} - \overline{\mathbf{X}}_j)^T, \tag{1.10}$$

with $j = 1, ..., g$ clusters and $i = 1, ..., n$ units per cluster. The total data matrix LF-T) is denoted by $\mathbf{X}_{ij}$, the between-cluster matrix, which contains the cluster means, is denoted by $\overline{\mathbf{X}}_j$, and the within-cluster data matrix LF-W, which contains the unit-wise deviations from these cluster means, is denoted as $(\mathbf{X}_{ij} - \overline{\mathbf{X}}_j)$. Moreover, $\overline{\mathbf{X}}$ designates a row vector with grand mean estimates and ^{T} marks a matrix transpose.

Unfortunately, however, the sample covariance matrix does not always yield an unbiased estimate of the population covariance matrix. For the within-cluster level, though, it holds true. The pooled within-cluster sample covariance matrix $\mathbf{S}_{pw}$ is an unbiased MLE of the population covariance matrix $\mathbf{\Sigma}_W$. For the between-cluster level, the unbiased MLE of the population covariance matrix $\mathbf{\Sigma}_B$ is a function of the sample covariance matrices of both levels (B. O. Muthén, 1990, 1994):

$$\hat{\mathbf{\Sigma}}_B = \frac{1}{c}(\mathbf{S}_B - \mathbf{S}_{pw}), \tag{1.11}$$

where c denotes the common cluster size. In the case of balanced data (i.e., balanced cluster sizes), $c = n$. Otherwise:

$$c = \left[N^2 - \sum_{j=1}^{g} n^2 \right] [N(g-1)]^{-1}, \tag{1.12}$$

When the data is unbalanced but consists of large numbers of clusters, the common cluster size is close to the average of the cluster sizes. Both LF covariance matrices with their all-units (co)variances, $\hat{\mathbf{\Sigma}}_B$ and $\hat{\mathbf{\Sigma}}_W$, can be found to the left in Figure 1.3. Note that, as the 'genuine' multilevel SEM is the LF approach, the indices $_{LF-B}$ and $_{LF-W}$ were simplified to $_B$ and $_W$ (or $_{pw}$ in the case of the pooled within-cluster estimator) to ensure consistency with Equation 1.8.

In WF, one sample covariance matrix is estimated off WF-T. Because WF-T is a single-level represented two-level data matrix, the MLE for single-level data is used. The single-level represented two-level sample covariance matrix reads:

$$\mathbf{S}_{WF-T} = \frac{1}{g}\sum_{j=1}^{g}(\mathbf{X.i}_j - \overline{\mathbf{X.i}})(\mathbf{X.i}_j - \overline{\mathbf{X.i}})^T, \tag{1.13}$$

where WF-T is designated by **X.i** and $\overline{\mathbf{X.i}}$ denotes a row vector with grand mean estimates. The sample covariance matrix is the population estimator, $\mathbf{S}_{WF-T} = \hat{\Sigma}_{WF-T}$. Note, however, that in Equation 1.13 the so-called biased MLE, which has g in the denominator in its leftmost term, is depicted. I focus on this estimator, and not the unbiased one, which has $g-1$ in the denominator, because it is the default in single-level SEM in *lavaan*[3]. The WF covariance matrix $\hat{\Sigma}_{WF-T}$ with its specific-units (co)variances can be found in Figure 1.3 to the right. Note that in chapter 4, it will be illustrated how the elements of the $p \cdot n$ "specific-units" (co)variances in $\hat{\Sigma}_{WF-T}$ correspond to the p "all-units" (co)variances in $\hat{\mathbf{\Sigma}}_B$ and $\hat{\mathbf{\Sigma}}_W$, σ^2_B and σ_B and σ^2_W and σ_W for both variables x_1 and x_2. This will be done by reformulating the "specific-units" (co)variances to the parameters of a random intercept-only model (which will be introduced in the next section).

Figure 1.3: Sample Covariance Matrices in Long Format (LF) and Wide Format (WF)

Long Format (LF) **Wide Format (WF)**

		cols: p	
		x_1	x_2
rows: p	x_1	-0.25	0.29
	x_2	0.29	-0.18

Between-Cluster Covariance Matrix $\hat{\Sigma}_B$

		cols: p	
		x_1	x_2
rows: p	x_1	1.92	-0.38
	x_2	-0.38	1.01

Within-Cluster Covariance Matrix $\hat{\Sigma}_W$

		cols: p · n			
		$x_{1.1}$	$x_{1.2}$	$x_{2.1}$	$x_{2.2}$
cols: p · n	$x_{1.1}$	9.35	-1.19	-1.49	4.05
	$x_{1.2}$	-1.19	3.50	-0.55	-1.46
	$x_{2.1}$	-1.49	-0.55	4.92	-1.38
	$x_{2.2}$	4.05	-1.46	-1.38	1.93

Total Covariance Matrix $\hat{\Sigma}_{WF-T}$

Estimated from example data set with $g = 4$, $n = 2$, $N = 8$, and $p = 2$.

Model Specification

Next, the model specifications for an random intercept-only model (e.g., J. J. Hox et al., 2017; Raudenbush & Bryk, 2002) in both approaches are presented. Put differently, the (co)variance structure of the p observed variables for the within- and between cluster parts, $\sigma^2_{x_1}$, $\sigma^2_{x_2}$, and $\sigma_{x_1x_2}$ for both levels, are modelled. These are equivalent to the LF covariance matrices $\mathbf{\Sigma}_W$ and $\mathbf{\Sigma}_B$.

In the LF approach, the same model specification for each level is applied to fit the random intercept-only model. The (co)variances, $\sigma^2_{x_1}$, $\sigma^2_{x_2}$, and $\sigma_{x_1x_2}$ for the within-cluster and between-cluster level, are modelled as (co)variances of the p observed variables from $\hat{\mathbf{\Sigma}}_W$ and

[3] See description of lav_matrix_cov() function in reference manual p.89, https://cran.r-project.org/web/packages/lavaan/lavaan.pdf, accessed on 24 January 2025.

$\hat{\boldsymbol{\Sigma}}_B$, respectively. The random intercept-only model in the LF approach is illustrated on the left side in Figure 1.4. The within-between decomposition that is taking place, reads:

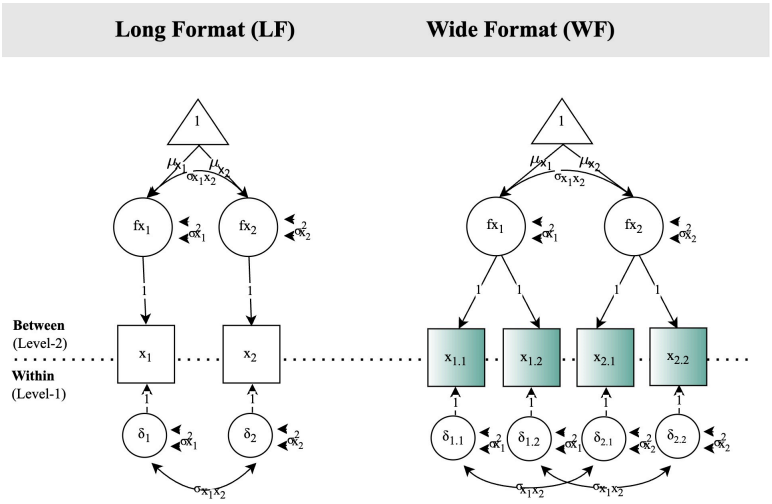
$$x_{ij} = \mu + \mu_j + \epsilon_{ij}, \tag{1.14}$$

where μ is the grand mean of x, μ_{0j} are the cluster means, and ϵ_{ij} are the deviations of each person in the cluster from the cluster mean. Note that a generic equation for x is shown and the index for the number of the variable is dropped to increase readability (i.e., x_{ij} instead of, for instance, x_{1ij} is used). The within and between components have the following distributions:

$$\mu_j \sim N(\mu, \sigma_b^2), \epsilon_{ij} \sim N(0, \sigma_w^2) \tag{1.15}$$

where σ_b^2 is the between-cluster variance and σ_w^2 is the within-cluster variance.

Figure 1.4: Model Specification in Long Format (LF) and Wide Format (WF)



Between-cluster parameters are located above the dashed line; within-cluster parameters below. At the within-cluster level, identical parameter labels indicate equality constrains. Model specification is contingent on the data set with $n = 2$ and $p = 2$.

In the WF approach, a restricted CFA is fit, where the multilevel random intercept-only model becomes a multivariate single–level model. On the right side in Figure 1.4 the random intercept-only model in the WF approach is presented. In contrast to the LF approach, all parameters are modelled as (co)variances of latent factors of the observed $p \times n$ specific-units variables from $\hat{\boldsymbol{\Sigma}}_{WF-T}$. To estimate the all-units (co)variances, several parameters are fixed or equality constrained. The WF approach offers an intuitive perspective on the between-within variance decomposition. The between-cluster part of every observed variable p is modelled by one common factor of the n specific-units variables in the WF data matrix. All factor loadings are set to 1. The within-cluster part of every observed variable p is modelled by unique factors.

The (co)variances of the n unique factors of each variable p are equality constrained. By the equality constraints, homoscedasticity of (co)variances of specific-units variables is modelled. The within-between decomposition in the single-level model reads:

$$x.i_j = \mu + \epsilon_j, \tag{1.16}$$

where μ are the grand means of $x.i$ and ϵ_j are the deviations of each person from these grand means. For instance, ϵ_3 is deviation of the person of the third cluster from each $x.i$. The random intercept-only model in the WF approach, which is fit as restricted CFA, can be seen on the right side in Figure 1.4.

2 Small Samples

It is one of the most often repeated phrases in statistics classes that sample sizes have to be large in order to obtain good estimates. More specifically, this holds true for traditional maximum likelihood estimation (MLE), which is the most often used type of estimation in psychology and the education sciences. As it relies on asymptotic theory, or in other words, finitely large samples, MLE is at risk of yielding non-converging models or parameter estimates with high sampling error in small samples (Hart & Clark, 1999; Jackson et al., 2013). Even more, MLE might produce estimates that are skewed away from the population parameters even if no other bias is present (Greenland et al., 2000); a problem which has been coined "small sample bias". This suggests that small samples are a severe hazard for empirical research.

In multilevel modelling, challenges arise when samples are small at any level. This causes MLE either to fail to converge or produce highly inaccurate estimates of higher level parameters (e.g., Finch & French, 2011; J. J. Hox & Maas, 2001; J. J. Hox et al., 2010a; Lüdtke et al., 2008, 2011; D. McNeish & Stapleton, 2016; D. M. McNeish & Stapleton, 2016; Meuleman & Billiet, 2009; Shin & Raudenbush, 2010; Stegmueller, 2013; Zitzmann, 2018; Zitzmann et al., 2015). Most often, these losses in accuracy have been found to concern bias in random effects, but several studies also found bias in fixed effects, for instance, in observed cluster means (Shin & Raudenbush, 2010; Zitzmann, 2018; Zitzmann et al., 2015), regression coefficients (D. McNeish & Stapleton, 2016), factor loadings and residual variances in measurement models (Meuleman & Billiet, 2009), contextual effects (Lüdtke et al., 2008, 2011), or cross-level interaction effects (Stegmueller, 2013). Moreover, these inaccuracies have been found to dependent more strongly on the sample size at the higher level (i.e., the number of clusters; Afshartous, 1995; Clarke, 2008; J. J. Hox & Maas, 2002; Kreft & Yoon, 1994; Maas & Hox, 2004b; Mok, 1995). Usually, a study design with larger numbers of smaller clusters is preferred over a design with smaller numbers of clusters with larger clusters (Mok, 1995). The cluster size appears to be only detrimental when marginally small. For example, Clarke (2008) found that higher level variance was overestimated when the cluster size equalled two, even when the number of clusters was large ($g = 200$). However, when the cluster size was increased to five, the bias disappeared. Thus, as other studies (J. J. Hox & Maas, 2001; Mok, 1995) confirmed, larger

cluster sizes can benefit estimation accuracy, but this advantageous effect is limited. In sum, evidence suggests that unless the cluster size is smaller than five, the number of clusters is the more important sample size in multilevel modelling. Thus, the "small sample bias" in multilevel analysis might be termed the "sparse clusters bias".

Whereas research of past decades has culminated in a plethora of sample size recommendations, the answer to the question of what is a sufficiently large sample size is ambiguous. For example, multilevel literature offers highly varying recommendations, ranging from just 8 or 10, over 20 to 30, 40, 50, or even 100 clusters to obtain converging and accurate models (Afshartous, 1995; J. J. Hox and Maas, 2001; J. J. Hox et al., 2010a; Kreft and De Leeuw, 1998; Lüdtke et al., 2008; Maas and Hox, 2004a; Meuleman and Billiet, 2009; Rabe-Hesketh and Skrondal, 2008; for a review see also D. Hox and McNeish, 2020; D. M. McNeish and Stapleton, 2016). It is noteworthy, that it has been pointed out that one needs larger numbers of clusters for obtaining accurate random effects than for fixed effects (see e.g., Afshartous, 1995; D. Hox & McNeish, 2020; J. J. Hox & Maas, 2001). However, these ad hoc conjectures are mainly based on simulation studies with varying magnitudes of sample sizes but small numbers of observed variables solely. Thus, recommendations are not generalizable to conditions with many observed variables (M. Yang et al., 2018). It has been shown that sample size recommendations depend on the number of observed variables and the type (e.g., measurement or structural model) and complexity (e.g., number of freely estimated parameters) of the model (e.g., Marcoulides et al., 2023). Thus, the literature diverges on the conclusion of precisely when the sample size becomes "too small", as "small samples" are only half of the story.

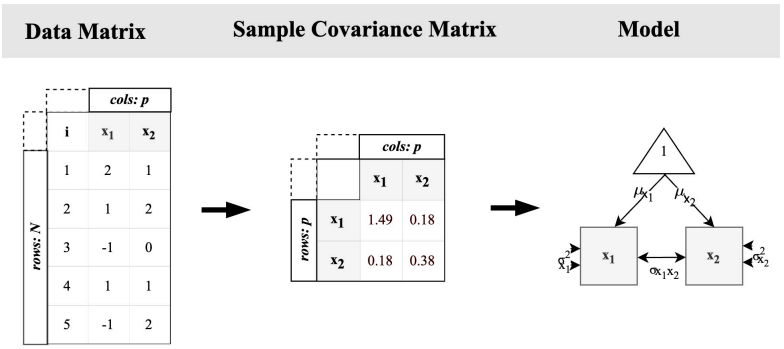
In single-level data, there is general agreement that conditions with a small sample size N paired with a large number of observed variables p, often abbreviated the other way around by "large p, small N", are troublesome. As a matter of fact, these conditions are frequently encountered in psychology and the education sciences, which is why it has received a great deal of attention in the literature (for an overview related to SEM, see, e.g., Deng et al., 2018; Marcoulides et al., 2023). In the following, it will be first reviewed why considering the sample size and the number of observed variables together has been found to be so important in single-level analysis. Subsequently, implications for multilevel analysis are derived.

2.1 Large p, small N in Single-Level Analysis

The term "large p, small N" already suggests that the attributes "large" and "small" become more tangible when these two magnitudes are related to each other. The number of observed variables is "large" in relation to the "small" sample they are collected from. In adjacent research field such as computational statistics and machine learning, the term high-dimensional, in which "dimensionality" refers to the number of observed variables, is more prominently used (Tibshirani & Hastie, 2007). Both terms have in common that they indicate that the number of observed variables can be considered large in one of two regards: the number of observed variables and the sample size could be of similar order, $p \sim N$, (see e.g., Ledoit &

Wolf, 2018), or the number of observed variables could be larger than the sample size, $p > N$, (see e.g., Pourahmadi, 2013; Sun, 2015; Tsukuma, 2016). While both scenarios have somewhat differing implications, the common theme is that when the dimensionality increases, the available data is likely to become sparse and estimation problems are the result.

Figure 2.1: Structural Equation Modeling: From Data Matrix to Sample Covariance Matrix to Model



Example data set with $N = 5$ and $p = 2$. A simple random intercept-only model is depicted.

Especially in the context of covariance structure modeling approaches such as SEM the ratio of the number of observed variables to the sample size, $p : N$, is of special relevance. To understand why, we need to elaborate first on the estimation procedure in SEM (see Figure 2.1). To estimate a specified model, first, the sample covariance matrix of the observed variables is estimated off the data matrix. Then, the model parameters are estimated by applying a fitting function, whose objective is to minimize the discrepancy between the sample covariance matrix and the model-implied covariance matrix (i.e., the model parameter estimates). Hence, the accuracy of the sample covariance matrix is important for the accuracy of the model parameter estimates.

Guidelines for the accuracy of the sample covariance matrix have been expressed in terms of $p : N$[1]. Minimum suggestions to properly estimate the sample covariance range from $1 : 5$ to $1 : 10$ (Allais, 1964; Everitt, 1975). However, these $p : N$ guidelines could not be consistently replicated in the context of covariance structure modeling. In factor analysis, there is mixed evidence for the importance of $p : N$ above N alone (e.g., Barrett & Kline, 1981; Cattell, 1978; Everitt, 1975; Gorsuch, 1983a), but, for instance, Arrindell and Van Der Ende (1985) suggests similar ratios than those for sample covariance matrices, more specifically a minimum of $1 : 10$, to guarantee factor stability.

Algorithms that estimate the model parameters use matrix algebra. Hence, the properties of involved matrices, such as those of the sample covariance matrix, are of crucial importance. While there are many relevant matrix properties, all can be traced back to one: the eigenvalues. Eigenvalues are a special set of scalars of a matrix. When at least one is zero, the matrix is

[1] Note that some sources denote $p : N$ the other way around as "subject-to-variables" ratio (SVT; see e.g., Beavers et al., 2013).

singular and thus, non-invertible. When at least one is zero or negative, the matrix is non-positive definite. When at least one is close to zero or the extrema are spread out largely, the matrix has a large condition number. When matrices are singular, non-positive definite or have large condition numbers (which by definition co-occurs), they are said to have distorted eigenstructures (compared to their population counterparts). Covariance matrices with distorted eigenstructures have been linked to lower convergence rates and less accurate parameter estimates in SEM (e.g., Arruda & Bentler, 2017; Depaoli & Clifton, 2015; Grilli & Rampichini, 2011; W. G. Hill & Thompson, 1978; Y. Huang & Bentler, 2015; Kelley, 1995; Lange et al., 1999; Lüdtke et al., 2008; Yuan & Bentler, 2017; Zitzmann, 2018; Zitzmann et al., 2015). Thus, the eigenvalues of the sample covariance matrix are important for the accuracy as well.

These sample eigenvalues have also been linked to $p : N$. Stein (1956) was the first to notice that the sample covariance matrix has a biased eigenstructure in small samples. In the context of eigenvalues, biased means that small eigenvalues are pushed downwards and large eigenvalues are pushed upwards compared to their population counterpart. It has been demonstrated, by analytical and empirical means, that the bias of the sample eigenvalues is of the order $p : N$ (Arruda & Bentler, 2017; Dempster, 1972; Hayashi et al., 2018; Schäfer & Strimmer, 2005; Stein, 1956, 1975). Thus, when p is large and N is small, $p \sim N$ or $p > N$, the ratio becomes close to 1 or even exceeds it which indicates a high level of eigenvalue bias. As mentioned above, with distorted eigenstructure, non-convergence and inaccuracy are more likely to occur. Moreover, as mentioned in the last chapter, the way that the traditional MLE is implemented in *lavaan* requires a positive definite sample covariance matrix (Hamaker et al., 2003; H. Singer, 2010; Van Montfort et al., 2018; Voelkle et al., 2012; Walther, Hecht, Nagengast, & Zitzmann, 2024), which, amongst others, necessitates that the supplied data matrix has just as many or less columns than rows, $p \leq N$. Otherwise, the model will not converge. However, this software restriction can be circumvented by using another estimator, more specifically, the full information maximum likelihood (FIML) estimator, which uses the raw data matrix instead (Hamaker et al., 2003; Trendafilov & Unkel, 2011; Unkel & Trendafilov, 2010; Voelkle et al., 2012). Thus, large $p \leq N$ which comes close to 1 rather poses threats to model estimation (and not convergence) by distorting the eigenstructure.

Finally, the combined effects of $p : N$ on matrix properties of the sample covariance matrix and on model estimation in SEM have been sparsely connected. In this regard, however, the study of Yuan and Chan (2008) is standing out. They found that both convergence rate and accuracy of model parameter estimates decrease with increasing $p : N$, and that altering the sample covariance matrix by adding a constant of the size $p : N$, which alters the distorted eigenstructure, convergence and estimation accuracy were improved. In summary, evidence suggests that $p : N$ is an important quantity for the sample covariance matrices and parameter estimates in single-level SEM but their relation needs to be further addressed.

2.2 Large p, small g in Multilevel Analysis

In two-level data, there is no unequivocal equivalent of the ratio of the number of observed variables to the sample size, $p : N$. Instead, there are different numbers of observed variables and sample sizes at both levels: at level-1 (within-cluster level), there are individual-level variables (e.g., age, sex, gender, location, motivation) and the total sample size, which is calculated as the numbers of cluster multiplied by the cluster size, $g \cdot n = N$, and at level-2 (between-cluster level) there are contextual variables (e.g., classroom, teacher, school) and aggregated individual-level variables (e.g., average age or motivation) and the number of clusters g. For means of simplification, I consider only observed aggregated individual-level variables at level-2 such that the number of observed variables is the same at both levels. Thus, the $p : N$ equivalents for each level are $p : N$ at level-1 and $p : g$ at level-2.

As mentioned earlier, in multilevel analysis, the major challenge is often posed by the sample size at the higher level, the number of clusters g. Thus, the occurrence of the "sparse clusters bias" might be narrowed down to "large p, small g" conditions in which $p : g$ at level-2 is of major importance. When scrutinizing the $p : g$ of the empirical studies recommending minimum cluster sizes earlier, we find that these range from $1 : 25$ (0.04) to $1 : 8$ (0.125). It appears that one can minimize the range of the recommendations substantially (earlier $g = 8$ to 100) by incorporating information on the number of observed variables. These numbers might be more suited for advising a minimum amount of clusters in research designs. However, generalization has to be cautious as the number of observed variables were set low throughout all these studies. Moreover, another factor might help explain the range of the recommendations: the variance at level-2 which is most often denoted by the ICC.

2.2.1 Large p, small g, small ICC

The ICC is suspected to have an impact on estimation accuracy as well. Some studies found that smaller ICCs demand for larger cluster sizes in order to guarantee convergence and accurate estimates of higher level parameters (Finch & French, 2011; J. J. Hox & Maas, 2001; J. J. Hox et al., 2010b; Lüdtke et al., 2011; Meuleman & Billiet, 2009; Muthen & Satorra, 1995; Zitzmann, 2018; Zitzmann, Wagner, et al., 2022), others found no relation (J. J. Hox & Maas, 2001). However, similar to small sample issues, whether the ICC has an influence might also depend on the type of parameter that one is scrutinizing. In a review, D. M. McNeish and Stapleton (2016) found that fixed effects are unaffected whereas random effects, such as ICC estimates, are influenced by the ICC in the population. Thus, evidence suggests that smaller ICCs result in less accurate estimates of higher level (co)variance parameters.

Moreover, there is evidence for an interaction of the number of clusters and the ICC. For instance, convergence issues have been found to be especially prevalent under conditions with small ICC and small numbers of clusters (Lüdtke et al., 2011). When coming back to the cluster size recommendations earlier, Maas and Hox (2004a) noted that there was only bias in parameter estimates when the ICC was small, or more particularly, the largest biased

manifested under conditions with both small numbers of clusters and small ICC. Moreover, Hedges and Hedberg (2013) advised to consider the expected ICC when planning studies because it plays important role in convergence and estimation accuracy. In this regard, it is useful to divide factors relevant in planning a study into sample and population characteristics. Whereas sample characteristics can be modified by the study design, population characteristics cannot. Referring to the large p, small g, small ICC conditions, the number of observed variables p and the number of clusters g can be set by the researcher and the ICC can only be accounted for by the sample characteristics, for instance, by using larger cluster sizes[2].

Evidence for an interplay of the number of clusters and the ICC is also given when examining matrix properties. With decreasing g and increasing p (thus, increasing $p : g$), and decreasing ICC, the probability of non-positive definite between-cluster covariance matrices increases (Bhargava & Disch, 1982; W. G. Hill & Thompson, 1978; Searle et al., 1992). This is reasonable: when variances are small compared to covariances, then on-diagonal elements in the sample covariance matrix are small compared to off-diagonal elements and this makes non-positive definiteness more likely. In turn, between-cluster covariance matrices with distorted eigenstructure (such as non-positive definiteness) have been related to less accurate between-cluster parameter estimates (Depaoli & Clifton, 2015; Grilli & Rampichini, 2011; W. G. Hill & Thompson, 1978; Lüdtke et al., 2011; Zitzmann, 2018; Zitzmann et al., 2015).

In sum, the combined evidence suggests that larger p: g are required when random effects are scrutinized in populations with low ICC. As a matter of fact, in the educational sciences, one is often interested in estimating random effects and ICCs are usually at the lower end (Adams et al., 2004; Gulliford et al., 1999). Thus, especially here, taking p: g instead of only g into account may foster optimal study designs that yield accurate model estimations.

Another important factor has to be noted: the data format. In the studies reported so far, both MLM and ML SEM have been utilised, and the former uses the data in WF whereas the latter usually uses the data in LF. When revisiting single-level SEM in Figure 2.1, we see that p and N relate to the dimensions of the data matrix. The number of columns is indicated by the number of observed variables p and the number of rows is indicated by the sample size N. Thus, $p : N$ can be expressed as $cols : rows$ which gives a figural notion of the data matrix by which the sample covariance matrix is estimated. The same data set results in data matrices with different ratios in the two data formats: in LF, p: $(g \cdot n = N)$ for LF-W and p: g for LF-B, an in WF, $(p \cdot n)$: g for WF-T. In LF, we see the p: N equivalents for level-1 (LF-W) and level-2 (LF-B) noted earlier. In WF, however, where "people are variables, too" (Mehta & Neale, 2005), there is only one data matrix which has larger $cols : rows$ than LF-B. While the empirical equivalence of both the LF and WF approaches to ML SEM have been demonstrated (Barendse & Rosseel, 2020; Mehta & Neale, 2005), only conditions with a small number of

[2]Note that in my first research project, I classified the number of observed variables p to the population characteristics because they are set for a certain (population) model. However, from a practical perspective, when models are too complex and do not converge, researchers may exclude certain observed variables from the analysis. Thus, I changed allocating p to sample characteristics.

observed variables and a large number of clusters at level-2 have been scrutinized, resulting in $cols \ll rows$ in both the LF and WF data matrices. In the last section, we discussed that larger $cols$: $rows$ are associated with larger eigenvalue bias, and smaller convergence rates and less accurate parameter estimates in single-level analysis. This suggests, that under conditions in which highly different $cols$: $rows$ in the LF and WF data matrices are the result, performances might differ. This might be especially true for the troublesome "large p, small g, small ICC" conditions in which random effects are estimated. This is what my first research project investigated.

3 Research Project (1) To be Long or to be Wide: How Data Format Influences Convergence and Estimation Accuracy in Multilevel Structural Equation Modeling

Background

The main objective of my first research project was to scrutinize the equivalents of the $p:N$ ($cols:rows$) effects for multilevel data. Motivated by evidence for a $p:N$-sized bias in eigenvalues in sample covariance matrices and larger $p:N$ being associated with lower convergence rates and less accurate estimates in single-level analysis, I was interested in whether these findings generalize to multilevel analysis where equivalents have not been investigated before. In multilevel data, one of two data formats, long format (LF) or wide format (WF), has to be chosen from, and both result in differing $cols:rows$. For both data formats approaches to multilevel SEM have been proposed and both can be utilised in *lavaan* in *R*. Whereas evidence from single-level data implies that their performances differ, studies comparing the performance of the LF and WF approach found no substantial differences. However, these included only conditions with small numbers of observed variables p, small cluster sizes n, and large numbers of clusters g, which resulted in small $cols:rows$ in both data formats. Thus, I investigated their performance under conditions of varying $cols:rows$ in both data formats. A special interest was in the "large p, small g, small ICC" conditions in which multilevel modeling generally encounters trouble with estimation. The $cols:rows$ effects were divided into two intertwined effects: (a) the effect of the data format, because the data format inherently leads to different $cols:rows$ (given the same sample), and (b) the effect of $cols:rows$ in each data format (given different samples). Whereas the effect of the data format answers which data format (approach) to use, the effect of $cols:rows$ answers which $cols:rows$ to aim at with our study design.

Method

To investigate the empirical equivalence of the matrix properties and model performances of both the LF and WF approaches, a simulation study was implemented. Two-level random intercept-only models for continuous, balanced data were scrutinized. Factors that were varied within the simulation study design were either related to population characteristics, the number of observed variables ($p = 2, 5, 10, 20$) and the ICC ($ICC = 0.05, 0.25, 0.50$), or sample characteristics, the cluster size ($n = 2, 5, 10, 20, 100$) and the number of clusters ($g = 2, 4, 5, 10, 20, 25, 40, 50, 100, 200$). This distinction was made to emphasize that population characteristics cannot be changed (though one could drop observed variables from modeling but model missspecification was not of interest) whereas sample characteristics can be changed. The number of observed variables p, the cluster size n and the number of clusters g were varied in such a way that the data matrices in both data formats, LF-W and LF-B in LF and WF-T in WF, resulted in $cols : rows$ with substantial differences. The ICC values were chosen to reflect the lower and upper levels of frequently encountered ICCs in the social sciences. The design of the simulation study was partially crossed and resulted in 240 conditions in which 1.000 data sets were replicated, respectively. Evaluation criteria for matrix properties, which indirectly captured the eigenvalue bias, included singularity and definiteness. For model performance, convergence rates, and relative root means squared error (RMSE), relative bias, and relative variance of (co)variance parameter estimates (i.e., random effects) were scrutinized. For the three measures of estimation accuracy, aggregates across parameter types (i.e., variances and covariances) were reported.

Results

At first, I examined which $cols : rows$ of the two data matrices in LF are determinative. The within-cluster (level-1) data matrix LF-W corresponds to $p : (g \cdot n)$ and the between-cluster (level-2) data matrix LF-B to $p : g$. Results suggested that the latter is more important for model performance. This is in line with evidence suggesting that the sample size at the higher level is the more important factor for the accuracy of higher level estimates (Afshartous, 1995; Clarke, 2008; J. J. Hox & Maas, 2002; Kreft & Yoon, 1994; Maas & Hox, 2004b; Mok, 1995). Thus, in the following, the crucial $cols : rows$ in LF correspond to $p : g$ (LF-B).

The findings suggested that there was an effect of data format on convergence, but not on estimation accuracy (even when $cols : rows$ in both approaches were substantially different; see results of $cols : rows$ effects in the next paragraph). Moreover, for estimation accuracy, I found an interaction effect of g and ICC for the level-2 (co)variance parameter. The least accurate, most biased, and inefficient parameter estimates were obtained when samples with small numbers of clusters g and small ICC values co-occurred.

Concerning the $cols : rows$ effects in both approaches, results indicated that these are different. For convergence, $cols < rows$ in LF-B ($p < g$) and $cols \leq rows$ in WF-T ($(p \cdot n) \leq g$) had to be satisfied. For estimation accuracy, the $cols : rows$ effect in the WF approach was less

strong. Both findings are in line the the results of the effect of the data format. In addition, evidence for an interaction of $cols:rows$ and ICC emerged here as well. Larger $cols:rows$ combined with smaller ICC were most detrimental.

Results of the matrix properties did not offer substantial surplus value above the $cols:rows$ for both convergence and estimation accuracy. For instance, in $cols < rows$ but also in $cols > rows$ non-positive definite $\hat{\boldsymbol{\Sigma}}_{LF-B}$ occurred, but only in the former case the models did not converge. This suggests that non-definiteness of $\hat{\boldsymbol{\Sigma}}_{LF-B}$ may be a necessary but no sufficient condition for non-convergence. Potentially, the underlying eigenvalue biases (which were only approximated in this study by $cols:rows$) may be more informative than any other matrix property.

To sum up, under $(p \cdot n) \leq g$ conditions, both approaches converged and the estimation accuracy was very similar which is in line with both approaches exhibiting different $cols:rows$ effects.

Discussion

A two-level data set can be arranged in two different data formats, LF and WF, and for both, SEM approaches to estimate multilevel models are readily available. Although past research demonstrated empirical evidence, only conditions with large sample sizes and small numbers of observed variables were observed. In investigated convergence and estimation accuracy under "large p, small g, small ICC" conditions with substantially different $cols:rows$ in the relevant data matrices in both approaches. Results suggested that as long as $cols < rows$ in the LF approach and $cols \leq rows$ in the WF approach, models did converge and estimation accuracy was very similar in both approaches even when $cols:rows$ differed substantially. however, these results have to be seen in perspective with certain limitations of the study. Firstly, the $cols:rows$ and matrix properties of the data matrices were only considered as proxies of the eigenvalue bias of the LF and WF covariance matrices. Thus, the exact sample eigenvalue biases are likely to differ and they might be more informative than the matrix properties. Secondly, only simple random intercept-only models were considered and thus, generalizability to more complex models is unclear. An especially promising venue for future research is also how the improvement of the eigenvalue structure in the LF and WF covariance matrices might improve model performance. This is what my second research project investigated. Before coming to the very project, the overarching topic of regularization is introduced.

4 Regularization

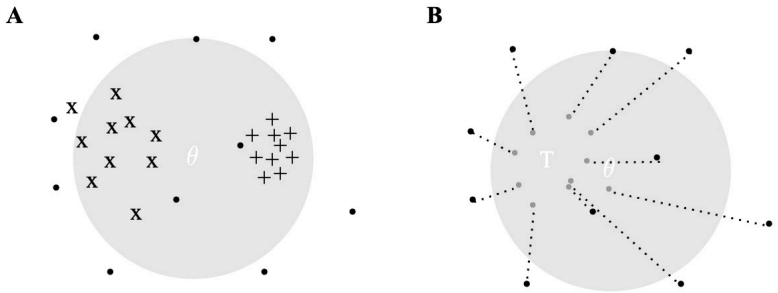
In multilevel analysis, challenges arise when sample sizes are small compared to the number of observed variables, especially at level-2. This is a condition that may cause MLE methods to either fail to converge or produce highly inaccurate estimates of level-2 parameter estimates (e.g., J. J. Hox & Maas, 2001; J. J. Hox et al., 2010a; Lüdtke et al., 2008, 2011; D. M. McNeish & Stapleton, 2016; Meuleman & Billiet, 2009; Shin & Raudenbush, 2010; Stegmueller, 2013; Zitzmann, 2018; Zitzmann et al., 2015). Unfortunately, collecting larger samples can be costly, time-consuming, or impractical for certain study designs, such as pilot studies with limited resources or when populations are small. This is especially true for sampling at the higher level in fields such as developmental psychology and the education sciences (see also D. Hox & McNeish, 2020, who pointed this out). In cluster sampling, first the level-2 units (e.g., classrooms) and then the level-1 units (e.g., students within classrooms) are sampled. When populations at level-2 are inherently small, such as school boards, then the issue becomes even more severe. There is sparse research on and little consensus about how to handle data analysis with very few clusters (see D. McNeish & Stapleton, 2016). Small ICCs, which are quite common in psychology and the education sciences (Adams et al., 2004; Gulliford et al., 1999), add to existing issues with non-convergence and accuracy of higher level parameter estimates (J. J. Hox & Maas, 2001; Lüdtke et al., 2011; Muthen & Satorra, 1995; Zitzmann, 2018; Zitzmann, Wagner, et al., 2022). Thus, under large p, small g, small ICC conditions, there is a need for an alternative approach that is straightforward to implement and can mend both convergence and accuracy issues. Regularization might be an appropriate candidate.

Regularization is a broad class of methods that encompasses several techniques which bias an estimator, such as MLE, in order to give reasonable answers in unreasonable situations (adapted from Bickel et al., 2006). The idea of regularization was originally developed by Tikhonov (1943) to address stability issues in inverse matrix problems and soon adopted by the statistical community. Thus, "unreasonable situations" refer to a variety of (overlapping) scenarios, among them non-convergence, inadmissible solutions, distorted eigenvalue structure of matrices, inaccurate estimates, and overfitting. Similarly, "reasonable answers" can mean differently, depending on the situation. It may mean to obtain a more well-behaved eigen-

strucure (e.g., invertible matrices) required for certain calculations or to improve accuracy of estimation in finite samples. As noted earlier, eigenstructure on one hand and convergence and estimation accuracy on the other hand have been linked (e.g., Arruda & Bentler, 2017; Depaoli & Clifton, 2015; Grilli & Rampichini, 2011; W. G. Hill & Thompson, 1978; Y. Huang & Bentler, 2015; Kelley, 1995; Lange et al., 1999; Lüdtke et al., 2008; Yuan & Bentler, 2017; Zitzmann, 2018; Zitzmann et al., 2015).

The first one to apply regularization is statistics has been Stein. He found counter-intuitive results when estimating independent means in multivariate models. Whereas estimating three or more means from normal distributions independently may seem optimal, Stein demonstrated that "shrinking" each individual estimate towards the overall mean yielded more accurate estimates (see James & Stein, 1961; Stein, 1956). This finding is known as *Stein's Paradox* because it defies the intuition that independent estimates should minimize error on their own. At this point it is noteworthy that several scholars suggested that shrinkage estimation can be viewed as a form of Bayesian estimation, an estimation with priors such as the overall mean. Whereas some relate it to Bayesian estimation with weakly informative priors (e.g., Anderson, 2003; Ledoit & Wolf, 2004; Tibshirani, 1996; Yuan & Chan, 2008), others emphasize parallels with empirical Bayes because of its data-driven nature (Greenland et al., 2000).

Figure 4.1: Properties of Estimators: Bias and Variance



A Three Estimators with Different Bias and Variance
B Bias-Variance Tradeoff Through Regularization

In statistical terms, regularization induces a so-called bias variance trade-off. Bias and variance are two fundamental properties regarding the accuracy of estimators. Bias of an estimator is the difference between the expected value of the estimator (i.e., the average estimate) and the population value of the parameter being estimated (θ). It tells us whether the estimator consistently over- or underestimates the population value. Variance of an estimator measures the spread or variability of the estimator's sampling distribution around the expected value (i.e., the average estimate) of the parameter being estimated (θ). High variance indicates that the estimator's value fluctuates significantly given the sample drawn. Both properties are illustrated in Figure 4.1. In Panel A, the estimator denoted by "+" has large bias and small variance whereas "•" has small bias and large variance. Traditionally, MLE tries to keep

both bias and variance minimal (though in finite samples, small sample bias exists). This is illustrated by the estimator denoted by "x". However, if we let go of the unbiasedness criterion, we can minimize the variance, thereby improving the accuracy of estimation. This is depicted in Panel B. The unbiased but largely variant estimates "•" get biased towards a target ("T") which is distinct from the parameter being estimated (θ). Thereby, a decrease in variance is traded for an increase in bias, overall improving accuracy. This admittance of bias can be viewed as adjustment for anticipated random error (that is, variance) in the estimate (Cole et al., 2014).

Whereas Stein used a shrinkage technique to refine the traditional estimator (i.e., using a conjugate estimate of a target and the traditional estimate), more modern approaches revolve around techniques such as constraining, restricting or penalizing. Whereas mathematically rather different, all have in common that they modify the parameter space to a certain degree. Constraining refers to approaches which set parameters to certain admissible values. For instance, Zitzmann, Walther, et al. (2022) suggested to set so-called Heywood cases (i.e., inadmissible variance estimates that are negative or zero) to zero and they justified this procedure by the very definition of MLE. Similar but slightly different, restricting refers to limiting the possible range of the parameter values[1]. For instance, Chen et al. (2001) gave recommendations on when to deal with Heywood cases by restricting them to positive values. In penalizing, a penalty term is added to the likelihood function such that more parsimonious models yield higher likelihoods, such as in the work of Chung et al. (2013) in which negative variances are avoided thereby. All these regularization techniques did not only deal with the inadmissible solutions but convergence rates and estimation accuracies were improved as well.

In several techniques, such as shrinkage and penalizing, the degree of regularization is controlled by the regularization parameter (sometimes also named after the technique such as shrinkage parameter). In Stein's pioneering work on regularization, the parameter was not defined explicitly as we commonly use it today. Instead, the traditional estimator of the mean was shrunken towards the grand mean by accounting for the number of observed variables (i.e., the number of means to be estimated) and their variances (James & Stein, 1961) which might be understood implicitly as regularization parameter. In more recent techniques the regularization parameter might be fixed, computed by closed-form solutions or estimated otherwise from the data.

To this date, regularization has been applied to a variety of statistical methods. Starting out with mean estimation (see James & Stein, 1961; Stein, 1956), it has been adapted to (co)variance estimation as alternative to the traditional sample covariance matrix (James and Stein, 1961; Stein, 1964; for an overview see e.g., Engel et al., 2017; Fan et al., 2016; Ledoit and Wolf, 2022; Pourahmadi, 2013), regression analysis (e.g., Hoerl & Kennard, 1970; Tibshirani, 1996; Van Hoa, 1985; Zou & Hastie, 2005), factor analysis (e.g., Ahn & Horenstein, 2013; George & Oman, 1996;

[1] Note that the terms "constraining" and "restricting" are sometimes used synonymously in the literature. The definitions given here enable to tell both techniques apart.

Jung & Takane, 2007; Zhao et al., 2020), and most recently SEM (e.g., Burghgraeve et al., 2021; P.-H. Huang et al., 2017; Jacobucci et al., 2016). In many research fields such as econometrics, statistics, and machine learning, regularization is part of the standard tool box of data analysis. In contrast, the application of or even awareness about regularization in psychology and the education sciences is rather limited (see also D. M. McNeish, 2015, who pointed this out)[2]. This is especially true for regularization in the context of SEM which we will turn to in the following.

4.1 Regularization of Covariance Matrices in Structural Equation Models

As discussed earlier, in SEM, first a sample covariance matrix is estimated from the data matrix, and then the model parameters are estimated thereof (revisit Figure 2.1 in chapter 2). Thus, there are two options for applying regularization: (1) to the sample covariance matrix ("input") or (2) to the model parameters ("output"). Under large p, small g, small ICC conditions, two unreasonable situations, distorted eigenvalue structures and inaccurate estimations (in the sample covariance matrices and the model matrices), often co-occur. Targeting the output can only mend issues caused by model matrices. For instance, Zitzmann (2018) proposed a shrinkage estimator for the between-cluster predictor matrix to deal with their distorted eigenstructure and the obtained between-cluster parameter estimates have been found to be more accurate than MLE. However, non-convergence cannot be dealt with by applying regularization to the model parameters. This can only be done by targeting the input. It is also noteworthy some authors even hypothesize that the distorted eigenstructure of the sample covariance matrix might be the underlying problem of the inaccurate model estimates (Arruda and Bentler, 2017; Y. Huang and Bentler, 2015; Yuan and Bentler, 2017; see also Zitzmann, 2018). Thus, targeting the sample covariance matrix is very promising as it has the potential to mend both convergence and accuracy issues of the model at once. Note at this point that the choice fell on the WF approach out of the two multilevel SEM approaches because it utilises single-level SEM which is more enhanced in its development in *lavaan* (version 0.6-15). More precisely, only here users may supply a sample covariance matrix instead of a data matrix which is a prerequisite to apply regularization to the sample covariance matrix in SEM.

Possible approaches to the regularization of sample covariance matrices that have been shown to improve eigenvalue structure (and thus, convergence) and estimation accuracy (of the sample covariance matrix) include ridge and shrinkage approaches. Both have in common

[2]Note that there are other terms for regularization which are more commonly used in psychology and the education sciences, such as "stabilization" (often used in the context of accuracy and model selection; see e.g., Ulitzsch et al., 2023; Zitzmann, 2018) and "smoothing" (prominent in the context of improving the eigenstructure of covariance matrices; see e.g., Lorenzo-Seva & Ferrando, 2021; Wothke, 1993). More recently, "regularization" found its way into the mainstream literature to denote matters concerned with improper solutions, model sparsity, and overfitting (see e.g., Arruda & Bentler, 2017; Jacobucci et al., 2016; Jung & Takane, 2007; Liang & Jacobucci, 2020; Orzek & Voelkle, 2023; Williams & Rodriguez, 2022). Whereas these terms partly refer to the differing regularization techniques, there is neither a strict usage of these terms nor any consistent taxonomy.

that they impose a well-behaved eigenstructure on the sample covariance matrix but the techniques to achieve this differ. Whereas ridge adds values, shrinkage pulls towards values. Both approaches will be briefly reviewed in the following.

4.1.1 Ridge Approach

The initial ridge approach, the ridge regression, has been proposed by Hoerl and Kennard (1970) to deal with non-orthogonal problems in regression analysis. More specifically, when predictors are highly correlated (i.e., multicollinear), the predictor matrix has distorted eigenvalue structure which in turn may lead to unstable estimates with ordinary least squares (OLS) estimation. Hoerl proposed a class of estimators that augment the on-diagonal (i.e., variances) of the predictor matrix by small positive quantities in order to obtain more favourable eigenvalues and more stable estimates.

Eventually, the idea of the ridge regression was introduced in SEM to deal with distorted eigenvalue structure of sample covariance matrices (Jöreskog & Sörbom, 1996). Beyond the distorted eigenstructure and instable estimates, non-convergence might occur when iterative algorithms require invertible matrices. As in the original work of Hoerl and Kennard (1970), the initially proposed constant was fixed. Yuan and Chan (2008) extended the idea by suggesting to determine the constant by p and N and to routinely fit a ridged sample covariance matrix (instead of the traditional sample covariance matrix) in SEM. Various further extensions have been proposed, for instance, determining the constant by more sophisticated formulas which account for the data at hand (e.g., Bentler & Yuan, 2011; Kamada et al., 2014; Yuan & Bentler, 2017; Yuan et al., 2011) or augment the off-diagonal (i.e., covariances) instead of the on-diagonal (sometimes referred to as "anti-ridging" approaches; e.g., Bentler & Yuan, 2011; Kamada et al., 2014; Yuan & Bentler, 2017). These ridge approaches have been shown to be effective in SEM as well: empirical findings suggest that ridging results in increased convergence rates and more accurate model estimates (e.g., Kamada & Kano, 2012; Kamada et al., 2014; Yuan & Chan, 2008).

4.1.2 Shrinkage Approach

In shrinkage estimation, born out of the work of James and Stein (1961) and Stein (1964), the population covariance matrix Σ is estimated as a weighted average of the sample covariance matrix and a pre-specified target matrix. The target matrix, similar to the population covariance matrix, has a well-behaved eigenstructure. The amount of weighting is controlled by the shrinkage parameter $\lambda \in [0,1]$. Let us illustrate the concept of weighting using the extremes: If $\lambda = 0$, no shrinkage is applied, and the sample covariance matrix will be kept. If $\lambda = 1$, we obtain the target matrix as the estimate of Σ. To clarify, "shrinkage" does not necessarily imply that the elements (of the sample covariance matrix) get smaller in absolute terms, but they are shrunken towards a certain value (of the target matrix). For example, if $\sigma_S = 0.1$ and $\sigma_T = 1$, then 0.1 is "shrunken" towards 1.

Broadly speaking, there are two types of shrinkage estimators, linear and non-linear ones, which differ (a) in the set of elements that are targeted and (b) the shrinkage intensity that is applied to these. Linear shrinkage is a convex combination of the sample covariance matrix and the target matrix. Put differently, the same shrinkage intensity is applied to each element of the sample covariance matrix. To this day, various estimators that build on Stein's idea of linear shrinkage have been proposed, for instance, by Banerjee and Monni (e.g., 2021), Fisher and Sun (2011), Gray et al. (2018), Ledoit and Wolf (2004), Schäfer and Strimmer (2005), Touloumis (2015), and Zhang (2022). However, because it might be disadvantageous to shrink all elements equally, non-linear shrinkage approaches were introduced. Non-linear shrinkage is a concave combination of the sample covariance matrix and the target matrix, that allows for different regularization intensities of elements. Similarly, several estimators, such as those by Dey and Srinivasan (e.g., 1985), James and Stein (1961), Ledoit and Wolf (2012, 2018), and R. Yang and Berger (1994) have been proposed.

Shrinkage estimates are more likely to be positive definite, non-singular and well-conditioned, and more accurate than the traditional sample covariance matrix, as demonstrated in studies such as Ledoit and Wolf (2004, 2022). However, so far, limited research addressed shrinkage of covariance matrices within SEM (e.g. Arruda & Bentler, 2017; Arruda, 2017; De Jonckere & Rosseel, 2023). So far, results suggest that condition numbers in the shrinkage estimate are more close to their population counterpart, and that models have higher convergence rates, lower bias in parameter estimates, and more accurate standard errors, but more research is needed.

Now that both ridge and shrinkage approaches have been introduced, it stands to reason which one is more suited for application in multilevel SEM under large p, small g, small ICC conditions. There is one main reasons why I focus on shrinkage and not on ridge approaches within my research. Ridge approaches treat all targeted elements of the covariance matrix equally. The same *absolute* quantity is added to each element in the on-diagonal (i.e., variances in the standard ridge approaches) or off-diagonal (i.e., covariances in the anti-ridge approaches). In contrast, even in linear shrinkage approaches, in which the same regularization parameter is applied to all targeted elements, the absolute differences between the traditional sample covariance matrix and the shrinkage estimate are different. Put differently, the same *relative* quantity is added to each element. Hence, shrinkage approaches target extreme values more which is very reasonably given that sampling error is likely to differ for each variable as well. The ridge and shrinkage approaches differ because of their objectives: in ridge, a well-behaved sample covariance matrix is desired to prevent non-convergence and in shrinkage, both a well-behaved and a more accurate sample covariance matrix are intended.

In the next section, the linear shrinkage estimator proposed by Touloumis (2015) will be introduced, as it appears suited for application in the context of multilevel SEM and an implementation in *R* is readily available.

Covshrink: A Linear Shrinkage Estimator of the Covariance Matrix

The popular linear shrinkage estimator by Ledoit and Wolf (2004) was refined by Touloumis (2015) through (1) expanding the set of target matrices, and (2) deriving consistent closed form solutions of the shrinkage parameters under "small N, large p" conditions in single-level data. The resulting, new estimators have been demonstrated to improve estimation accuracy of covariance matrices in terms of MSE as compared with the traditional sample covariance matrix under these conditions (Touloumis, 2015). The general equation for the linear shrinkage estimator reads:

$$\hat{\mathbf{S}}^{*} = (1 - \hat{\lambda})\mathbf{S} + \hat{\lambda}\mathbf{T}, \tag{4.1}$$

in which $\mathbf{S}$ is the unbiased MLE of the $p \times p$ population covariance matrix, $\mathbf{T}$ is the target matrix, and $\hat{\lambda}$ is the shrinkage parameter. The target matrix is one of three diagonal matrices: the equal target matrix $\hat{v}\mathbf{I}_p$ which contains the overall mean of the sample variances in the on-diagonal (originally proposed by Ledoit & Wolf, 2004), the identity matrix $\mathbf{I}_p$ which has ones in the on-diagonal, or the unequal target matrix $\mathbf{D}_S$ which adopts the sample variances in the on-diagonal. This implies that the on-diagonal elements (i.e., variances) of the shrinkage estimate are pulled towards the mean of the sample variances (in case of the equal target matrix $v\mathbf{I}_p$), one (in case of the identity matrix $\mathbf{I}_p$), or are left unaltered (in case of the unequal target matrix $\mathbf{D}_S$). Across all target matrices, off-diagonal elements (i.e., covariances) of the shrinkage estimate are systematically pulled towards zero. For each target matrix, there is a different formula for the shrinkage parameter. The closed form solution for the shrinkage parameter of the equal matrix $\hat{v}\mathbf{I}_p$, in which $\hat{v} = Y_{1N/p}$, is:

$$\hat{\lambda}_E = \frac{Y_{2N} + Y_{1N}^2}{NY_{2N} + Y_{1N}^2 + \frac{p-N+1}{p} Y_{1N}^2}, \tag{4.2}$$

for the shrinkage parameter of the identity matrix $\mathbf{I}_p$:

$$\hat{\lambda}_I = \frac{Y_{2N} + Y_{1N}^2}{NY_{2N} + Y_{1N}^2 - (N-1)(2Y_{1N} - p)}, \tag{4.3}$$

and for the shrinkage parameter of the unequal target matrix $\mathbf{D}_S$:

$$\hat{\lambda}_U = \frac{Y_{2N} + Y_{1N}^2 - 2Y_{3N}}{NY_{2N} + Y_{1N}^2 - (N-1)Y_{3N}}, \tag{4.4}$$

in which Y_{1N}, Y_{2N}, and Y_{3N} are combinations of U-statistics (for their estimation, see pp. 5 and 12 in Touloumis, 2015). The optimal shrinkage intensity, which minimizes the MSE between

the population covariance matrix and the respective shrinkage estimator, is approximated by sample-based unbiased and ratio-consistent estimators (Touloumis, 2015). The resulting biased shrinkage estimators of Σ are $\hat{\mathbf{S}}_E^*$ (equal target matrix), $\hat{\mathbf{S}}_I^*$ (identity target matrix), and $\hat{\mathbf{S}}_U^*$ (unequal target matrix). Because of the closed form of the shrinkage parameters, the approach is computationally fast regardless of the number of observed variables p. In addition, the obtained estimates have a more well-behaved eigenstructure, for instance, matrices are non-singular and well-conditioned (Touloumis, 2015).

WFcovshrink: Shrinkage Estimation of the Covariance Matrix in the WF Approach

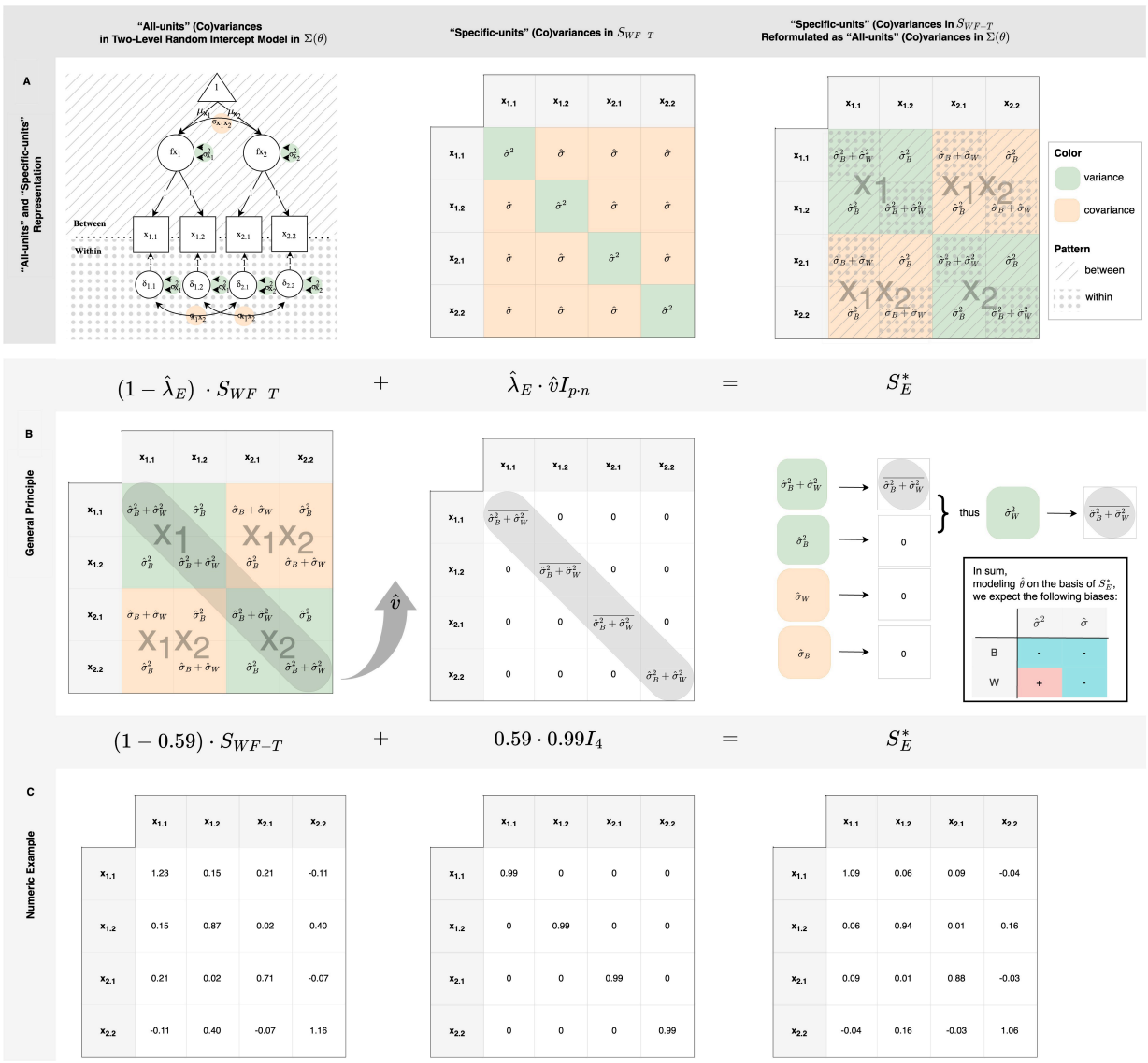
In the following, it will be outlined how WFcovshrink, or in other words, how plugging in covshrink in the WF approach, alters elements of the sample covariance matrix and model parameters. Here, the single-level represented two-level sample covariance matrix $\mathbf{S}_{WF-T}$[3] with its (co)variances of the $p \cdot n$ "specific-units" variables is replaced by the shrinkage estimate Equation 4.1 for whose estimation N is replaced by g and p by $p \cdot n$ in Equation 4.2, Equation 4.3, and Equation 4.4. The mode of action of WFcovshrink is illustrated in more detail by example of the equal target matrix and the resulting shrinkage estimate $\hat{\mathbf{S}}_E^*$ in Figure 4.2. In Panel A, the general framework of the WF approach, how $\hat{\theta}$ is modelled by $\hat{\Sigma}_{WF-T}$, is outlined. Panel B and C detail the WFcovshrink approach. In Panel B, the principal of how the shrinkage estimate $\hat{\mathbf{S}}_E^*$ alters $\hat{\theta}$ is shown, and in Panel C, an empirical example is given.

In Panel A, the general framework of the WF approach is outlined. To the left, the model specification of the two-level random intercept-only model is depicted in which in the WF approach, a restricted CFA is fitted on basis of the $p \cdot n$ "specific-units" variables in the data matrix WF-T. In the middle, the original (co)variances of the respective sample covariance matrix $\mathbf{S}_{WF-T}$ are shown. To the right, the reformulation of these (WF) $p \cdot n$ "specific-units" (co)variances as (LF) p "all-units" (co)variances that are estimated thereof in the random intercept-only model are illustrated. The (co)variances of $x_{1.1}$ and $x_{1.2}$ (see upper left green block) are used to model the variances of one common and two unique factors which correspond to the between- and within-cluster variances of x_1. Their variances contribute to the between- and within-cluster variances, whereas their covariance contributes only to the between-cluster variance via the common factor. Similarly, the (co)variances of $x_{2.1}$ and $x_{2.2}$ (see lower right green block) are used to model between- and within-cluster variances of x_2. The covariances of $x_{1.1}$ and $x_{1.2}$ with $x_{2.1}$ and $x_{2.2}$, respectively (see lower left and upper right orange blocks), are used to model the covariances of the two common factors and every n-th unique factor of each common factor which correspond to between- and within-cluster covariances of x_1 and x_2.

In Panel B, the principal of how the WFcovshrink approach operates is shown. To the left,

[3] Note that in the implementation of the linear shrinkage estimator by Touloumis (2015) in *R* in the *ShrinkCovMat* package only the unbiased MLE of the sample covariance matrix, can be used. In contrast to the normal theory derived, biased MLE (see Equation 1.13 in chapter 1), the denominator changes from g to $g-1$. Since the the default of single-level SEM in *lavaan* (i.e., the WF approach) is the biased MLE, I did an empirical investigation and found no differences for convergence and only slight differences for estimation accuracy (Walther, Hecht, & Zitzmann, 2024).

Figure 4.2: Operation Principle of Shrinkage Estimation of the Covariance Matrix in the WF Approach



$\mathbf{S}_{WF-T}$ contains (co)variances of $p \cdot n$ "specific-units" variables. In Panel A, these are reformulated as (co)variances of p "all-units" variables modelled in the random intercept-only model. Panel B introduces the principle of how the shrinkage estimate with the equal target matrix alters estimates of the random intercept-only model. In Panel C, an empirical example with number of clusters $g = 50$, cluster size $n = 2$, and number of observed variables $p = 2$ (sample characteristics), and $\sigma_B^2 = 0.05$, $\sigma_W^2 = 0.95$, $\sigma_B = 0.015$, and $\sigma_W = 0.285$ for both variables x_1 and x_2 (population characteristics), is given. The figure is taken from "Shrinking Small Sample Problems in Multilevel Structural Equation Modeling via Regularization of the Sample Covariance Matrix" by J.-K. Walther, M. Hecht, and S. Zitzmann, 2024, *Structural Equation Modeling Journal*, 32(1), p. 45-65. https://doi.org/10.1080/10705511.2024.2380919. CC BY.

the reformulated $\mathbf{S}_{WF-T}$ (as shown in Panel A rightmost) is depicted again. Its on-diagonal elements (grey bar) are averaged ($\hat{v}$) and used as the on-diagonal elements in the equal target matrix $\hat{v}\mathbf{I}_p$ ("equal variances"). In reformulated terms, $\hat{v}$ equals the grand mean of the total variances of both variables x_1 and x_2 ($\overline{\hat{\sigma}^2_B + \hat{\sigma}^2_W}$). The off-diagonal elements of $\hat{v}\mathbf{I}_p$ are set to zero. To the right, an overview of the directions in which the sample (co)variances in $\mathbf{S}_{WF-T}$ are pulled by shrinkage estimation are given. On-diagonal elements are shrunken towards $\hat{v}$. In reformulated terms, this means that total variances ($\hat{\sigma}^2_B + \hat{\sigma}^2_W$) are shrunken towards the grand mean of the total variances ($\overline{\hat{\sigma}^2_B + \hat{\sigma}^2_W}$). Off-diagonal elements, or in reformulated terms, between-cluster variances ($\hat{\sigma}^2_B$), within-cluster covariances ($\hat{\sigma}_W$), and between-cluster covariances ($\hat{\sigma}_B$), are pulled towards zero. When total variances are pulled towards the grand mean and between-cluster variances are pulled towards zero, it follows that within-cluster variances ($\hat{\sigma}^2_W$) are pulled towards the grand mean of the total variances ($\overline{\hat{\sigma}^2_B + \hat{\sigma}^2_W}$), too. The biases that are expected from utilising the shrinkage estimate for the estimation of the model parameters are depicted in the box in the right lower corner. Between-cluster variances and between- and within-cluster covariances are expected to have downward biases, whereas within-cluster variances are expected to show upward biases. Consequently, estimates of the ICC are anticipated to be more conservative than those derived by the (unregularized) WF approach.

In Panel C, an empirical example of shrinkage estimation is given. To the left, the sample covariance matrix $\mathbf{S}_{WF-T}$ can be seen. In the middle, the equal target matrix $\hat{v}\mathbf{I}_p$ is depicted in which the mean of the sample variances in $\mathbf{S}_{WF-T}$ are $\hat{v} = 0.99$, or in reformulated terms, the grand mean of the total variances $\overline{\hat{\sigma}^2_B + \hat{\sigma}^2_W}$ of x_1 and x_2. The empirical data consists of number of clusters $g = 50$, cluster size $n = 2$, and number of observed variables $p = 2$, and thus, the $cols : rows$ of the WF data matrix (WF-T), $(p \cdot n) : g$, are relatively small with $4 : 50 = 0.08$. and it lies within the minimum $cols : rows$ recommendations, $0.04 - 0.125$, derived for single-level data earlier. In line with this, the shrinkage parameter is medium-sized, $\hat{\lambda} = 0.59$ and the shrinkage estimate is composited a bit more by the target matrix than the sample covariance matrix. To the right, the resultant shrinkage estimate $\mathbf{S}^*_E$ is given. With the aid of the reformulated $\mathbf{S}_{WF-T}$ (see Panel A rightmost or Panel B leftmost) and calculus, it can be seen that the shrinkage estimates are for the most part closer to the population parameters then those of the sample covariance matrix:

- for the between-cluster variance of x_1 $\sigma^2_B = 0.05$, 0.06 in contrast to 0.15
- for the within-cluster variance of x_1 $\sigma^2_W = 0.95$, 1.03 in contrast to 1.08
- for the between-cluster variance of x_2 $\sigma^2_B = 0.05$, -0.03 in contrast to -0.07
- for the within-cluster variance of x_2 $\sigma^2_W = 0.95$, 0.91 in contrast to 0.78
- for the between-cluster covariance of x_1 and x_2 $\sigma_B = 0.015$, -0.04 in contrast to -0.11.

Only estimates of the within-cluster covariance of x_1 and x_2 $\sigma_W = 0.285$, are more far away in the shrinkage estimate (0.13) than in the sample covariance matrix (0.32).

Next, the shrinkage estimate and sample covariance matrix can be used to estimate model parameters in the WFcovshrink and the unregularized WF approach, respectively. These can be seen in Table 4.1. For all except one estimate the WFcovshrink approach yielded estimates which are closer to their population counterparts (and the one exception, the within-cluster variance of x_2, only differs on the second decimal to the estimate of the unregularized WF approach). Now the ICCs can be calculated thereof and compared with their population counterparts (0.05 for both x_1 and x_2). The estimates for x_1 are similarly off in both approaches, $0.06/(0.06+0.95) = 0.06$ in the WFcovshrink approach, and $0.15/(0.15+0.88) = 0.14$ in the unregularized WF approach. Note that WFcovshrink approach underestimated whereas the unregularized WF approach overestimated the ICC. For the estimate of x_2, the WFcovshrink approach was closer with $-0.04/(-0.04+1.02) = -0.04$ in contrast to $-0.08/(-0.08+1.01) = -0.08$ in the unregularized WF approach. Here, both approaches underestimated the ICC in the population. To sum up, with one exception the WFcovshrink yielded more accurate estimates than the unregularized WF approach. Nonetheless, this was just singular evidence. Whether this holds true for other sample sizes and target matrices remains to be put to test. This is what my second research project addressed.

Table 4.1: Model Parameter Estimates of Random Intercept-Only Model for Example Data Set

Approach	Between			Within		
	$\sigma^2_{x1} = 0.05$	$\sigma^2_{x2} = 0.05$	$\sigma_{x1x2} = 0.015$	$\sigma^2_{x1} = 0.95$	$\sigma^2_{x2} = 0.95$	$\sigma_{x1x2} = 0.285$
$\hat{\theta}_{WF}$	0.15	-0.08	-0.05	0.88	1.01	0.35
$\hat{\theta}_{WFcovshrink_E}$	0.06	-0.04	-0.02	0.95	1.02	0.15

Example data set with number of clusters $g = 50$, cluster size $n = 2$, and number of observed variables $p = 2$. The table is taken from "Shrinking Small Sample Problems in Multilevel Structural Equation Modeling via Regularization of the Sample Covariance Matrix" by J.-K. Walther, M. Hecht, and S. Zitzmann, 2024, *Structural Equation Modeling Journal*, 32(1), p. 45-65. https://doi.org/10.1080/10705511.2024.2380919. CC BY.

5 Research Project (2) Shrinking Small Sample Problems in Multilevel Structural Equation Modeling via Regularization of the Sample Covariance Matrix

Background

When the number of clusters is small in contrast to the number of observed variables at the higher level, non-convergence and inaccuracy of model parameter estimates may result. However, regularization approaches in the context of multilevel SEM, with the exception of ridge, are most often applied to the model parameters ("output" of SEM) which is only able to tackle accuracy issues. Even though particularly promising, sparse research has been done in regularizing the sample covariance matrix ("input" of SEM) so far. My second research project thoroughly examined the performance of linear shrinkage estimation of the covariance matrix as a way of handling small sample sizes and low ICCs in multilevel SEMs. More precisely, the shrinkage estimation was part of a two-stage SEM estimation approach in which the traditional sample covariance matrix was replaced by a regularized covariance matrix. All three target matrices that were proposed by Touloumis (2015) for their linear shrinkage estimator were scrutinized. All were hypothesized to foster convergence and accuracy of between-cluster parameter estimates. This two-stage approach was adopted to the WF approach to multilevel SEM (Barendse & Rosseel, 2020; Mehta & Neale, 2005; Walther, Hecht, Nagengast, & Zitzmann, 2024). This was done for the pragmatic reason that single-level SEM (which the WF approach uses) is more enhanced in its development in *lavaan* (version 0.6-15). More precisely, only here users may supply a sample covariance matrix instead of a data matrix.

Method

In order to examine the performance of the proposed two-stage approach, WFcovshrink, as compared with the standard LF and WF approaches, a simulation study was conducted. Two-level random intercept-only models for continuous, balanced data under various conditions were simulated. These conditions were varied by several factors in the simulation design

that belonged to either sample or population characteristics. The sample characteristics, factors which may be altered by the researcher, were comprised of the number of clusters ($g = 4, 10, 30, 50, 100$, the cluster size ($n = 2, 5, 10$), and the number of observed variables ($p = 2, 5, 10$). The following numbers of clusters were included: 4, 10, 30, 50 and 100. The population characteristics, factors which cannot by altered by the research design, included the between-cluster variance as indicated by the ICC ($ICC = 0.05, 0.25$) and the correlation at each level ($\rho_W = \rho_B = 0.10, 0.30$). From these, variances and covariances of the population covariance matrix at both the between- and within-cluster level were calculated. Fully crossing these factors, a simulation design with 360 conditions was yielded and for each 1.000 data sets were replicated. Evaluation criteria for performance were convergence rates, and relative bias and root mean squared error (RMSE) of (co)variance parameter estimates (i.e., random effects) of between- and within-cluster level. For the latter two, values were investigated individually for each parameter type (i.e., variances and covariances), aggregated across parameter types, and the typical conjugate parameter, the ICC.

Results

Replicating the findings from the first research project, the operationalization of small samples in combination with large numbers of observed variables ($p : g$ in LF and $(p \cdot n) : g$ in WF approaches) has only been shown to be determinative for convergence, but not for estimation accuracy. As shown in my first research project, $cols < rows$ in LF-B ($p < g$) on the LF approach and $cols \leq rows$ in WF-T ($(p \cdot n) \leq g$) in the standard WF approach had to be satisfied for models to converge. In contrast, all models, irrespective of the $cols$: $rows$ in WF-T, converged in the WFcovshrink approach with all three target matrices.

In line with the bias-variance tradeoff, the WFcovshrink approaches had increased biases compared to the unregularized approaches. The observed direction of the biases of the (co)variances at bot levels reflected the hypothesized ones: at the between-cluster level, both variances and covariances exhibited a downward bias, whereas at the within-cluster level, a downward bias in covariances and an upward bias in variances was found. As a consequence, all regularized approaches yielded downward biased estimates of the ICC.

Similarly in line with the tradeoff, these biases came often with an improvement in accuracy. The WFcovshrink approaches consistently yielded more accurate between-cluster estimates across all simulated conditions, with the most significant improvements observed under conditions of small g and small n or small ICC. In contrast, the results suggest that the effect on the accuracy of the within-cluster parameter estimates depended on the cluster size n and the correlations of the observed variables at the within-cluster levels. Unexpectedly, more accurate estimates were obtained when the cluster size was very small ($n = 2$) or when the correlations at the within cluster level was small ($\rho_W = \rho_B = 0.1$). As the cluster size or correlation at the within-cluster level increased, the estimates tended to be somewhat less accurate than those of the traditional one-step approaches. The results showed that the ICC estimates inherited advantages of both the between- and within-cluster level: in very small

cluster sizes or when the population ICCs were marginal, estimates were more accurate than the traditional methods. Overall, the accuracy gains of the WFcovshrink approaches were largest under large p, small g, very small n, and small ICC conditions. The performance of all three target matrices of WFcovshrink was very similar.

Discussion

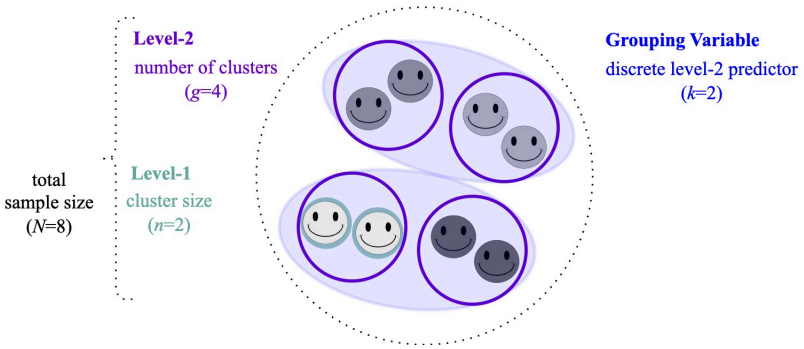
Small sample sizes at any level, but especially at the higher level, often give rise to challenges with convergence and accuracy of higher level parameter estimates in multilevel SEM. The proposed two-stage approach, WFcovshrink, was quit effective in handling these challenges. Models always converged and inaccurate parameter estimates at the between-cluster level, and for very small cluster sizes even those at the within-cluster level, were shrunken significantly. However, although the findings are encouraging, it is important to acknowledge certain limitations. Concerning the set-up of the simulation study, the variances of all observed variables at each level (and consequently, the ICCs) were identical, and only balanced cluster sizes and simple random intercept-only models were considered. This might have lead to overly optimistic results. Moreover, the application of the approach as it is, without worsening estimates of within-cluster parameters and ICC, is limited to conditions with very small cluster sizes ($n = 2$). The usage of other shrinkage estimators, or other ways of determining the shrinkage parameter merit further attention in future research. Nevertheless, this study stands out as one of the pioneering efforts to integrate shrinkage estimation of the covariance matrix in the SEM framework, especially in the multilevel modeling context.

6 Heterogeneous Variances

The assumption that data in psychology and the education sciences is homogeneous and merely representing a single population is often unrealistic. Instead, heterogeneous populations ought to be expected. 'Heterogeneous' comes from the Greek 'heterogenes', composed of 'heteros' (different) and 'genos' (give birth), which translates to "diverse in kind or nature". In a statistical sense, populations may be heterogeneous in two regards. On one hand, heterogeneity may relate to variation in characteristics in *one* population. Individuals in a population differ from one another on one or more variables of interest. For example, a population that includes people of different ages, sexes and genders, income levels, or cultural backgrounds is heterogeneous with respect to those characteristics. Variability of predictors and outcomes in linear regression analysis is a necessary requirement to estimate statistical models properly. When there is a lack of variance, non-convergence or biased estimates might result (J. J. Hox & Maas, 2001; Lüdtke et al., 2008, 2011; Muthen & Satorra, 1995; Zitzmann, 2018; Zitzmann et al., 2015). Thus, heterogeneity within a population may be seen as necessary requirement for any kind of statistical modeling. On the other hand, heterogeneity may relate to variation in distributional parameters across *multiple* populations. Every population has a mean and a variance of a certain variable and these might differ across populations. For example, the distributions of income levels differ across countries with respect to both mean and variability (Anand & Segal, 2015). Statistical models have to account for these differences in order to accurately estimate desired effects. With MLM, one accounts for (or is even interested in) the differences in (between-cluster) means. However, the variability of populations is supposed to be similar, or in other words, homogeneity of variances is assumed. This second kind of heterogeneity is the one I am interested in.

Homogeneity of variance is a standard assumption in statistics. Within the general MLM framework, it is assumed that (within-cluster) residual variances and random effects (co)variances are constant across clusters. For instance, the variability of students' performance in a math test is supposed to be equal across all sampled classes. However, multiple scenarios might be imagined in which this homogeneity assumption is likely to be violated. For example, the type of school might be related to the variability of student's performance in the math

Figure 6.1: Two-Level Multigroup Data



test: the performance of students from high schools might be less variable and more similar than that of students from grammar schools. In this case, the grouping variable would be at level-2 (see Figure 6.1), and the heterogeneous variances would be at level-1. Empirical evidence supports the claim as heterogeneity of variance is a frequently observed phenomena (Goldstein, 2005). To make this more tangible, Keselman et al. (1998) reviewed articles from prominent educational and behavioral science journals and calculated variance ratios (VR) of variables of interest. The VR indicates the ratio of the largest variance to that of the smallest variance of groups. In the case of homogeneity, the VR would be close to 1. Values substantially larger than 1 suggest heterogeneity. Keselman reported a median variance ratio (VR) of 2.25 which suggests that the group with the largest variance (e.g., grammar schools) showed variability more than twice the size compared to the group with the smallest variance (e.g., high schools). Despite this findings, a recent evaluation of reporting practices in multilevel research (Luo et al., 2021) showed that only 4.5% of studies checked the homogeneity assumption, for instance, by calculating the VR. Thus, homogeneity of variances appears to be often assumed, but less frequently observed and even checked in practice. Instead, heterogeneity of variances ought to be considered.

Whether heterogeneity of variance is considered a nuisance or an avenue depends on our research focus, but either way, it has to be accounted for. On one hand, if one is merely interested in means (e.g., of heterogeneous variances), then one has to be aware of that unaccounted heterogeneity may induce downward biases in standard errors of between-cluster parameters (F. L. Huang et al., 2022; Korendijk et al., 2008). Standard procedures to deal with these are, for example, using robust standard errors (e.g., Huber, 1967; White, 1982; see also Maas and Hox, 2004a), resampling techniques (e.g., Zitzmann et al., 2023; see also Zitzmann et al., 2024), or applying non-linear transformations to the dependent variable (e.g., Hodges, 1998). When planning a study in which one expects variances to be heterogeneous, calculating adequate sample sizes for the different populations a priori is suggested (Candel & van Breukelen, 2015). On the other hand, heterogeneous variance components might be of substantive interest in our research. For instance, analysing heterogeneous within-cluster variances in students' performance across schools can reveal differences in curricula, tracks or teaching effectiveness. To quantify these heterogeneous (within-cluster) variances, Hedeker

and Mermelstein (2007) and B. T. West et al. (2022) suggested to calculate group-specific ICCs (e.g., one ICC for high schools and one for grammar schools). These may facilitate to decide whether certain between-cluster variables (e.g., school type) are relevant for the variability of a given outcome (e.g., students' test performance) or not. These types of findings may offer a valuable increment to mean tendencies alone, help limit potential variables that give rise to heterogeneity of math achievement by exploring in which variables the two school types differ, and they would have substantial implications on predictability across groups (i.e., larger variances lead to lower predictability).

6.1 Modeling Heterogeneous Variances

Advanced statistical techniques have to be employed when heterogeneous variances are sought to be modeled. Broadly speaking, there are two main frameworks (which we are already familiar with) that are suited to model heterogeneous variances for multilevel data: multilevel models (MLM) with heterogeneous variances and multilevel multigroup SEM. MLM with heterogeneous variances (also known as HET, dispersion models, or scale location models; see e.g., Raudenbush & Bryk, 1987; Skrondal & Rabe-Hesketh, 2004) are prominent in longitudinal research in which inter-individual differences in intra-individual change is the subject of investigation. Although several approaches have been proposed, for instance, (Hedeker et al., 2008) suggested to relax the constant level-1 variance assumption by modeling the variance as a log-linear function of level-1 and level-2 predictors, (for a review see Leckie et al., 2014), the main disadvantage of MLM with heterogeneous variances is (as with general MLM) that one cannot model more than one dependent variable simultaneously and account for measurement error. However, complex relations among latent variables are the main focus of interest in the fields of psychology and the education sciences. In multilevel multigroup SEMs (ML MG SEM; see e.g., B. O. Muthén et al., 1997), one can do so. They are frequently employed to test for measurement invariance in CFAs across groups (e.g., school type, countries, measurement occasions), which is, for instance a prerequisite for cross-group comparisons such as group mean differences (Vandenberg & Lance, 2000). Multilevel multigroup SEM will be introduced more thoroughly in the following.

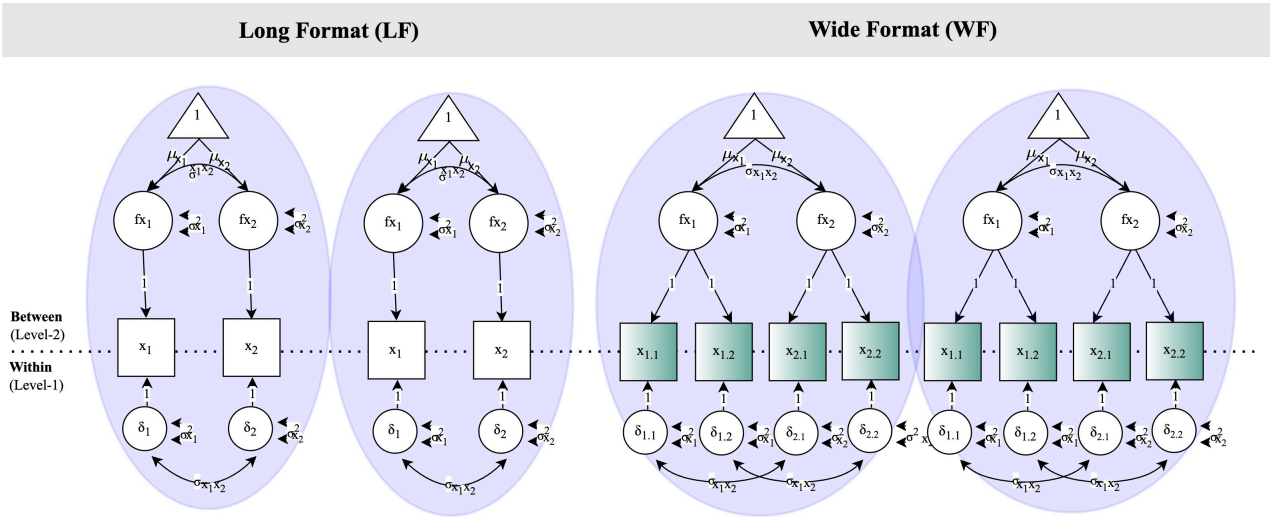
6.1.1 Multilevel Multigroup Structural Equation Modeling

Multilevel multigroup SEM integrates multilevel modeling, multigroup analysis, and structural equation modeling which enables researchers to explore complex relationships among variables at different levels while simultaneously comparing these relationships across multiple groups. For each group, a separate model is fitted or more precisely, the same model is estimated in which certain parameters are estimated freely in each group. Often times, this more complex, heterogeneous model is then compared with the simpler, homogeneous model to assess which model fits the data better. For instance, to test measurement invariance across groups, the fit of a model with freely estimated factor loadings is compared with that of

a model with equality constrained measurement models across groups.

Group-specific heterogeneous variances can be modeled at the between- or the within-cluster level. For this, the across-population invariance assumptions of λ_B and B_B, or λ_W and B_W (see Equation 1.5, Equation 1.6 and Equation 1.7 in chapter 1) are relaxed. In other words, the models at the between- or the within cluster-level are allowed to differ by group. Similarly, the grouping variable can be at any level. For instance, a between-cluster grouping variable might be school type and a within-cluster grouping variable might be gender or socio-economic status (SES). Note, however, that within-cluster grouping variables require more sophisticated modeling techniques. Standard procedures, such as ML MG SEM, do not take into account that the level-2 data is not independent between level-1 groups. Thus, the assumption of independent level-2 units is violated. Moreover, the dependency between units of different level-1 groups within the same cluster is not modelled. There may be intricate dependencies that are specific to each cluster which can be modelled with more advanced approaches such as the approach by Ryu (2014) who proposed an adapted model specification and fitting function. Hence, and because I am more interested in contextual (i.e., between-cluster) variables, my research is limited to between-cluster grouping variables. To come back to the earlier example of the random intercept-only model, in Figure 6.2, to the left, the standard multilevel multigroup SEM is illustrated, in which model parameters may vary in the two groups. In contrast, in the standard multilevel SEM, one model would be fitted across groups.

Figure 6.2: Two-Level Multigroup Models in Long Format (LF) and Wide Format (WF)



Between-cluster parameters are located above the dashed line; within-cluster parameters below. At the within-cluster level, identical parameter labels indicate equality constrains. Blue circles display the groups ($k = 2$). In this figure, both between- and within-cluster parameters across groups are equality constrained. Model specification is contingent on the data set with $n = 2$, $p = 2$ and $k = 2$.

Barendse and Rosseel (2020) and Mehta and Neale (2005) demonstrated that a multilevel structural equation model can be fitted within a single-level framework (see chapter 1). I was

motivated by similar considerations: when a multilevel SEM can be estimated as a single-level CFA, then a multilevel multigroup SEM may be estimated as a single-level multigroup CFA. In my third research project, I extended the WF approach by multigroup modeling and altered the model specification to allow for group-specific variances. Thus, a multilevel SEM for each group is specified and certain equality constrains across groups are set, see the right side in Figure 6.2. The focus of my third research project lays on models with heterogeneous within-cluster (co)variances stratified by a between-cluster predictors. However, models with different assumptions on heterogeneity at both levels as stratified by a between-cluster variable can be estimated with the proposed approach as well.

7 Research Project (3) How to Estimate Multilevel Multigroup Structural Equation Models In A Single-Level Framework in R

Background

Heterogeneity of variance is more than a statistical nuisance when variability is a focal point in research. In multilevel modeling, for instance, the inclusion of discrete variables at the between-cluster level may lead to the detection of differences between variances at the within-cluster level and these heterogeneous variances have the potential to inform research and practice. For example, when students are nested within classes, modeling the school type as group (i.e., assuming different populations for different school types) may reveal differences in variability in students' performance in a test, such as lower variance for students at high schools compared to grammar schools, which is crucial information in the field of educational effectiveness. Along the lines of 'people are variables too', the present research article demonstrates how a single-level formulation of multilevel structural equation models, the wide format approach (Barendse & Rosseel, 2020; Mehta & Neale, 2005), can be used in combination with multigroup modeling in order to obtain heterogeneous (co)variance estimates. Evidence for the proposed WFmultigroup approaches' accuracy is given by means of a simulation study and its application is showcased with an empirical illustration with the *lavaan* package in *R*.

Method

Within the simulation study, the proposed WFmultigroup approaches' accuracy and bias was scrutinized under conditions of heterogeneous within-cluster (co)variances grouped by discrete between-cluster variables. Within the simulation design, the number of clusters ($g = 200, 500, 1000$), the cluster size ($n = 2, 10, 30$), the variance ratio ($VR = 2, 5$), and the variance at the between-cluster level ($\sigma^2_B = 0.05, 0.25$) were varied. Only bivariate ($p = 2$) relations were under investigation. In each condition, two-level random intercept-only models

with heterogeneous within-cluster (co)variances for two groups ($k = 2$) were considered.

For reasons of reproducibility, a step-by-step illustration on how to estimate a multilevel multigroup SEM as a single-level restricted multigroup CFA in *lavaan* using an open access data set of the Programme for International Assessment of Student Assessment (PISA) was provided. In accordance with the simulation study, the chosen example model was a bivariate random intercept-only models with heterogeneous within-cluster (co)variances grouped by the discrete between-cluster variable country in which two countries were compared. In the main body of the article, only the code for the model specification was presented but the code for all prior steps, such as data subsetting, inspection of missing data, and multiple imputation, as well as model specifications of models with homogeneous within- and between-cluster (co)variances, heterogeneous between-cluster (co)variances, and heterogeneous within- and between-cluster (co)variances was given in the appendix.

Results

Generally, the accuracy of proposed WFmultigroup approach was good. All investigated conditions resulted in biases within the acceptable range of |10%|(L. K. Muthén & Muthén, 2002). Moreover, it is noteworthy conditions with larger VRs showed more accurate and less biased between-cluster parameter estimates. This was especially true when cluster sizes were small. For instance, when $g = 200$ and $n = 2$, under $VR = 2$ the relative RMSE was 150%, whereas under $VR = 5$ it dropped to half. At the within-cluster level, smaller numbers of clusters and smaller cluster sizes were related to less accurate estimates as well, but accuracy was better than at the between-cluster level and bias was close to zero (e.g., Depaoli & Clifton, 2015; Finch & French, 2011; J. J. Hox & Maas, 2001; J. J. Hox et al., 2010a; Lüdtke et al., 2011; Muthen & Satorra, 1995; Zitzmann et al., 2016). The ICC estimate, composed of the between- and within-cluster variance estimates, inherited both their strengths and weaknesses: smaller numbers of clusters, smaller cluster sizes, smaller between-cluster variances, and smaller VRs resulted in less accurate and more negatively biased estimates (as the between-cluster parameter estimates) but the magnitude of inaccuracy and bias was less strong (as for the within-cluster parameter estimates). To obtain accurate and unbiased estimates in bivariate random intercept-only models under the conditions observed, using a sample of $g = 100$ and $n = 10$ for every group is recommended.

In the empirical illustration, a detailed account on how a bivariate random intercept-only model of creative activities at school and growth mindset in Albania and Ireland can be modeled with the WFmultigroup approach using *lavaan* was given. Empirical results unveiled different variances but quite similar covariances: for creative activities at school 1.73 in Albania and 0.57 in Ireland ($VR = 3.04$), for growth mindset 1.68 in Albania and 0.74 in Ireland ($VR = 2.27$), and for their covariances -0.02 in Albania and 0.04 in Ireland. Of special importance for practitioners are two strategies employed in the data investigation and preparation phase. Firstly, I suggested inspecting the univariate distributions of raw data and cluster means of the observed data and of simulated data in order to aid in model specification. When the

distributions of raw data and cluster means in the groups are congruent, then homogeneous variances at both levels but also heterogeneous variances at level-2 might be the cause. When one group has more clinched distributions in the raw and group mean data, this might indicate heterogeneous level-1 variances. When these distributions are farther offset (i.e., they have different modes), then heterogeneous variances at both levels are likely the reason. Secondly, multiple imputation by chained equations was used to deal with the large amount of missings which were to a large extent introduced by unbalanced cluster sizes when formatting to WF. Imputation was done in LF and separately for each group.

Discussion

Heterogeneous variances require advanced modeling techniques to be accounted for. My third research project empirically evaluated and illustrated how heterogeneous (within-cluster) (co)variances stratified by between-cluster grouping variables can be estimated with a single-level multigroup restricted CFA. The simulation study brought evidence that the proposed WF-multigroup approach yields accurate and unbiased estimates of a bivariate random intercept-only model under conditions of moderately large numbers of clusters and cluster sizes. In the empirical illustration its implementation in *R* with the package *lavaan* was demonstrated.

Nevertheless, the WFmultigroup (just as the standard WF) approach has limitations that should be noted. Firstly, it might be inadequate when large cluster sizes and/or large numbers of groups are concerned. The minimum requirement for convergence due to how MLE is implemented in *lavaan* is $(p \cdot n_k) \leq g_k$. This requirement may be circumvented, however, by resorting to the "genuine" (LF) multilevel multigroup SEM with its less restrictive requirement, $p \leq (g_k \cdot n_k)$, or to full information maximum likelihood (FIML) estimation which uses the raw data instead of the sample covariance matrices (and thus, does not require certain matrix properties as the implementation of MLE). Secondly, when the amount of missing values is substantial and/or when the cluster sizes are highly unbalanced while the number of clusters is small, then multiple imputation of the data might be questionable. In the utilised empirical example data set, up to 72% of missing values were imputed which was justified by the large existent data ($N = 3,398$ and $g = 274$), evidence for the data being MAR, and the results of the sensitivity analysis. However, in other settings, this procedure may not be warranted. Future research ought to investigate multiple imputation of a large sample of heterogeneous, clustered data with unbalanced numbers of clusters, highly differing cluster sizes and large amounts of missings.

For future research, two further avenues might be of special interest. Firstly, more complex models that use heterogeneous within-cluster (co)variances as predictors or outcomes could be explored. For instance, D. McNeish (2021) proposed a framework to estimate location scale models as multilevel SEM in which different models for both mean (location) and variance (scale) of outcomes can be specified. The WFmultigroup approach could extend these scale location models by allowing for group-specific and time-specific heterogeneous variances. Secondly, the effect of the VR could be investigated more thoroughly. In the simulation study,

the accuracy of between-cluster parameter estimates was larger when the VR was larger which might be related to the factor analytic modeling within the WF approach. More precisely, between-cluster (co)variances are estimated as common factor (co)variances that are equality constrained across groups. Thus, for larger VR, the ratio of common (i.e., between-cluster) to unique (i.e., within-cluster) variances of the indicators (i.e., the $p \cdot n$ "observed" variables in the WF data matrix) in the second group increased as well, and thereby, the amount of communality of the indicators across both groups increased (MacCallum et al., 1999) found that larger commonalities required smaller sample sizes for factor recovery. Hence, future research could scrutinize this hypothesis and whether the effect is unique to the WFmultigroup approach.

8 Discussion

In psychology and the education sciences, observational data is often of "hierarchical" nature in which lower level units are nested within higher level units, such as students nested within schools. This dependence in the data structure has to be accounted for to safeguard accurate estimates at any level. Multilevel structural equation modeling (SEM) is a powerful tool which not only enables to model the effects at the different levels, but also to model measurement error and complex relations among variables simultaneously. Just as multilevel data can be arranged in two data formats, long (LF) and wide (WF) format, there are two frameworks to utilize multilevel SEM, the LF and the WF approach. The LF approach is the "genuine" multilevel SEM approach and the WF approach uses a single-level restricted confirmatory factor analysis (CFA) framework. In the current version of the SEM package *lavaan* (0.6-15; Rosseel, 2012) in the free statistical software *R*, the WF approach has more options to be modified because it uses a single-level framework that has undergone enhanced development with regard to methodology and programming. Within the scope of the present thesis, I suggested a (1) more precise definition of "small samples" for multilevel data and validated the equivalent performance of the wide format (WF) approach to multilevel SEM here, and I extended the WF approach under conditions of (2) small samples and (3) heterogeneous variances with focus on convergence and estimation accuracy.

8.1 Summary

Within my first research project, I answered the questions *"What are small samples in multilevel data?"* and *"Is the performance of the LF and WF approaches equivalent in small samples?"*. For this, I built on research from single-level analysis. Several scholars expanded information on samples sizes by information on the number of observed variables. This large p, small N conditions in single-level analysis were transferred to large p, small g conditions in multilevel analysis. Moreover, the significance of the variance at the between-cluster level, usually conceptualized by the Intraclass Correlation (ICC), was taken into account. Smaller ICCs demand for larger numbers of clusters g. Thus, small sample conditions in multilevel analysis have been contoured as "large p, small g, small ICC" conditions. The number of observed variables

p and the number of clusters g are sample characteristics that can be changed in the study design whereas the ICC belongs to population characteristics that can only be compensated for by the sample characteristics. I noted that $p : N$ is equatable to $cols : rows$ of the data matrix in single-level data. In multilevel data, though, the $cols : rows$ differs depending on the data format, long format (LF) or wide format (WF), and in the former, it additionally differs depending on the type of data matrix, within- or between-cluster. I found that in LF, the between-cluster data matrix with $p : g$ is of primary relevance for convergence whereas in WF, the relevant ratio is $(p) : g$. I extended existing empirical evidence and scrutinized whether the LF and WF approaches are equivalent under "large p, small g, small ICC" conditions in which these ratios differ substantially. Hence, the contributions of the first research project are recommendations for minimum sample sizes for converging multilevel models given certain numbers of observed variables, ICCs and the implementation of MLE in *lavaan* and to know when the LF and WF approaches are applicable, $p > g$ in LF and $(p) \leq g$ in WF. Under conditions in which both approaches converge, namely $(p) \leq g$, their estimation accuracy has been found to be comparably.

Within my second research project, I explored the question *"How to deal with small samples?"*. For this, I adapted the WF approach by regularization of the covariance matrix in order to obtain converging and more accurate models. More specifically, a two-stage estimation approach, WFcovshrink, was proposed in which the traditional sample covariance matrix was substituted by a shrinkage estimate. The linear shrinkage estimator by Touloumis (2015) was utilised for this purpose. The performance of WFcovshrink in contrast to the traditional, unregularized LF and WF approaches was explored under conditions of large p, small g, small ICC and under less problematic conditions. The approach always achieved convergence by overcoming the $cols \leq rows$ restriction imposed by the implementation of MLE in *lavaan*. Moreover, it resulted in more accurate between-cluster parameter estimates throughout, and under conditions of small cluster sizes even in more accurate within-cluster parameter estimates. However, as the within-cluster parameter estimates got less accurate when the cluster sizes were increased, the proposed WFcovshrink approach is, so far, only advisable for conditions with very small cluster sizes. Notwithstanding, the research project was a pioneering success in applying regularization to covariance matrices in multilevel SEM to mitigate estimation problems related to small sample sizes.

Within my third research project, the question *"How to model heterogeneous variances?"* was investigated. I proposed an approach that enables to estimate heterogeneous variances in multigroup multilevel SEM within a single-level framework. On that account, the single-level restricted CFA (WF approach) was extended to a single-level multigroup restricted CFA (WFmultigroup). The proposed approach was backed up by a simulation study and its application in *R* with the package *lavaan* illustrated by means of an openly accessible empirical example data set. The focus was on heterogeneous within-cluster (co)variances stratified by between-cluster grouping variables. The simulation study yielded evidence for the approaches' accuracy and unbiasedness under conditions of moderately large numbers of clusters and cluster sizes. The empirical illustration included suggestions for model specification (i.e.,

at which levels heterogeneity of variances ought to be modeled) by means of descriptive statistics and handling of missing data by means of multiple imputation. The research project provided new avenues for teaching staff and researchers when dealing with group-specific heterogeneous within-cluster (co)variances in multilevel analysis.

Open Science principles such as replication, accessibility and transparency were adhered to in all of my research projects. Throughout all three research endeavours, several existing findings could be replicated: (a) the number of clusters g is the more crucial sample size for convergence and estimation accuracy of between-cluster parameters in multilevel data (Afshartous, 1995; Clarke, 2008; J. J. Hox & Maas, 2002; Kreft & Yoon, 1994; Maas & Hox, 2004b; Mok, 1995), (b) smaller ICCs are related to lower convergence rates and less accurate estimates of random between-cluster parameters (Finch & French, 2011; J. J. Hox & Maas, 2001; J. J. Hox et al., 2010b; Lüdtke et al., 2011; D. M. McNeish & Stapleton, 2016; Meuleman & Billiet, 2009; Muthen & Satorra, 1995; Zitzmann, 2018; Zitzmann, Wagner, et al., 2022), and (c) estimates of random between-cluster parameters are more affected by inaccuracy than random within-cluster parameters (Finch & French, 2011; J. J. Hox & Maas, 2001; J. J. Hox et al., 2010a; Lüdtke et al., 2008, 2011; D. McNeish & Stapleton, 2016; D. M. McNeish & Stapleton, 2016; Meuleman & Billiet, 2009; Shin & Raudenbush, 2010; Stegmueller, 2013; Zitzmann, 2018; Zitzmann et al., 2015). To guarantee accessibility and transparency of workflows, analysis was done in the free and open-source software *R*, code snippets were appended in the articles and the complete scripts for the simulation studies and related figures and tables uploaded to Github (https://github.com/demianJK?tab=repositories), and open access data was used for empirical illustration. Moreover, all three articles were published open access through funding of the library of the University of Tübingen.

8.2 Limitations

Now that the contributions of my research projects are summarised, their limitations have to be acknowledged in a similar fashion. The following points of criticism mainly concern the designs of the simulation studies. These were very similar throughout all three research endeavours. Thus, generalization beyond the range of the conditions studied should be undertaken with caution. Including new conditions and investigating open questions might offer some interesting avenues for future research.

8.2.1 Random Intercept-Only Models

The most simple two-level models, namely random intercept-only models, were examined. In these, the covariance structure of the sample covariance matrices at both levels in LF, $\hat{\mathbf{\Sigma}}_B$ and $\hat{\mathbf{\Sigma}}_W$, and the cluster means at the between-cluster level (i.e., random intercepts) were estimated. While this is the standard model that is estimated to assess whether MLM is necessary, more complex models that include directed paths among variables are commonly utilised. More complex models and model misspecification, for instance, have been related

to increased occurrence of non-convergence and non-positive definite covariance matrix errors (see e.g., Peugh, 2010). It would be, for example, interesting to examine whether the WFcovshrink approach can minimize these issues in more complex and misspecified models as results from the standard model suggest. Moreover, it might be sensible to reconsider the equivalence of the LF and WF approaches when more complex models and model misspecification are concerned. For example, Barendse and Rosseel (2020) considered a factor model which was either correctly specified or not and found that the WF approach yielded slightly less biased between-cluster factor loadings (in balanced data) than the LF approach when the model was misspecified (−3% in contrast to −9% relative bias). The authors attribute this to an disadvantage of the marginal maximum likelihood (MML) estimator in the LF approach compared to the similarly performing diagonally weighted least squares (DWLS) and pairwise maximum likelihood (PML) estimators in the WF approach. However, when contrasting the DWLS estimator in both approaches, then a slight difference in relative bias of |1%| in favour of the WF approach hints at potential differences in estimation accuracy. While this difference may seem trivial, it might become more substantial under conditions more problematic for estimation, such as large p, small g, small ICC conditions. One may further hypothesize potential reasons why the WF approach might have an advantage. For instance, models specified in the WF approach have more degrees of freedom because the number of observed parameters is higher than the number of freely estimated parameters (the latter is the same in LF an WF model specifications). The sample covariance matrix $\hat{\mathbf{\Sigma}}_{WF-T}$ has the dimensions $(p \times n) \times (p \times n)$ and every observed variable p is modelled as common factor from the n specific-units variables whose (co)variances are equality constrained. Thus, there might be some kind of "model size effect" concerned when comparing the accuracy of the LF and WF approaches. That the accuracy of the WF approach has not been found to be superior in my research despite the differences in degrees of freedom might have been attributable to the standard models employed. In the literature, the model size effect has been conceptualized in different ways (see e.g., Shi et al., 2018), amongst others as the number of observed variables p, the number of freely estimated parameters q, and a function of both, the degrees of freedom df. However, as Shi et al. (2018) note, these factors covary: when the number of observed variables in a model is changed, so do the number of freely estimated parameters and the degrees of freedom. Thus, in future research, a sensible research design in which it can be ruled out which factor is of importance, need to be constructed. An alternative explanation for the less biased between-cluster factor loadings could be the single-level factor modeling of multilevel parameters in the WF approach. In the misspecified condition, a non-zero effect on the first item at the between-cluster level was left out in modeling. In the LF approach, the item (i.e., observed variable) is manifest whereas it is modelled as common factor by n specific-units variables in the WF approach. Thus, bias might be pushed into the unique factors (i.e., within-cluster components) and related parameters. Unfortunately, Barendse and Rosseel (2020) did not report bias of within-cluster parameter estimates in details and thus, the hypothesis remains to be put to test by future research.

8.2.2 Random Effects

Only random effects of the random intercept-only model, that is, variances and covariances of observed variables at both levels, were investigated. Fixed effects, in the case of the model under investigation, cluster means at the between-cluster level, were not examined. The focus was on random parameter estimates because most evidence found small sample bias in these (e.g., Finch & French, 2011; J. J. Hox & Maas, 2001; J. J. Hox et al., 2010a; Lüdtke et al., 2008, 2011; D. McNeish & Stapleton, 2016; D. M. McNeish & Stapleton, 2016; Meuleman & Billiet, 2009; Shin & Raudenbush, 2010; Stegmueller, 2013; Zitzmann, 2018; Zitzmann et al., 2015). However, some studies also reported biases in observed cluster means (Shin & Raudenbush, 2010; Zitzmann, 2018; Zitzmann et al., 2015) and other fixed effect parameter estimates in more complex models, such as regression coefficients, factor loadings and residual variances in measurement models, contextual effects, and cross-level interaction effects (Lüdtke et al., 2008, 2011; D. McNeish & Stapleton, 2016; Meuleman & Billiet, 2009; Stegmueller, 2013). Equivalence of the LF and WF approaches with respect to fixed effects might deserve a closer look. As outlined in the last subsection, slight differences in biases in factor loadings at the between-cluster level have been found under conditions of large samples (Barendse & Rosseel, 2020). Future research might consider exploring these under conditions less favourable for estimation accuracy, such as large p, small g, small ICC conditions. Beyond that, a potential extension of the WFcovshrink approach to fixed effects to improve accuracy might be conceivable. This will be discussed later.

8.2.3 Balanced Data

Analysis was restricted to balanced data. This made for a more simple simulation design at the price of neglecting an issue that many practitioners face when collecting data. It is generally known that unbalanced data is costlier to deal with (see e.g., McDonald & Goldstein, 1989), and highly unbalanced data may be especially problematic in the WF approach (see also Barendse & Rosseel, 2020, who pointed this out). For instance, as has been seen in the empirical illustration in my third research project, a substantial spread of cluster sizes, varying from 1 to 45, introduced a non-negligible amount of missing data in the WF data matrix (WF-T). I dealt with this issue by means of multiple imputation after checking the data and the missing mechanisms but this and other procedures have to be validated by future research. Alternatively, one might deal with unbalanced data by case-varying parameters, for instance, different factor loadings for every cluster as suggested by Bauer (2003). However, the major drawback of this procedure is that one cannot formally test the fit of these models by using conventional likelihood-ratio statistic because they are not summarized by a single covariance matrix and mean vector (Raudenbush, 2001). Moreover, when unbalancedness is substantial, models may not converge without compensating for the missings, as has been noted in my third research project. Notably, even with smaller percentages of unbalanced data, undesirable consequences might arise. For instance, Barendse and Rosseel (2020) found higher relative bias (i.e., about 10%) in the standard errors of factor loadings for the WF approach in the

conditions with cluster sizes varying from 3 to 9. However, my research focused on matters of estimation problems such as convergence and accuracy but not on inference. Thus, future research should scrutinize the effects of unbalanced data on different performance outcomes and consider potential remedies to deal with these, especially for the WF approach.

8.2.4 Estimation Problems

The scope of all three research projects within my thesis was limited to estimation problems such as non-convergence and inaccuracy of estimates. For reasons of practical applicability, however, other performance measures, such as standard errors, would have to be validated for the WF approach and its extensions as well. At least concerning the equivalence of the LF and WF approaches, there is some evidence that their standard errors are comparable. For instance, Barendse and Rosseel (2020), who inspected the performance of both the LF and WF approaches for a one factor model, used robust standard errors (in M*plus*) for both approaches and found their biases to be similarly low across all conditions. However, the investigated conditions did not include large p, small g, small ICC conditions, but only moderately large samples and small numbers of observed variables in which the relevant data matrices in both approaches (LF-B and WF-T) had $cols << rows$. Future research might put these under investigation. More importantly, however, the accuracy of the standard errors of the WFcovshrink approach need to be assessed. For instance, McQuitty (1997) scrutinized the standard errors obtained when ridge, another type of regularization of the sample covariance matrix similar to shrinkage estimation, was applied and found them to be biased upwards. The author considered two possible (intertwined) explanations. Firstly, larger ridge constants caused the on-diagonal elements (i.e., variances) to be relatively more important than the off-diagonal elements (i.e., covariances). When the the ridged sample covariance matrix was used to fit a CFA, the inter-item correlations were underestimated and standard errors of the factor loadings were inflated. Secondly, as the on-diagonal elements (i.e., variances) were increased by the (larger) ridge constants, this simultaneously increased the standard errors of the variances of the observed variables. In the case of linear shrinkage estimation, it is not necessarily the same case that variances are increased while covariances are decreased as the direction of the shrinkage depends on the relation of the elements of the sample covariance matrix and the target matrix. Nevertheless, it is unlikely that standard errors are unbiased when elements of the sample covariance matrix are altered in the shrinkage estimate. Hence, future research should scrutinize this matter and potentially propose corrections for standard errors when shrinkage estimates are used in SEM.

8.3 Avenues for Future Research

In the following, further avenues for future research are suggested. These ideas build on the most vital findings, questions, and hypothesis that emerged within the three research projects. They are intended to challenge the reader and stir more questions than to provide ready-made

instructions.

8.3.1 Cols:Rows Conceptualization

The generalization of the $p : N$ effect from single-level to multilevel analysis, the $cols : rows$ conceptualization, which was studied within my first and second research project, indicated that the ratio is of primary relevance for convergence in multilevel SEM with MLE as implemented in *lavaan*. Effects on convergence deserve to be further studied in multilevel analysis

Firstly, it might be of special interest for practitioners to explore minimum $cols : rows$ requirements for other implementations of MLE, such as full information likelihood (FIML) estimation, and whether these are transferable to other statistical software, such as M*plus*. For instance, McCoach et al. (2018) compared five common software solutions for MLM and found differences in terms of convergence rates. This suggest that substantial differences in the implementation of multilevel SEM software might also exist.

Secondly, it remains to be answered whether the minimum $cols : rows$ suggestions are valid when more complex models, such as measurement models, are investigated. Multicollinearity (i.e., large covariances/correlations among observed variables), but also the opposite direction, weak factor loadings and communality (i.e., small covariances/correlations among observed variables), as well as a small number of indicators for common factors have been shown to make non-convergence more likely (Boomsma, 1985; MacCallum et al., 1999; Yuan & Chan, 2008). In addition, fitting large models is inherently more unstable (Breiman, 1996), and an extended conceptualization that incorporates model complexity might prove useful. For example, a large area of research in single-level analysis was dedicated to the "model size effect" which, amongst others, has been conceptualized as the number of freely estimated parameters q (Shi et al., 2018) and the number of participants per freely estimated parameter $N : q$ (Herzog et al., 2007; Jackson, 2001, 2003). In the context of SEM, where first a sample covariance matrix and then model parameters are estimated, one could hypothesize that $p : N$ is more suited for the sample covariance matrix whereas $p : q$ is more suited for model parameters. However, these considerations remain to be put to test in the context of multilevel SEM in future research.

Thirdly, including $cols : rows$ in optimal design research might prove useful. Commonly, optimization focuses on maximization of power but as Hecht et al. (2023) emphasized, in very small samples, simulated power might be biased by estimation problems and non-convergence. Including an optimization criterion that at least rules out non-convergence due to software requirements, such as the $cols \leq rows$ requirement in *lavaan*, might be of practical significance. For instance, in the case of the LF approach, substituting or amending the design parameters 'number of observed variables' and 'number of clusters' by $p : g$ or including $p < g$ as constraint for the optimization may foster excluding research designs that would give rise to non-converging models. For instance, the user-friendly Shiny App

OptDynMo proposed by Hecht et al. (2023) for dynamic longitudinal models could be amended by the ratio $p < T$ (in which T are the number of time points). However, first, research would need to address whether the requirement holds true for (software of) dynamic panel models as well. In addition, as the optimal design calculation is based on the likelihood ratio statistics proposed by Satorra and Saris (1985), it could be adapted for any SEM, including multilevel SEM, in which $p < g$ for the LF approach and $(p \cdot n) < g$ for the WF approach might be included. Moreover, optimal design research could benefit from including the decomposition of the between- and within-cluster variance, the ICC. My research replicated earlier findings that smaller ICCs are related to larger percentages of non-convergence and inaccuracy of between-cluster estimates of random parameters (Finch & French, 2011; J. J. Hox & Maas, 2001; J. J. Hox et al., 2010b; Lüdtke et al., 2011; D. M. McNeish & Stapleton, 2016; Meuleman & Billiet, 2009; Muthen & Satorra, 1995; Zitzmann, 2018; Zitzmann, Wagner, et al., 2022), and thus, expected ICC values should be taken into account when planning studies. This claim is supported by other scholars who suggested to include the ICC in study planning because it is pivotal for convergence and estimation accuracy (Hedges & Hedberg, 2013).

8.3.2 Further Application of the WF Approach

The empirical evidence for the equivalence of the LF and WF approaches has been extended to large p, small g, small ICC conditions, and potential advantages of the WF approach have been hypothesized. Thus, one might consider whether the WF approach might be utilised for other kinds of hierarchical data. Just as clustered data, such as students nested within classrooms, longitudinal data has a hierarchical structure. Here, time points (T, level-1 units) are nested within individuals (N, level-2 units). Thus, there is a large potential of applying the WF approach to the analysis of longitudinal data. For instance, growth curve models (GCM) can be estimated within a MLM or SEM framework (see e.g. Bryk & Raudenbush, 1987; D. R. Rogosa & Willett, 1985; J. D. Singer, 2003; Steele, 2008), and these kind of models are called multilevel growth models or MLM of change. Whereas GCM in a MLM framework uses the data in WF, in a SEM framework, the model uses the data in LF and thus, might be reformulated into the WF approach. GCM in a MLM framework has the advantage that individual variation in the timing of measurements, otherwise known as unequally spaced measurements, unbalanced data, attrition or non-monotone patterns of missingness, can be dealt with better (Steele, 2008). In a MLM framework, cluster sizes are not required to be equal but in a SEM framework, the missing data must be dealt with, for instance, by using full information maximum likelihood (FIML) estimation or multiple imputation. However, when variables of interest are not directly observable, then GCM in a SEM framework has to be employed and reformulating this traditional LF approach into the WF approach might be beneficial for certain scenarios.

First some general introduction (see e.g., Bryk & Raudenbush, 1987): multilevel growth models concern the analysis of change over time. As in the modeling of clustered data, variance is decomposed into within and between parts. With longitudinal data, however, in which level-1

units are time points and level-2 units are individuals, there is within- and between-*individual* variation. The former concerns how individuals vary relative to their own average and the latter concerns how individuals change compared to others. More specifically, at the within-individual level, the outcome of individuals at time points are modeled as a function of a systematic growth trajectory, which is often represented by polynomial terms, plus residual error. The (co)variance structure assumed for the growth model depends on the functional form assumed for the individual growth model and on the amount of variance and covariance among the individual growth parameters. For the residual error (i.e., level-1), most often a simple structure is assumed which puts that errors are independently and normally distributed with a mean of zero and constant variance. At the between-individual level (i.e., level-2), the variation of growth parameters is modeled.

Using the WF approach potentially offers another advantage for the modeling of the error term. In traditional MLM, unaccounted heteroscedasticity in residual error (i.e., level-1) variance may result in less efficient parameter estimates (for regression models cf. White, 1980). To account for this, extensions of the traditional MLM, such as MLM with heterogeneous variances (see e.g., Hoffman, 2007; Leckie et al., 2014; Raudenbush & Bryk, 1987) may be employed. Alternatively, the WF approach might offer a feasible alternative for modeling time-specific residual (co)variances. Here, equality constraints across time points can be easily relaxed. The data matrix in WF (WF-T) has the dimensions $g \times (p \cdot n)$ in the case of clustered data in which g is the number of clusters and n is the cluster size and $N \times (p \cdot T)$ in the case of longitudinal data in which N is the number of individuals and T is the number of time points. In the modeling of the WF approach, equality constraints across the n or respectively T specific-units variables of each observed variable p are commonly set. These represent the assumption of homoscedasticity of (co)variances of the $p \cdot n$ (clustered data) or respectively $p \cdot T$ (longitudinal data) specific-units variables. Heteroscedasticity, in contrast, indicates that the variance of a variable changes in dependence of another variable, such as the index in a cluster (clustered data) or time (longitudinal data). For clustered data in WF, in which, for instance, $x_{1.1}$ is the observed variable x_1 for every 1^{st} unit in each cluster, the homoscedasticity assumption is reasonable. The index (i.e., ordering) of the units in the clusters ought to be arbitrary. Thus, there is no reason to assume heteroscedasticity. In contrast, for longitudinal data, in which, for instance, $x_{1.1}$ is the observed variable x_1 for every 1^{st} time point of each individual, the homoscedasticity assumption may not be reasonable. Here, homoscedasticity is a form of the standard stationarity assumption in longitudinal modeling which indicates that statistical properties, such as mean and variance, of a variable do not change over time. Heteroscedasticity, or in other words, non-stationarity of (co)variances, in contrast would allow (co)variances at different time points to differ, which is reasonable in designs in which interventions or environmental shocks are present. Future research ought to scrutinize the utility of the WF approach for multilevel growth models and compare its performance with existing methods for modeling time-specific (co)variances.

8.3.3 Extensions and Further Applications of the WFcovshrink Approach

Extensions: Other Shrinkage Estimators and Regularization of Means

The WFcovshrink approach has been shown to be successful in improving convergence rates and improving accuracy but still, it might be extended in two ways, by customizing the regularization of the (co)variances and by including regularization of the cluster means, in order to enhance accuracy properties even more. Both have been demonstrated to mitigate inaccuracies in between-cluster parameter estimates. However, whereas the former only affects estimates of random parameters, the latter has been shown to influence estimates of both fixed and random parameters.

Although the WFcovshrink approach has proved to result in more accurate between- and even within-cluster parameter estimates in large p, small g, small ICC conditions, evidence suggest that regarding the within-cluster parameter estimates, effects hold true only for conditions with very small cluster sizes. The employed linear shrinkage approach was designed for single-level data and not for single-level representations of multilevel data. Thus, the target matrix and its shrinkage parameter were not optimal under this conditions. Although several target matrices have been proposed by scholars, it may be preferable to customize these anew as single-level represented multilevel covariance matrices have not been considered in the literature before. In the employed shrinkage estimator, for instance, on-diagonal values (i.e., sum of between- and within-cluster variances) were shrunken towards a non-zero value whereas off-diagonal values (i.e., between-cluster variances and within- and between-cluster covariances) were shrunken towards zero which increased the risk of inflated, less accurate within-cluster variances. Once could prevent this, for instance, by shrinking the between-cluster variances less intensely than the between- and within-cluster covariances in the off-diagonal. This would suggest a target matrix with different values in the off-diagonal. It might be furthermore fruitful to consider non-linear shrinkage estimators in which different shrinkage intensities are applied to the elements of the covariance matrix. This way, between-cluster variances and between- and within-cluster covariances in the off-diagonal could be shrunken differently. For example, when between-cluster variances are shrunken less (compared to the other elements in the single-level represented multilevel covariance matrix), then within-cluster variances might not be inflated severely (because these are estimated by subtracting between-cluster variances in the off-diagonal from the between- and within-cluster variances in the on-diagonal). Moreover, the closed-form solutions of the shrinkage parameter ought to include information on the sample sizes at both levels. In the way the shrinkage estimator was employed, only the sample size at the higher level, the number of clusters, was considered but not the cluster size. This might have led to inflated shrinkage parameters under conditions of smaller numbers of clusters but larger cluster sizes (although the former has been shown to be the more important quantity in multilevel data when it comes to estimation accuracy). Implementing these suggestions might yield estimators that are more appropriate for small ICC but future research should similarly consider target matrices that are more suited for diverse properties of multilevel population covariance matrices, such as

larger ICCs. In other research areas of when other kinds of data, such as longitudinal data, is concerned, the ratio of level-2 to total variance may increase, making a target matrix with (close to) zero off-diagonal values inappropriate. Moreover, the conceptualization of level-2 may be important as findings from Gulliford et al. (1999) suggest. They investigated ICCs of lifestyle risk factors and health outcomes and found their ICCs to be contingent on the conceptualization of level-2. At the district health authority level, ICCs were found to be below 0.01, at the postal code sector level, below 0.05, and at the household level, they fell in the range of $0.0-0.3$. This implies that the "lower" the chosen level-2 conceptualization, the larger the ICC. This claim is supported by Hedges and Hedberg (2013) who found that the higher the level, the less variability is expected. In sum, prior research on ICCs of the variables of interest should also inform the choice of the target matrix.

Examining whether an extended WFcovshrink approach can minimize inaccuracies in fixed effects might be another important avenue for future research. In the basic random-intercept models employed throughout my research, fixed effects were only cluster means (which were not scrutinized within my research) but in more complex models, this may include parameters such as regression weights and factor loadings. It has been noted that unreliability of cluster means yields biases in slopes and variance components, particularly contextual effects, and this is especially true when ICCs are small (Grilli & Rampichini, 2011; Lüdtke et al., 2011). Regularization of the sample covariance matrix does not affect the cluster means, but methods to attenuate the measurement error in the cluster means, such as the doubly latent approach by Lüdtke et al. (2011), have been proposed. This approach was designed to tackle two sources of error, namely sampling (through sampling of items) and measurement error (through sampling of persons). Observed variables are decomposed into different components, indicator-specific means, within- and between-cluster factor loadings and indicator-specific error sources at the within- and between-cluster level, respectively, which are modelled separately. This correction has been shown to outperform traditional MLE when the number of clusters is small. Thus, a combination of the doubly latent approach with regularization of the covariance matrix might create substantial synergy in accuracies in large p, small g, small ICC conditions which future research could look into.

Further Applications: Adaptation for Longitudinal Data

Just as the WF approach, the WFcovshrink approach might be borrowed for longitudinal data analysis. One may either utilise the complete WFcovshrink approach, a single-level restricted CFA with a two-stage estimation approach that replaces the sample covariance matrix by a shrinkage estimate, or only supply the shrinkage estimate to another modeling approach that uses the data in WF.

The complete WFcovshrink approach may be adopted to the WF approach reformulation of GCMs in a SEM framework suggested earlier. When the number of individuals (i.e., level-2 units) is sparse, then replacing the sample covariance matrix by a shrinkage estimate might mitigate potential estimation problems. For instance, Laird et al. (1987) noted that when

the model parameter matrix is not positive semi-definite (psd means that eigenvalues are positive or zero), which may emerge because of sparse data, then models may not converge. It would be interesting to scrutinize whether providing a shrunken covariance matrix with a well-behaved eigenstructure may solve this issues. Moreover, the shrinkage estimate may lead to more accurate model parameter estimates just as in multilevel modeling in my second research project. For GCMs, evidence suggests that at least 50 level-2 units with traditional MLE (Maas & Hox, 2005) or 20 level-2 units with Bayesian estimation (J. J. Hox et al., 2012) ought to be used. Accuracy of these methods ought to be compared to those of the WFcovshrink approach. However, the results of my research also suggested that the linear shrinkage estimator proposed by Touloumis (2015), which was plugged in the WFcovshrink approach, may not be the best choice when more than a few time points are included. When $T = 2$, residual (i.e., level-1) variances were more accurate. Although GCMs usually contain few time points T (i.e., level-1 units), less accurate residual variances may result when $T = 5$. Reducing the number of time points to two, though, is no viable option for many reasons, for instance, as these provide an inadequate basis for modeling change (Bryk & Weisberg, 1977; D. Rogosa et al., 1982). Moreover, we have to keep in mind that generalization from WFcovshrink used for multilevel SEM to WFcovshrink used for GCM in a SEM framework has to be done with caution as only the sample covariance matrix is the same but the modeling approaches differ. More suitable shrinkage estimators for longitudinal data, such as the non-parametric one proposed by Wu (2003), may be sought after.

Regularization of the WF sample covariance matrix may offer an avenue for other longitudinal modeling approaches that use the data in WF (without necessarily using the WF approach to multilevel SEM). Here, for instance, GCMs in the MLM framework might benefit from substituting the sample covariance matrix by a shrinkage estimate when the number of individuals (i.e., level-2 units) is small and the model parameter matrix is not psd. Another avenue for the application of shrinkage estimation of the WF covariance matrix might be dynamic panel models, such as autoregressive cross-lagged (ARCL) panel models. These are prone to face problems with convergence under conditions of high autocorrelations (i.e., multicollinearity) and heteroscedasticity because a higher risk of distorted eigenstructure is introduced (Dufour & King, 1991). For example, high autocorrelations mark high off-diagonal values (i.e., variances) and when these are large compared to on-diagonal values (i.e., variances), then eigenvalues are spread out more widely making undesirable properties such as non-positive definiteness more likely. To foster convergence, several procedures have been suggested, for instance, setting equality constraints in autoregressive and cross-lagged effects across time points (Orth et al., 2021) or using the heteroscedasticity- and autocorrelation-consistent covariance matrix estimator proposed by K. D. West (1997). However, instead of adapting the model side, we may adapt the sample side by means of regularization of the sample covariance matrix. When applying the linear shrinkage estimator by Touloumis (2015), on one hand, certain desirable shifts take place. Autoregressive equivalents in the sample covariance matrix, such as the covariance of variable x_1 at time points one and two, $x_{1.1}$ and $x_{1.2}$, are in the off-diagonal and thus, are shrunken towards zero. Similarly, cross-lagged

equivalents in the sample covariance matrix, such as the covariance of variable x_1 at time point one and variable x_2 at time point two, $x_{1.1}$ and $x_{2.2}$, are in the off-diagonal and also are shrunken towards zero. On the other hand, the same cautionary note as fin the paragraph before applies here. The utilised shrinkage estimator may be rather advisable for very small numbers of level-1 units (i.e., time points) which may restrict the application in ARCL models and other shrinkage estimators for longitudinal data may be more advisable. In sum, the application of shrinkage estimation of the covariance matrix in WF appears promising but the proposed ideas remain to be assessed in future research.

8.4 Resumé

The research done in the present thesis validated and extended multilevel SEM in the WF approach for the frequently occurring estimation problems non-convergence and inaccuracy (especially of between-cluster parameter estimates) under conditions of small samples and heterogeneous variances. Tribute to Open Science principles was paid: analysis was done in the free and open-source software *R* and made public, and existing findings could be replicated. In conclusion, the research contributed to providing sound methodology which is amongst others needed for decision making in educational politics. May future research harness the utility of the WF approach and its extensions even further.

A Appendix

A.1 List of Publications

Publications Relevant to This Thesis

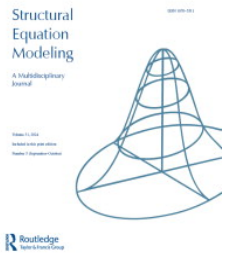
1. **Walther, J. K.**, Hecht, M., Nagengast, B., Zitzmann, S. (2024). To Be Long or To Be Wide: How Data Format Influences Convergence and Estimation Accuracy in Multilevel Structural Equation Modeling. Structural Equation Modeling: A Multidisciplinary Journal, 31(5), 759–774. https://doi.org/10.1080/10705511.2024.2320050
2. **Walther, J. K.**, Hecht, M., Zitzmann, S. (2024). Shrinking Small Sample Problems in Multilevel Structural Equation Modeling via Regularization of the Sample Covariance Matrix. Structural Equation Modeling: A Multidisciplinary Journal, 32(1), 45-65. https://doi.org/10.1080/10705511.2024.2380919
3. **Walther, J. K.**, Hecht, M., Nagengast, B., Zitzmann, S. (2025). Multilevel Multigroup Structural Equation Modeling In A Single-Level Framework. Structural Equation Modeling: A Multidisciplinary Journal, 1–32. https://doi.org/10.1080/10705511.2024.2434596

Other Publications Not Relevant to This Thesis

1. Zitzmann, S., **Walther, J. K.**, Hecht, M., Nagengast, B. (2022). What Is the Maximum Likelihood Estimate When the Initial Solution to the Optimization Problem Is Inadmissible? The Case of Negatively Estimated Variances. Psych, 4(3), 343–356. https://doi.org/10.3390/psych4030029
2. Hecht, M., **Walther, J. K.** *(Shared First Authorship)*, Arnold, M., Zitzmann, S. (2023). Finding the Optimal Number of Persons (N) and Time Points (T) for Maximal Power in Dynamic Longitudinal Models Given a Fixed Budget. Structural Equation Modeling: A Multidisciplinary Journal, 31(3), 535–551. https://doi.org/10.1080/10705511.2023.2230520

A.2 Publications

In the following, the three publications relevant to this thesis are attached.



Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hsem20

To Be Long or To Be Wide: How Data Format Influences Convergence and Estimation Accuracy in Multilevel Structural Equation Modeling

Julia-Kim Walther, Martin Hecht, Benjamin Nagengast & Steffen Zitzmann

To cite this article: Julia-Kim Walther, Martin Hecht, Benjamin Nagengast & Steffen Zitzmann (2024) To Be Long or To Be Wide: How Data Format Influences Convergence and Estimation Accuracy in Multilevel Structural Equation Modeling, Structural Equation Modeling: A Multidisciplinary Journal, 31:5, 759-774, DOI: 10.1080/10705511.2024.2320050

To link to this article: https://doi.org/10.1080/10705511.2024.2320050

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC

View supplementary material

Published online: 27 Mar 2024.

Submit your article to this journal

Article views: 891

View related articles

View Crossmark data

Citing articles: 2 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=hsem20

STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
2024, VOL. 31, NO. 5, 759–774
https://doi.org/10.1080/10705511.2024.2320050

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

To Be Long or To Be Wide: How Data Format Influences Convergence and Estimation Accuracy in Multilevel Structural Equation Modeling

Julia-Kim Walther[a], Martin Hecht[b], Benjamin Nagengast[a,c] and Steffen Zitzmann[d]

[a]University of Tübingen; [b]Helmut Schmidt University; [c]Korea University; [d]Medical School Hamburg

ABSTRACT

A two-level data set can be structured in either long format (LF) or wide format (WF), and both have corresponding SEM approaches for estimating multilevel models. Intuitively, one might expect these approaches to perform similarly. However, the two data formats yield data matrices with different numbers of columns and rows, and their *cols* : *rows* is related to the magnitude of eigenvalue bias in sample covariance matrices. Previous studies have shown similar performance for both approaches, but they were limited to settings where $cols \ll rows$ in both data formats. We conducted a Monte Carlo study to investigate whether varying *cols* : *rows* result in differing performances. Specifically, we examined the $p : N$ (*cols* : *rows*) effect on convergence and estimation accuracy in multilevel settings. Our findings suggest that (1) the LF approach is more likely to achieve convergence, but for the models that converged in both, (2) the LF and WF approach yield similar estimation accuracy, which is related to (3) differential *cols* : *rows* effects in both approaches, and (4) smaller *ICC* values lead to less accurate between-group parameter estimates.

KEYWORDS

Accuracy; convergence; eigenvalues; multilevel SEM

Covariance matrices are an integral part of multivariate statistics (T. W. Anderson, 2003). In structural equation modeling (SEM), the covariance matrix of the observed variables can be expressed as a function of the model parameters. To estimate a specified model, the sample covariance matrix of the observed variables is estimated first. Then, the model parameters are estimated by applying a fitting function, whose objective is to minimize the discrepancy between the sample covariance matrix and the model-implied covariance matrix. The model parameters are estimated by algorithms that use matrix algebra. Thus, the properties of involved matrices, such as the sample covariance matrix, are of importance.

Among matrix properties, eigenvalues appear to be the most important ones. Eigenvalues are a special set of scalars of a matrix and their characteristics inform us about other important matrix properties. At least one zero eigenvalue indicates singularity (which implies non-invertibility). At least one non-positive eigenvalue shows non-positive definiteness. At least one eigenvalue close to zero or largely spread out extrema result in a large condition number (which equals the ratio of the largest to the smallest eigenvalue; see Golub & Van Loan, 2013). These unfavorable matrix properties, which can easily be detected by the eigenvalues, are linked to lower convergence rates and less accurate estimations (e.g., Boomsma, 1985; Golub & Van Loan, 2013; Hill & Thompson, 1978; Kelley, 1995; Lange et al., 1999; Zitzmann, 2018; Zitzmann et al., 2015).

Commonly, the sample covariance matrix is estimated by the standard maximum likelihood (ML) estimator, which yields biased eigenvalue estimates (Stein, 1956). It has been shown analytically and empirically that the magnitude of the bias of the sample eigenvalues is comparable to the ratio of the number of observed variables to the sample size $p : N$ (Arruda & Bentler, 2017; Dempster, 1972; Hayashi et al., 2018; Schäfer & Strimmer, 2005; Stein, 1956, 1975)[1]. More precisely, "biased" means that small eigenvalues are pushed downwards, and large eigenvalues are pushed upwards compared to their population counterpart. As a consequence, the ratio of the largest to the smallest eigenvalue gets larger, and it is more likely that at least one is zero or negative (even when all population eigenvalues are positive). As previously mentioned, this makes lower convergence rates and less accurate estimations more likely.

As a matter of fact, large p, small N settings, are a frequent state of affairs in the social sciences, which has received much attention (for an overview related to SEM, see, e.g., Deng et al., 2018; Marcoulides et al., 2023). Some

CONTACT Julia-Kim Walther julia-kim.walther@uni-tuebingen.de Hector Research Institute of Education Sciences and Psychology, University of Tübingen, 72072 Tübingen, Germany.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10705511.2024.2320050.

[1]Note that Hayashi et al. (2018) pointed out that when p is negligibly small relative to N, the magnitude of the bias is $1 : N$. However, when p is not negligibly small relative to N, which is commonly the case, the magnitude of the bias is $p : N$.

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

research in this area focused on effects of $p : N$ on various outcomes.[2] For example, Yuan and Chan (2008) found that both convergence rate and accuracy of estimation decrease with increasing $p : N$. Further, they proposed the "ridge method", which adds a constant of size $p : N$ to the on-diagonal of the sample covariance matrix to improve its eigenvalue structure. This resulted in higher convergence rates and more efficient model parameter estimates. In the context of model fit and inference, Yuan et al. (2018) found that larger $p : N$ led to more biased likelihood ratio statistic, and Xing and Yuan (2017) proposed corrections for biased model fits based on these statistics. Further, some researchers suggested that the fundamental problem with test statistics in large p, small N settings might be the biased sample eigenvalues (Arruda & Bentler, 2017; Huang & Bentler, 2015; Yuan & Bentler, 2017). In sum, previous research supports the notion that $p : N$ is an important factor in convergence, model estimation, model fit and inference, and that the effect of $p : N$ is connected to the eigenvalue bias.

It is interesting to note that in the investigated single level settings, the number of observed variables p corresponds to the number of columns, and the sample size N corresponds to the number of rows of the data matrix by which the sample covariance matrix is estimated. Thus, $p : N$ can be expressed as *cols* : *rows*. This emphasizes the relation between data matrix, sample covariance matrix, its eigenvalues, and model performance. However, many data sets in the social sciences have a hierarchical data structure. With two-level data (e.g., students nested within classes, clients nested within therapists, employees nested within teams), the number of observed variables and sample size has no one-to-one relation to columns and rows. Firstly, the two levels of data, level-1 (e.g., students, clients, employees) and level-2 (e.g., classes, therapists, teams), have different numbers of observed variables and sample sizes. Secondly, the same data set can be arranged in two different formats, long format (LF) and wide format (WF), that result in data matrices with inherently different dimensions. LF leads to longer data matrices (i.e., more rows), whereas WF leads to wider data matrices (i.e., more columns). However, to the best of our knowledge, possible equivalents of the $p : N$ (*cols* : *rows*) effect in multilevel settings have not been investigated before.

There are multiple approaches to estimate a multilevel SEM. We focus on the multilevel SEM approach by Muthén (1990, 1994) which uses the data in LF, and the single level restricted CFA approach by Barendse and Rosseel (2020) and Mehta and Neale (2005) which uses the data in WF. Both approaches are readily available in the lavaan package (Rosseel, 2012) for the statistical software *R*. Whereas past research demonstrated the analytical and empirical equivalence of both approaches (Barendse & Rosseel, 2020; Mehta & Neale, 2005), the empirical evidence included only settings with a small number of observed variables at both levels, small level-1 sample sizes, and, most notably, large level-2 sample size. In other words, the data matrices in both data formats had *cols* $\ll$ *rows*, implying small biases of the eigenvalues. Little is known about how the approaches perform when *cols* : *rows* differs across data matrices of both formats. We are interested in examining two (intertwined) effects on convergence and model estimation here: (a) the effect of the data format, because the data format inherently leads to different *cols* : *rows*, and (b) the effect of *cols* : *rows* in each data format. Whereas the effect of the data format answers which data format (approach) to use, the effect of *cols* : *rows* answers which *cols* : *rows* to aim at with our study design. We conducted a Monte Carlo study to investigate these matters. The present article is organized as follows. First, we introduce the data matrices and sample covariance matrices in each data format. Second, we discuss the two SEM approaches. Finally, we present the results of the present study, discuss its implications, and provide suggestions for future research.

1. How Data Format Influences the Representation of Data Set and (Co)variances

We first clarify the terms data set, data format, data matrices, and (sample) covariance matrices. To this end, we refer to the example in Figure 1. The *data set* (see Panel A) specifies the data that we collect in a given setting. Relevant information are the sample sizes and the number of observed variables at both levels. At level-2, we have the number of groups g. At level-1, we have the group size n, and the total sample size $N = g \cdot n$. For means of simplification, we restrict our example to the same observed variables p at both levels. In other words, we only look at level-2 variables that are aggregates of level-1 variables. Models that include the same variable at both levels are often referred to as contextual analysis models (e.g., Boyd & Iversen, 1979; Raudenbush & Bryk, 2002). In our example, we observed two groups ($g = 2$) with two units each ($n = 2$), resulting in a total sample size of $N = g \cdot n = 4$ units. For every level-1 unit we observed two variables ($p = 2$), x_1 and x_2, which we aggregate to obtain level-2 variables. The *data matrix* (see Panel B) specifies the data set in matrix form. It has two dimensions, columns and rows. The dimensions of the data matrix depend on the sample sizes and the number of observed variables, and, importantly, on the *data format*. We can arrange our data set either in LF or in WF. The data format further determines which sample covariance matrices are estimated. The *sample covariance matrix* contains the variances and covariances of the observed variables (i.e., of the columns of the respective data matrix). Note that the sample covariance matrix ($\mathbf{S}$) is not always the unbiased population covariance matrix estimator ($\hat{\mathbf{\Sigma}}$). This is why we refer to the entirety of covariance matrices of the observed variables as *covariance matrices* (see Panel C).

[2]Note that a large strain of research in large p, small N settings focused on the "model size" effect, which has been conceptualized in many different ways: as the effect of the number of observed variables p (Shi et al., 2018, 2019), the number of freely estimated parameters q (Shi et al., 2018), the number of participants per (freely estimated) parameter $N : q$ (Herzog et al., 2007; Jackson, 2001, 2003), or a function of both p and q, the degrees of freedom df (Herzog et al., 2007; Shi et al., 2018). Because it has not been conceptualized as $p : N$, we do not go into detail about the model size effect here.

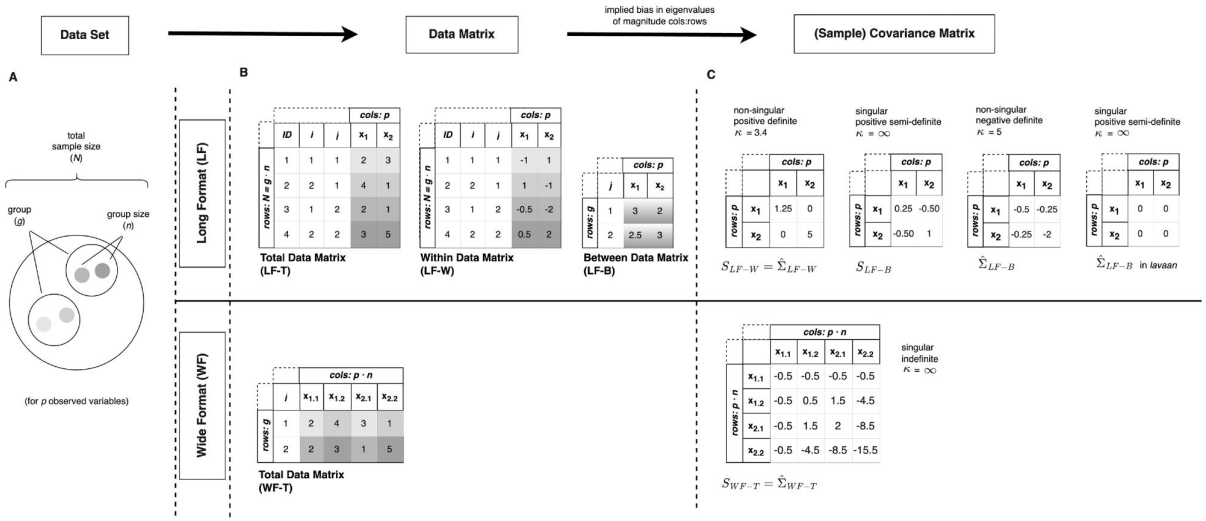


Figure 1. Representation of data set and (Co)variances. Example data set with number of groups $g = 2$, group size $n = 2$, and number of observed variables $p = 2$. In the WF approach, p is split into n specific-units variables (e.g., $x_{1.2}$ is x_1 for every 2nd unit in the group). We have coding variables for each unit (ID), units within groups (i), and groups (j). The grey shades indicate different units. κ = condition number. $\hat{\Sigma}_{LF-B}$ in lavaan = $\hat{\Sigma}_{LF-B}$ with negative variances and related covariances set to 0.

Next we consider the data matrices and covariance matrices more closely. Note that the procedures we present in the following correspond to the multilevel modeling approaches by Muthén (1990, 1994) and Mehta and Neale (2005) and their implementation in lavaan that we investigated in our study, and may not generalize to other approaches.

1.1. Long Format: Multilevel Representation of Data Set and (Co)variances

When we arrange the data set in long format (LF), the total raw data matrix (LF-T) has p columns and $N = g \cdot n$ rows. We decompose the total data matrix (LF-T) into the between-group data matrix (LF-B) and the within-group data matrix (LF-W), see the upper part of Panel B of Figure 1, to separate the total (co)variance of each variable(s) into between-group and within-group components. For this, the group means (on level-2) are estimated and subtracted from the value of their respective level-1 units for every observed variable. The group means constitute the data matrix LF-B with p columns and g rows. The deviations from these group means constitute the data matrix LF-W with p columns and $g \cdot n$ rows. Both LF-W and LF-B include all units for every p, resulting in *all-units* (co)variances[3]. One sample covariance matrix for each level using LF-W and LF-B is estimated.

The ML estimators for the sample covariance matrices of the two levels, the pooled within-group estimator in Equation (1) and the between-group estimator in Equation (2) (Muthén, 1990, 1994), read:

$$\mathbf{S}_{LF-W} = \frac{1}{N-g}\sum_{j=1}^{g}\sum_{i=1}^{n}(\mathbf{X}_{ij} - \bar{\mathbf{X}}_j)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_j)^T, \quad (1)$$

$$\mathbf{S}_{LF-B} = \frac{n}{g-1}\sum_{j=1}^{g}(\bar{\mathbf{X}}_j - \bar{\mathbf{X}})(\bar{\mathbf{X}}_j - \bar{\mathbf{X}})^T, \quad (2)$$

with $j = 1, \ldots, g$ groups and $i = 1, \ldots, n$ units per group. $\mathbf{X}_{ij}$ denotes the raw data in LF (i.e., LF-T), $\bar{\mathbf{X}}_j$ denotes group mean estimates (i.e., LF-B), $(\mathbf{X}_{ij} - \bar{\mathbf{X}}_j)$ denotes unit-wise deviations from these group mean estimates (i.e., LF-W), $\bar{\mathbf{X}}$ denotes a row vector with grand mean estimates, and ^{T} is the matrix transpose.

For the within-group level, the unbiased ML estimator of the population covariance matrix is the pooled within-group sample covariance matrix, $\mathbf{S}_{LF-W} = \hat{\mathbf{\Sigma}}_{LF-W}$. However, for the between-group level, the unbiased ML estimator of the population covariance matrix is a function of the sample covariance matrices of both levels (Muthén, 1990, 1994):

$$\hat{\mathbf{\Sigma}}_{LF-B} = \frac{1}{c}(\mathbf{S}_{LF-B} - \mathbf{S}_{LF-W}), \quad (3)$$

where c denotes the common group size, and in the case of balanced data, $c = n$. The *cols* : *rows* implied biases in the eigenvalues are $p : (g \cdot n)$ for $\mathbf{S}_{LF-W} = \hat{\mathbf{\Sigma}}_{LF-W}$ and $p : g$ for $\mathbf{S}_{LF-B}$ and $\hat{\mathbf{\Sigma}}_{LF-B}$. However, note that $\mathbf{S}_{LF-B}$ and $\hat{\mathbf{\Sigma}}_{LF-B}$ are influenced by more factors than p and g, which we will discuss in the next section, and that in lavaan, negative variances and related covariances in $\hat{\mathbf{\Sigma}}_{LF-B}$ are set to zero[4]. Thus, the eigenvalues of these LF-B covariance matrices differ from what we would expect from *cols* : *rows* alone. All LF covariance matrices are shown in the upper part of Panel C of Figure 1.

[3]Barendse and Rosseel (2020) introduced the term *unit-specific* with regards to the WF covariance matrix. To avoid confusion with within-group effects and make the difference between LF and WF more succinct, we use the terms *specific-units* (co)variances for the WF covariance matrix, and *all-units* (co)variances for the LF covariance matrices instead.

[4]We found no information on this procedure in the reference manual or presentations but discovered it by chance. The potential reason is discussed later.

1.1.1. Why Between-Group Covariance Matrices Have Even More Problems with Unfavorable Matrix Properties

As outlined, LF-B has larger *cols* : *rows* than LF-W, more specifically, $p : g$ in contrast to $p : (g \cdot n)$, which implies larger bias in the eigenvalues of $\mathbf{S}_{LF-B}$ than of $\mathbf{S}_{LF-W}$. Hence, non-singularity, non-positive definiteness, and high condition numbers are more likely. However, the eigenvalues of $\mathbf{S}_{LF-B}$ and $\hat{\mathbf{\Sigma}}_{LF-B}$ are likely not only influenced by the *cols* : *rows* of LF-B. For example, the probability of being non-positive definite increases not only with increasing p and decreasing g (which links to the *cols* : *rows*) but also with decreasing group sizes n and intraclass correlation[5] (*ICC*; Bhargava & Disch, 1982; Hill & Thompson, 1978; Searle et al., 1992). Further, $\hat{\mathbf{\Sigma}}_{LF-B}$ is estimated as the difference of $\mathbf{S}_{LF-B}$ and $\mathbf{S}_{LF-W}$, see Equation (3), and the subtraction often results in a non-positive definite covariance matrix even when both sample covariance matrices are positive definite (Bhargava & Disch, 1982; Hill & Thompson, 1978). Non-positive definiteness indicates zero or negative eigenvalues, which is related to singularity and high condition numbers. In sum, the fact that the occurrence of non-positive definiteness is related to larger p, and smaller g, n, and *ICC*, suggests that the exact bias term of $\mathbf{S}_{LF-B}$ and $\hat{\mathbf{\Sigma}}_{LF-B}$ may be different than $p : g$. We will elaborate more on this assumption shortly.

1.2. Wide Format: Single Level Representation of Data Set and (Co)variances

When we arrange the data set in WF, the total raw data matrix (WF-T) has $p \cdot n$ columns and g rows, see the lower part of Panel B of Figure 1. Here we do not separate the total (co)variance in within- and between-group components. However, each variable p is split into n *specific-units* variables, or as Mehta and Neale (2005, p.1) put it, "people [n] are variables too." One sample covariance matrix is estimated from WF-T.

We estimate the sample covariance matrix with the estimator for single level data. The single level represented two-level sample covariance matrix with specific-units (co)variances reads:

$$\mathbf{S}_{WF-T} = \frac{1}{g}\sum_{j=1}^{g}(\mathbf{X}.\mathbf{i}_j - \overline{\mathbf{X}.\mathbf{i}})(\mathbf{X}.\mathbf{i}_j - \overline{\mathbf{X}.\mathbf{i}})^T, \tag{4}$$

where $\mathbf{X}.\mathbf{i}$ denotes the raw data in WF (i.e., WF-T) and $\overline{\mathbf{X}.\mathbf{i}}$ denotes a row vector with grand mean estimates. $\mathbf{S}_{WF-T}$ is the so-called biased ML estimator[6] for $\mathbf{\Sigma}_{WF-T}$. The *cols* : *rows* implied bias in the eigenvalues of $\mathbf{S}_{WF-T}$ is $(p \cdot n) : g$. We see $\mathbf{S}_{WF-T}$ in the lower part of Panel C of Figure 1.

[5]Hox et al. (2017) define the *ICC* as amount of between-group variance out of the total variance (i.e., sum of between- and within-group variance), $\sigma_B^2/(\sigma_B^2 + \sigma_W^2)$.

[6]In the unbiased ML estimator, the denominator would be $g - 1$. However, we focus on the biased estimator because it is the default in lavaan (see Rosseel et al., 2023, reference manual p.81 accessed on 16 September 2023, lav_matrix_cov function).

1.3. Summary and Comparison

All LF and WF covariance matrices are estimated by ML. The *cols* : *rows* of LF-W, LF-B, and WF-T, are $p : (g \cdot n), p : g$, and $(p \cdot n) : g$. Thus, the order of the magnitude of *cols* : *rows* is LF-W < LF-B < WF-T. However, it is unclear whether the *cols* : *rows* bias in sample eigenvalues in single level settings is directly applicable to multilevel settings. As we pointed out earlier, evidence suggests that the between-group covariance matrices are not only influenced by the *cols* : *rows* of LF-B. However, no study investigated the exact bias term of the eigenvalues of LF and WF covariance matrices before. Within the scope of this study, we are not interested in the exact term of the bias. Rather we use *cols* : *rows* as a proxy of the eigenvalue bias, because we assume that it is the main influence on the bias. Let us support this assumption by looking at the other matrix properties of the LF and WF covariance matrices, which are shown in Panel C of Figure 1. $\hat{\mathbf{\Sigma}}_{LF-W}$ is non-singular, positive definite, and has a small condition number. We see that $\hat{\mathbf{\Sigma}}_{LF-B}$ has only negative variances. Thus, in lavaan all elements are set to zero which results in a singular, positive-semi definite matrix with an infinite condition number. $\hat{\mathbf{\Sigma}}_{WF-T}$ is singular, indefinite, and has an infinite condition number. Note that negative variances in $\hat{\mathbf{\Sigma}}_{WF-T}$ are kept which might contribute to inadmissible values ("Heywood cases"). In sum, the LF covariance matrices are assumed to have a smaller eigenvalue bias than the WF covariance matrix.

2. How Data Format Influences Multilevel SEM Approaches

It has been shown that ML estimation in multilevel SEM (Muthén, 1990, 1994) is analytically and in certain settings empirically equivalent to single level restricted confirmatory factor analysis (CFA) models (Barendse & Rosseel, 2020; Mehta & Neale, 2005). The multilevel SEM approach uses the data matrix in LF. The single level approach uses the data matrix in WF. More specifically, Mehta and Neale (2005) demonstrated analytical equivalence of the LF and WF approach for continuous, unbalanced data with full information ML (FIML) estimation, and provided an empirical example which used the WF approach (in Mplus). Barendse and Rosseel (2020) demonstrated empirical equivalence of the LF and WF approach for discrete, balanced and unbalanced data with marginal ML (MML) and pairwise ML (PML) estimation in a Monte Carlo study (in lavaan and Mplus). With very small group sizes $n = 3$ (but large g) the WF approach even resulted in somewhat less biased parameter estimates. However, all empirical data sets contained only small p and n, and large g (and thus large N), which resulted in *cols* $\ll$ *rows* in all involved data matrices, and no significant differences in the *cols* : *rows* of the data matrices of both formats. In contrast, different *cols* : *rows* imply different magnitudes of eigenvalue bias. Thus, data sets that result in strongly different *cols* : *rows* in both formats imply differences in convergence and estimation accuracy. In lavaan, the multilevel SEM (LF) approach is yet implemented only for continuous data. Thus, we restricted our analysis to continuous, balanced data and standard

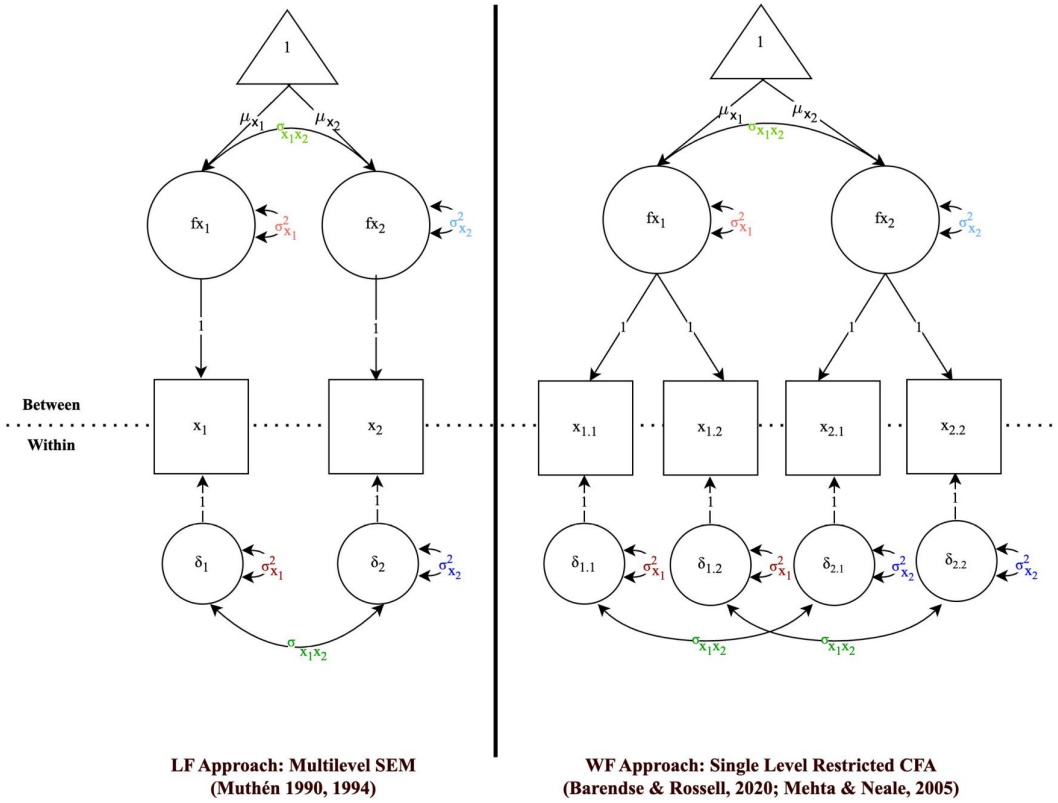


Figure 2. Example model specification in the LF and WF approach for a contextual intercept-only model. Example data set with group size $n = 2$, and number of observed variables $p = 2$. In the WF approach, p is split into n specific-units variables (e.g., $x_{1.2}$ is x_1 for every 2nd unit in the group; see also Figure 1), and identical parameter labels indicate equality constraints in the model. Across both approaches, the same parameters have the same color. Across both levels, matching parameters have similar color. Affiliation of parameters to the between- and within-group level is indicated by location above and below the dashed line.

normal-theory ML estimation for both the LF and WF approach[7]. We do not want to go into detail about the model estimation and fitting functions here. The interested reader is referred to Mehta and Neale (2005). Instead, we want to underline how the data format influences the modeling approaches, more specifically, the model specifications and minimum data set requirements.

2.1. Model Specification

Whereas the same multilevel model can be estimated in the LF and WF approaches, their model specifications differ. For reasons of simplicity, we only consider so called "intercept-only models" (e.g., Hox et al., 2017; Raudenbush & Bryk, 2002) of all-units (co)variances. In other words, we model the p variables covariance structure for the within- and between-group level which is equivalent to the LF covariance matrices. In the LF approach, we have the same model specification at each level. The (co)variances at both levels are modeled as (co)variances of the p observed variables (from $\hat{\boldsymbol{\Sigma}}_{LF-W}$ and $\hat{\boldsymbol{\Sigma}}_{LF-B}$, respectively). In the WF approach, the model specification differs at both levels. In contrast to the LF approach, all parameters are modeled as (co)variances of latent factors of the observed $p \times n$ specific-units variables (from $\hat{\boldsymbol{\Sigma}}_{WF-T}$). To estimate the all-units (co)variances, the between-group parameters are modeled as common factors, with factor loadings set to 1, and the within-group parameters are modeled as unique factors, with equality constraints added to (co)variances of the unique factors of the same variable p. Note that with the equality constraints, homoscedasticity of specific-units variables is modeled. Figure 2 depicts the model specifications for a contextual intercept-only model for both approaches, using the earlier example data set. Note that in both approaches, the mean structure at the between-group level has to be included in order to discern the total (co)variance into within- and between-group parts. In the "genuine" multilevel approach (LF), this is done implicitly. In the restricted single level CFA approach (WF), we have to add it. Example R code for the specification and estimation of both models is presented in the Online Supplemental Material.

2.2. Software Requirements

We want to consider two requirements of lavaan because they are connected to the (sample) covariance matrices and their matrix properties. The first requirement concerns both approaches; the second concerns only the LF approach.

[7]The default estimator in both single level and multilevel SEM in lavaan is the standard normal-theory ML estimator (see Rosseel et al., 2023, reference manual p.53f; see also https://lavaan.ugent.be/tutorial/est.html, accessed on 16 September 2023). Note also that both approaches use the quasi-Newton algorithm (Jak et al., 2021).

2.2.1. Minimum Data Set

One requirement of lavaan is that the data matrices we give as input, LF-T and WF-T, respectively, need to have *cols* $\leq$ *rows*. This requirement is likely based in the fact that sample covariance matrices have undesirable matrix properties in these settings. For example, when *cols* $>$ *rows*, at least one sample eigenvalue becomes zero and the sample covariance matrix turns singular (e.g., Duncan et al., 1997; Gorsuch, 1983; Wothke, 1993). Whereas the ML estimators in the LF and WF approach does not require the covariance matrices to be non-singular (i.e., invertible), singular matrices are non-positive definite and result in an infinite condition number, which might have a negative impact on convergence and estimation accuracy. Due to the different *cols* : *rows* of LF-T and WF-T, $p : (g \cdot n)$ and $(p \cdot n) : g$, the minimum data set requirements for both approaches differ. Two points are noteworthy here. Firstly, larger group sizes n are advantageous for LF-T, but disadvantageous for WF-T. Secondly, there are settings where we can only use the LF approach because $p < (g \cdot n)$ (in LF-T) is more easily satisfied than $(p \cdot n) < g$ (in WF-T). Our example data set with $g = 2$, $n = 2$, and $p = 2$, would result in 2 : 4 in LF-T and 4 : 2 in WF-T. Thus, we could only use the LF approach to analyze it. Note that the *cols* $\leq$ *rows* requirement generally supports the importance of considering *cols* : *rows*.

2.2.2. Definiteness of $\hat{\boldsymbol{\Sigma}}_{LF-B}$

In the LF approach, there is one further requirement. Model estimation with the quasi-Newton algorithm (the default) fails with an error when $\hat{\boldsymbol{\Sigma}}_{LF-B}$ is negative definite (i.e., has only negative eigenvalues; see e.g., Rosseel, 2018). This might be the reason why lavaan sets negative variances and related covariances in $\hat{\boldsymbol{\Sigma}}_{LF-B}$ to 0, because this seems to prevent that all eigenvalues are negative. Thus, we do not have to worry about this requirement, but keep in mind, that $\hat{\boldsymbol{\Sigma}}_{LF-B}$ can be altered.

3. This Study

The aim of this study is to investigate the equivalent of the $p : N$ (*cols* : *rows*) effect on convergence and estimation accuracy in multilevel SEM. To this end, we scrutinize the LF and WF approach in settings which result in different *cols* : *rows* of the data matrices of both data formats. We use *cols* : *rows* as a proxy of the eigenvalue bias of the LF and WF covariance matrices. Two (intertwined) effects are considered: (a) the effect of the data format, with its inherently different *cols* : *rows*, and (b) the effect of *cols* : *rows* in each data format. By investigating (a) the effect of the data format, we want to learn which data format (and related approach) to prefer in a given setting. The *same* setting results in inherently different *cols* : *rows* in the data matrices of both approaches. For example, the setting $g = 10$, $n = 2$, and $p = 2$ results in 2 : 20 in LF-W, 2 : 10 in LF-B, and 4 : 10 in WF-T. Unless all data matrices have *cols* $\ll$ *rows*, we assume that the LF approach, which has inherently smaller *cols* : *rows* in its data matrices, outperforms the WF approach. We are also interested in (b) the effect of *cols* : *rows* in each data format, to gather insight into optimal study design. On that account, *different* settings that result in the same *cols* : *rows* in both data formats are investigated. Out of the multiple data matrices in LF, we focus on LF-B. We do so because LF-B does not have to satisfy *cols* $<$ *rows* in lavaan, which implies more possible variation in *cols* : *rows*, and thus has the largest (i.e., most problematic) *cols* : *rows* among the data matrices in LF. To continue the example, to have 4 : 10 in LF-B (like in WF-T), we need another setting, for example where $p = 4$ (and $g = 10$ and $n = 2$ stay the same).

3.1. Method

In the following, we outline the data generation and evaluation criteria of the Monte Carlo study. As pointed out earlier, we only considered intercept-only models at the within- and between-group level. As a consequence, the data-generating model equals the data analysis model. All computations were performed on an AMD Ryzen Threadripper PRO 3975WX 32-cores (3.50 GHz) CPU on a Windows 10 (Version 20H2) platform. Data generation and analysis was conducted using R version 4.3.1 (R Core Team, 2023)[8]. The R code for data, analysis, tables, and figures is available at https://github.com/demianJK/LF_WF_SEM.

3.1.1. Data Generation

For the data generation, we drew from a multivariate normal distribution with population means fixed to zero and varying population covariance matrices and sample characteristics. Data generation was done in LF and separately for level-1 and level-2 data (see also example LF model in Figure 2). With respect to the *population covariance matrices*, we varied the number of observed variables p and the *ICC*. The range of the ICC was informed by common values in the social sciences (Gulliford et al., 1999). The total variance of each observed variable, $\sigma^2_B + \sigma^2_W$, was constrained to be 1. Hence, $\sigma^2_B = ICC$ and $\sigma^2_W = 1 - \sigma^2_B$. The variances for each level were then used to compute the covariances, respectively, with a fixed correlation of .30. Note that these variances and covariances are also the parameters in our intercept-only models. Applying a fully-crossed design for p and the *ICC*, we arrived at 12 different population conditions. With respect to the *sample characteristics*, we varied the total sample size N, the numbers of units per group n, and the number of groups g. We first set N and n, computed $g = N/n$ and then excluded all conditions where $N \geq$

[8]We used the following R packages: *broom* version 1.0.5 (Robinson et al., 2023), *car* version 3.1-2 (Fox et al., 2023), *cowplot* version 1.1.1 (Wilke, 2020), *dplyr* version 1.1.0 (Wickham, Chang, et al., 2023; Wickham, François, et al., 2023), *effectsize* version 0.8.3 (Ben-Shachar et al., 2023), *ggbreak* version 0.1.1 (Yu & Xu, 2022), *ggplot2* version 3.4.1 (Wickham, Chang, et al., 2023; Wickham, François, et al., 2023), *gridExtra* version 2.3 (Auguie & Antonov, 2017), *huxtable* version 5.5.2 (Hugh-Jones, 2022), *lavaan* version 0.6–14 (Rosseel et al., 2023), *lsr* version 0.5.2 (Navarro, 2021), *MASS* version 7.3-58.2 (Ripley et al., 2023), *patchwork* version 1.1.2 (Pedersen, 2022), *stringr* version 1.5.0 (Wickham & R Studio, 2022) *tidyr* version 1.3.0 (Wickham et al., 2022), and *xlsx* version 0.6.5 (Dragulescu & Arendt, 2020).

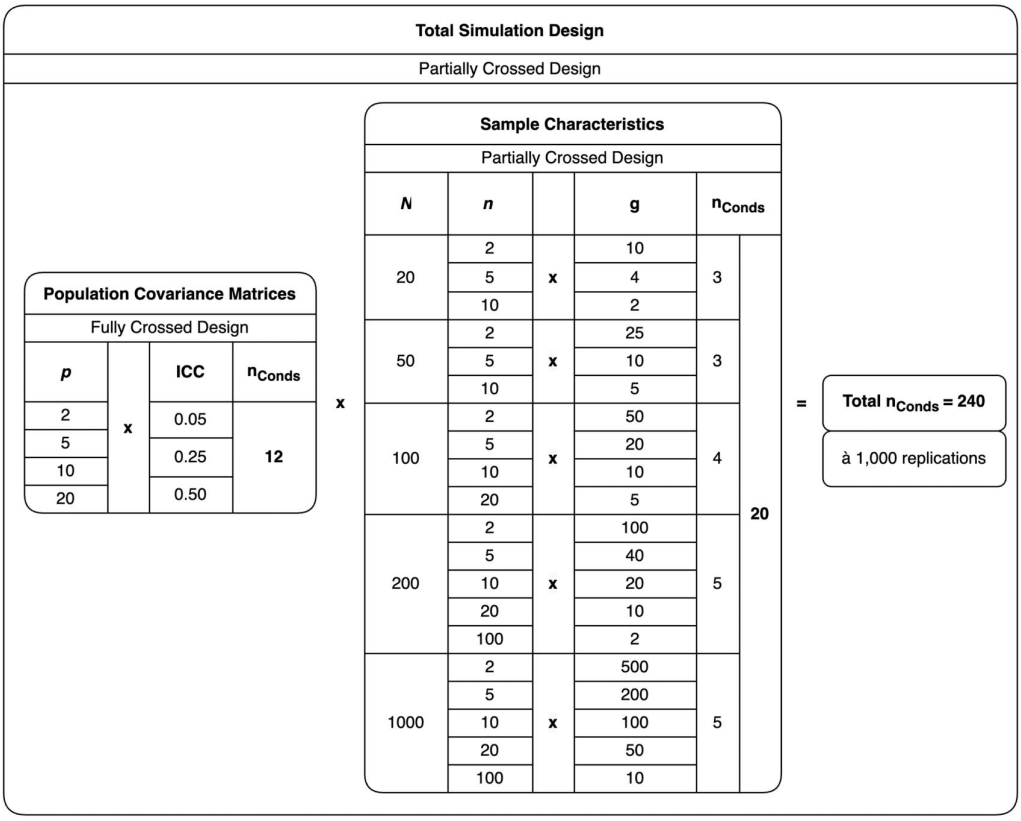


Figure 3. Simulation design.

n, $g < 2$, and where g was not an integer. Applying the described partially crossed design for N, n, and g, we arrived at 20 different sample conditions. Note that we chose the variations of p, n, and g to create conditions where LF-T has $cols \leq rows$ (and thus, LF-B and WF-T have $cols > rows$, $cols = rows$, or $cols < rows$). We have done this because of the $cols \leq rows$ requirement of lavaan for LF-T and WF-T. When LF-T has $cols > rows$, then WF-T has $cols > rows$, too, and neither approach could be applied. Moreover, we did this to investigate whether the LF approach leads to good results in settings where it is applicable but the WF approach is not (i.e., LF-T has $cols \leq rows$, but WF-T has $cols > rows$). Finally, we partially crossed the population and sample conditions and arrived at a total of 240 simulation conditions. For each condition, we simulated 1000 data sets. The complete simulation design is shown in Figure 3.

3.1.2. Evaluation Criteria

To assess model performance, we included convergence rate, and for estimation accuracy, relative root mean squared error (RMSE), relative bias, and relative variance[9]. The RMSE assesses the overall accuracy of an estimator. It is defined as the square root of the mean of the squared differences between the estimates and the population value. Bias is a measure to assess the extent with which an estimator targets the population value. It is defined as the mean of the differences between the estimates and the population value. Variance is a measure of the efficiency of an estimator. It is defined as the mean of the differences between the estimates and the mean of all estimates. We used the relative versions of these parameters dividing them by the respective population value (which is determined by the ICC). We further multiplied them by 100 to arrive at percentages. We did this to investigate potential differences in accuracy, bias, and efficiency with respect to the ICC. For example, when $RMSE = 0.1$ for both $\theta = ICC = 0.5$ and $\theta = ICC = 0.1$, this mounts to relative RMSE of 20% and 100%, respectively, which suggests that the estimation of the smaller ICC is less accurate. Note that we only consider estimation accuracy of (co)variances but not of means of the intercept-only models.

3.2. Results

In the following, we summarize the results of (a) the effect of the data format, and (b) the effect of $cols : rows$ on convergence and estimation accuracy criteria in each data format. For this purpose, we plotted the results of the Monte Carlo study aggregated (a) by the sample size at level-2, the

[9]We further investigated the aforementioned matrix properties, non-singularity, definiteness, and the condition number. However, results suggested that these matrix properties did not offer critical information above $cols : rows$. Thus we do not report them in the main findings. Further information can be found in Figure 1 in the Online Supplemental Material.

number of groups g (same data set, different data formats), and (b) by *cols* : *rows* of LF-B and WF-T (different data sets, same data format). Out of the sample sizes at the two levels, level-1 group size n, and level-2 number of groups g, we decided to plot the latter because it equals the number of *rows* (i.e., observations) in both LF-B and WF-T. Table 1 shows descriptive statistics of the convergence and estimation accuracy criteria. Note that for the latter, we used only simulation conditions that resulted in 100% convergence in both approaches to minimize differences in Monte Carlo error in both approaches and thus, facilitate comparison of both approaches. An overview of the results of the simulation study clustered by the factors of the simulation design (i.e., 240 simulation conditions for each approach) can be found in Figure 2 in the Online Supplemental Material.

3.2.1. The Effect of Data Format

3.2.1.1. Convergence. In Figure 4, convergence rates aggregated by sample size at level-2 g are depicted. It is evident that with increasing g, average convergence rates increased with a seemingly logarithmic trend in the LF and WF approach. Nevertheless, there was an effect of data format on convergence, as the LF approach was more likely to converge in small and moderate sample sizes. For instance, the average convergence rate in $g = 100$ was 100% in the LF approach, and 80% in the WF approach. Further, for the WF approach, we see that the convergence trend was non-monotonous with g. Whereas $g = 25$ had an average convergence rate of 75%, $g = 40$ and $g = 50$ had lower ones. The non-monotonous trend of the WF approach suggests that other terms than g might also be relevant for convergence. We will scrutinize this matter in the following section when investigating the *cols* : *rows* effects.

3.2.1.2. Estimation Accuracy. Results for the estimation accuracy of between-group parameters are shown in Figure 5. The relative RMSE (Panel A), relative bias (Panel B), and relative variance (Panel C) decreased with larger numbers of groups g in a very similar fashion for both approaches. Slight differences were only present for settings with small g and small *ICC* where variability across simulation conditions was large. Overall, this suggests that there was no substantial effect of the data format on estimation accuracy of between-group level parameters. However, the findings revealed a connection between the estimation accuracy of between-group level parameters and the magnitude of the *ICC*. Specifically, the smaller the *ICC*, the more inaccurate, biased, and inefficient were the between-group parameters. Besides main effects of g and *ICC*, the results further imply an interaction effect. The least accurate, most biased, and inefficient results were obtained in settings with small g and small *ICC*.

Similar patterns have been found for the estimation accuracy of the within-group parameters that are depicted in Figure 6. However, in contrast to the between-group parameters, an effect of the *ICC* was only present for the relative variance, as was an interaction effect of g and *ICC*. Furthermore, the within-group parameters were overall more accurate, less biased, and more efficient than the between-group parameters.

3.2.1.3. Summary. Our findings suggest that there was an effect of data format on convergence, but not on estimation accuracy, despite the different *cols* : *rows* (and implied magnitudes of eigenvalue biases) in both approaches. Further, we found an interaction effect of g, *ICC*, and parameter level on estimation accuracy. More specifically, the least accurate, most biased, and inefficient parameter estimates were

Table 1. Descriptive statistics of the model performance criteria.

Criterion	M	SD	Median	Min	Max
Convergence rate[a]	49.28	49.33	49.25	0.00	100.00
Relative RMSE[b]	84.83	101.44	47.76	14.04	571.09
Relative bias[b]	−4.12	12.42	−1.27	−105.16	5.04
Relative variance[b]	10.36	20.33	3.10	0.48	144.17

Unit for all criteria is %.
[a]$N = 480$ (all simulation conditions for both approaches).
[b]$N = 156$ (all simulation conditions that converged 100% in both approaches).

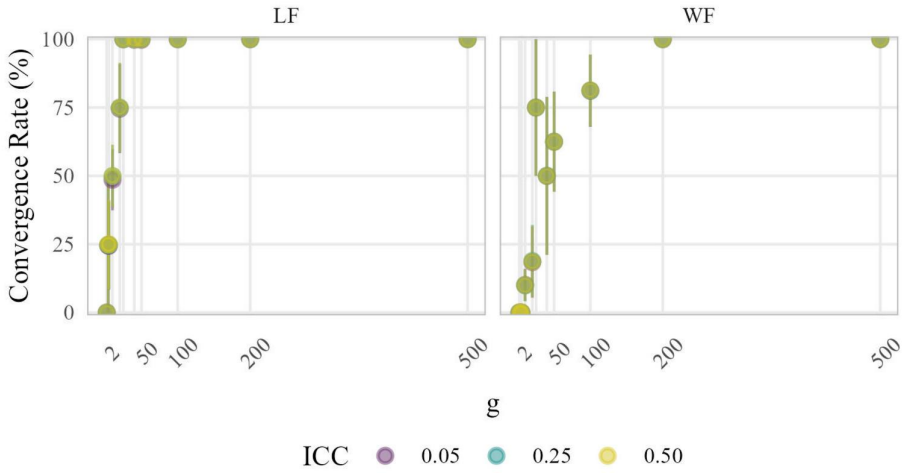


Figure 4. Convergence aggregated by sample size at level-2. Points indicate means; lines indicate means ± standard errors (i.e., variability across simulation conditions). The sample size at level-2 g corresponds to the *rows* of both LF-B and WF-T.

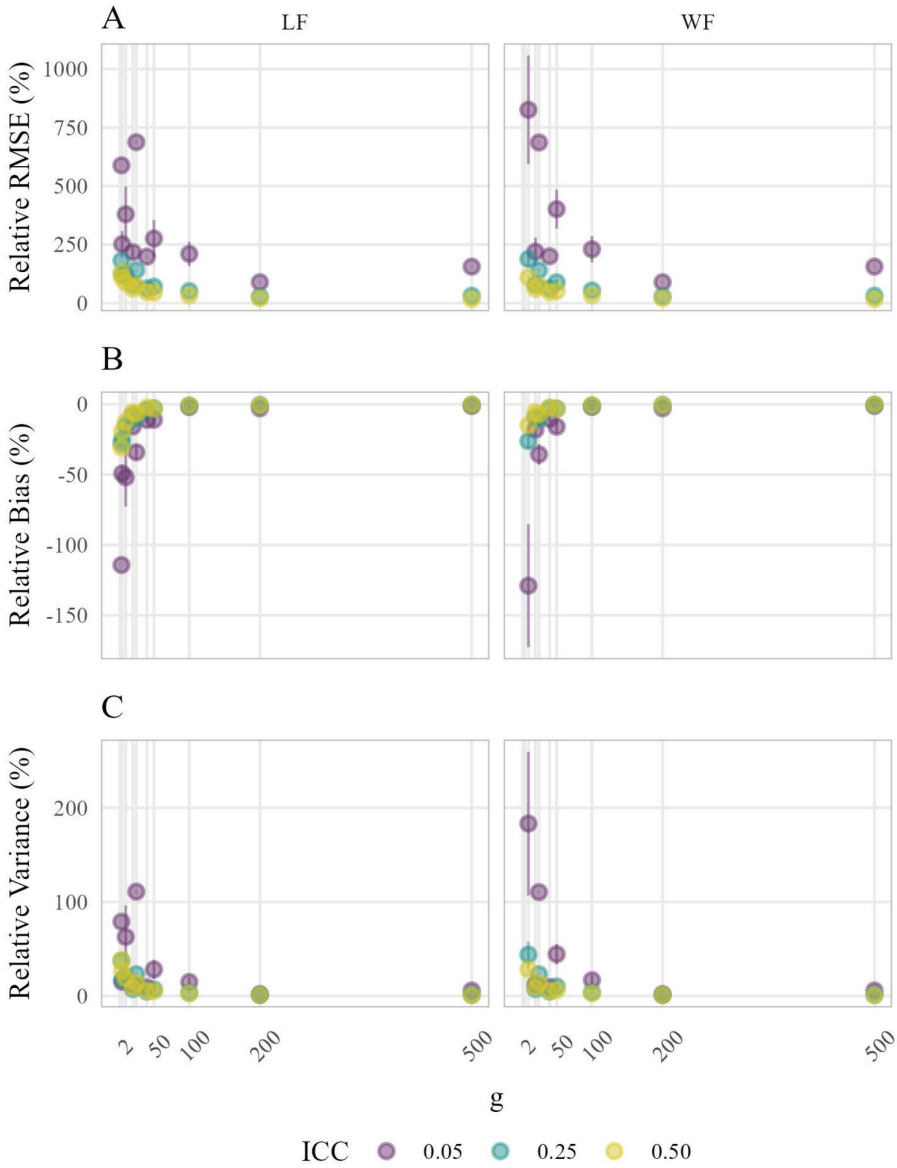


Figure 5. Estimation accuracy of between-group parameters aggregated by sample size at level-2. Points indicate means; lines indicate means ± standard errors (i.e., variability across simulation conditions). The sample size at level-2 g corresponds to the *rows* of both LF-B and WF-T.

obtained in samples with small numbers of groups, small *ICC* values, and at the between-group level.

3.2.2. The Effect of Cols : Rows in Each Data Format

3.2.2.1. Convergence. In Figure 7, convergence rates aggregated by *cols* : *rows* of LF-B and WF-T, $p : g$ and $(p \cdot n) : g$, are shown. For *cols* < *rows*, both approaches led to average convergence rates of $\approx 100\%$. For *cols* = *rows*, however, results for both approaches differed. In LF, this always led to non-convergence. In WF, half of the models converged. For the LF approach, this extends the minimum data set requirements given by lavaan and informs about optimal study design. We require *cols* < *rows* in the input data matrix LF-T ($p < (g \cdot n)$) to use lavaan, and our results suggest that we additionally need to satisfy *cols* < *rows* in LF-B ($p < g$) to get a converging model in the LF approach. In other words, we have to design our study in such a way that the number of level-2 variables is smaller than the number of groups. The reason why $p < g$ is required might be connected to the eigenvalue bias in $\hat{\boldsymbol{\Sigma}}_{LF-B}$, which might have been non-negligible when *cols* = *rows*[10]. However, we cannot confirm that because we used *cols* : *rows* only as a proxy of the eigenvalue bias. Comparing convergence rates aggregated by *cols* : *rows* and aggregated by sample size at level-2 (i.e., *rows*; see Figure 4), it can be seen that the former exhibited smaller variability and thus, gave more reliable information on convergence. It further suggests that

[10]Note that the matrix properties singularity and non-definiteness of $\hat{\boldsymbol{\Sigma}}_{LF-B}$ were non-informative here (for more information see Figures 1, 3, and 4 in the Online Supplemental Material).

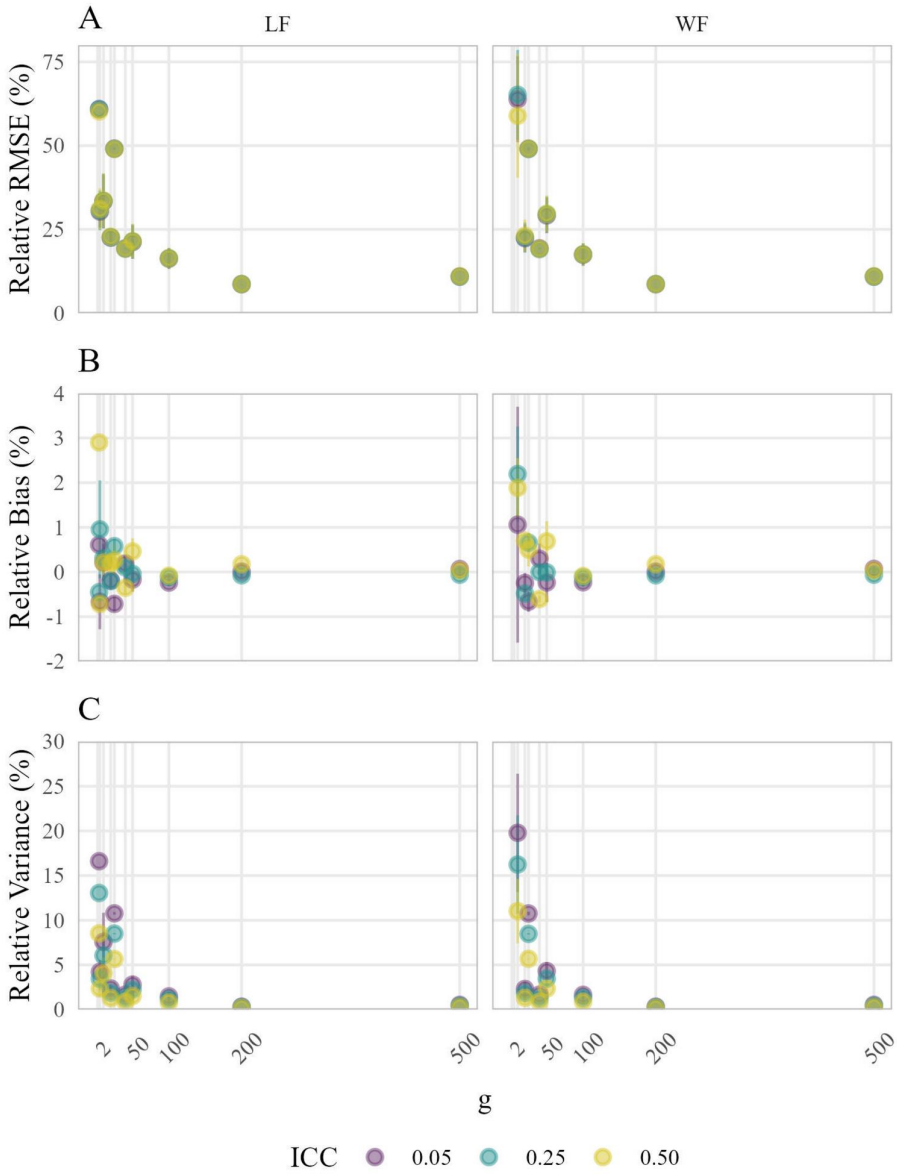


Figure 6. Estimation accuracy of within-group parameters aggregated by sample size at level-2. Points indicate means; lines indicate means ± standard errors (i.e., variability across simulation conditions). The sample size at level-2 *g* corresponds to the *rows* of both LF-B and WF-T.

the effect of data format on convergence was attributable to the differences in *cols* : *rows* in both approaches.

3.2.2.2. Estimation Accuracy. We saw in the last section that there was no noticeable effect of data format on estimation accuracy. Put differently, the estimation accuracy in both approaches was very similar for a given number of groups *g* (i.e., *rows* of LF-B and WF-T) despite different *cols*, and thus, *cols* : *rows* of LF-B and WF-T. This suggests that there must have been different effects for *cols* : *rows* in each data format. In Figure 8, the relative RMSE (Panel A), relative bias (Panel B), and relative variance (Panel C) by the *cols* : *rows* in each data format are shown. We indeed see that the *cols* : *rows* effect is less steep in the WF approach. For example, similar relative RMSE, relative bias, and relative variance where obtained with an *cols* : *rows* of 0.2 in LF-B in the LF approach and 0.4 in WF-T in the WF approach. Within each data format, larger *cols* : *rows* resulted in larger relative RMSE, relative bias, and relative variance. Further, there was an interaction with the *ICC*. Larger *cols* : *rows* and smaller *ICC* values led to more inaccurate, biased, and inefficient estimation, with the most problematic results in settings with large *cols* : *rows* and small *ICC* values. To continue the example from above, a *cols* : *rows* of 0.2 in LF-B in combination with the smallest *ICC* 0.05 yielded approximately a six times higher relative RMSE than large *ICC* values of 0.50 in the LF approach. Note however, that these *cols* : *rows* effects, and interaction effects of *cols* : *rows* and *ICC* were not strictly monotonous. We will take up the issue in the discussion.

For the estimation accuracy of the within-group parameters, which is depicted in Figure 9, there was a slightly

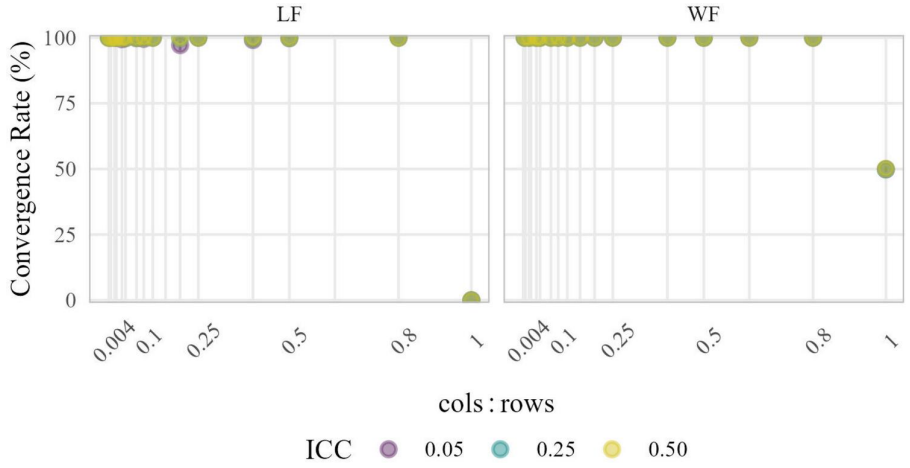


Figure 7. Convergence aggregated by *Cols* : *Rows*. Points indicate means; lines indicate means ± standard errors (i.e., variability across simulation conditions). The *cols* : *rows* of LF-B and WF-T, $p : g$ and $(p \cdot n) : g$, are depicted.

increasing trend in relative RMSE (Panel A) and relative variance (Panel C) with increasing *cols* : *rows*. For the relative bias (Panel B), there was no systematic trend. In addition, the relative variance showed small effects of the *ICC*. Smaller *ICC* resulted in somewhat larger relative variance. Overall, as for the between-group parameters, the *cols* : *rows* effect of the WF approach was less steep. Compared to the between-group parameters, the effects of *cols* : *rows* and the *ICC* on the estimation accuracy of the within-group parameters were very small.

3.2.2.3. Summary. The results showed that there were differential effects of *cols* : *rows* on convergence and estimation accuracy in both data formats (approaches). With regard to convergence, we found that *cols* < *rows* in LF-B ($p < g$) and *cols* ≤ *rows* in WF-T ($(p \cdot n) \leq g$) had to be satisfied which expands the minimum data set requirements of lavaan. For estimation accuracy, the effect in the WF approach was less strong. For both approaches, increasing *cols* : *rows* and having a smaller *ICC* was detrimental. These effects were more pronounced for the between-group parameters and smaller *ICC* values.

4. Discussion

One can arrange a two-level data set in two different data formats, LF and WF. In the two data formats, the involved data matrices, have inherently different *cols* : *rows*. The *cols* : *rows* is the $p : N$ equivalent, which depends on g, n, and p in multilevel settings, and implies a magnitude of bias in the sample eigenvalues. For both data formats, SEM approaches to estimate multilevel models with standard ML exist. Past research has provided evidence for analytical and empirical equivalence in settings with large samples sizes at level-2 where *cols* ≪ *rows* in all data matrices, and thus, the assumed bias in eigenvalues was negligibly small. Using a Monte Carlo study, we included settings with small sample sizes at level-2 where *cols* : *rows* differs in all data matrices. We investigated the effect of the data format (with the same data set in different data formats), and the effects of *cols* : *rows* (with different data sets in the same data format).

Regarding the effect of the data format, we found only an effect on convergence. In particular, the LF approach with its inherently smaller *cols* : *rows* was more likely to converge. The estimation accuracy of both approaches did not differ substantially by the choice of the data format. Thus, the results of our study extend the evidence of the empirical equivalence of the LF and WF approaches to settings where both the sample size at level-1 n, and the sample size at level-2 g, are small.

Regarding the *cols* : *rows*, we found differential effects on convergence and estimation accuracy in both data formats (approaches). Concerning the former, the LF approach requires that the number of variables is smaller than the number of groups ($p < g$), whereas the WF approach requires that the number of variables multiplied by group size is smaller than or equal to the number of groups ($(p \cdot n) < g$; which is equivalent to the lavaan requirement). *Cols* : *rows* (number of observed variables to sample size at level-2) includes more information than *rows* alone (sample size at level-2). It informs us about minimum requirements on study design, and which data format (approach) to use for converging models.

In accordance with the effect of data format, we found differential *cols* : *rows* effects on estimation accuracy. Within each data format (approach), smaller *cols* : *rows* resulted in models that yield more accurate, less biased, and more efficient parameter estimates. However, the WF approach had a less steep effect. Thus, the difference in *cols*, and thus, in *cols* : *rows*, in WF-T compared to LF-B (by the factor n) was not, as expected, detrimental for the estimation accuracy. Instead, the factor n was likely responsible for the less steep effect. Apparently, n appears to simply operate distinctly in both data formats (approaches). In the LF approach, the sample size at level-1 n influences the accuracy of the first and second-order moments at level-2. In the WF approach, the sample size at level-1 n (together with p) determines the number of "observed variables", but also the

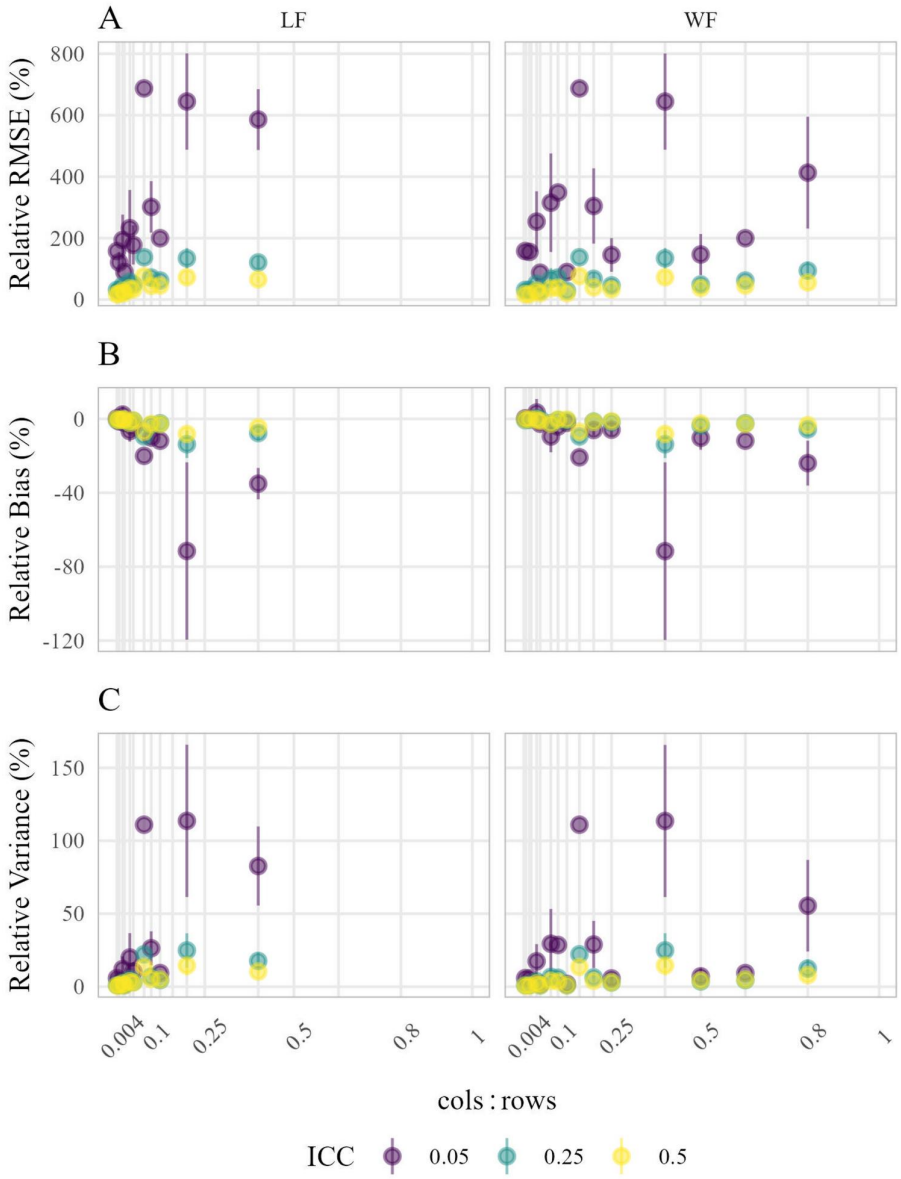


Figure 8. Estimation accuracy of between-group parameters aggregated by *Cols* : *Rows*. Points indicate means; lines indicate means ± standard errors (i.e., variability across simulation conditions). The *cols* : *rows* of LF-B and WF-T, $p : g$ and $(p \cdot n) : g$, are depicted.

number of equality constrains between related model parameters. Having more relations between "observed variables" with equality constrains similarly yields more accurate estimates. In sum, we assumed an effect of data format on estimation accuracy, based on the same *cols* : *rows* effect on estimation accuracy in both data formats (approaches). However, we found no effect of data format on estimation accuracy, but different *cols* : *rows* effects on estimation accuracy in both data formats (approaches).

Besides the main effects of data format and *cols* : *rows*, we found further noteworthy effects. Most notable is the interaction between *g* (or *cols* : *rows*) and magnitude of the *ICC* and parameter level on estimation accuracy. In particular, the smaller *g*, or the larger the *cols* : *rows*, and the smaller the *ICC* values, the more inaccurate, biased, and inefficient are the between-group parameters. Again, these findings reveal valuable insight into optimal study design. When between-group parameters are of interest, and prior studies found marginal *ICC* values, then future studies should have relatively large *g* (or *cols* : *rows*).

4.1. Limitations and Directions for Future Research

Firstly, we used the *cols* : *rows* of a data matrix only as a proxy of the eigenvalue bias of the LF and WF covariance matrices. Our findings showed that within each data format, larger *cols* : *rows* resulted in lower convergence rates, and less accurate between-group parameter estimates in both approaches, which suggests that eigenvalue biases might have increased with increasing *cols* : *rows*. However, we did

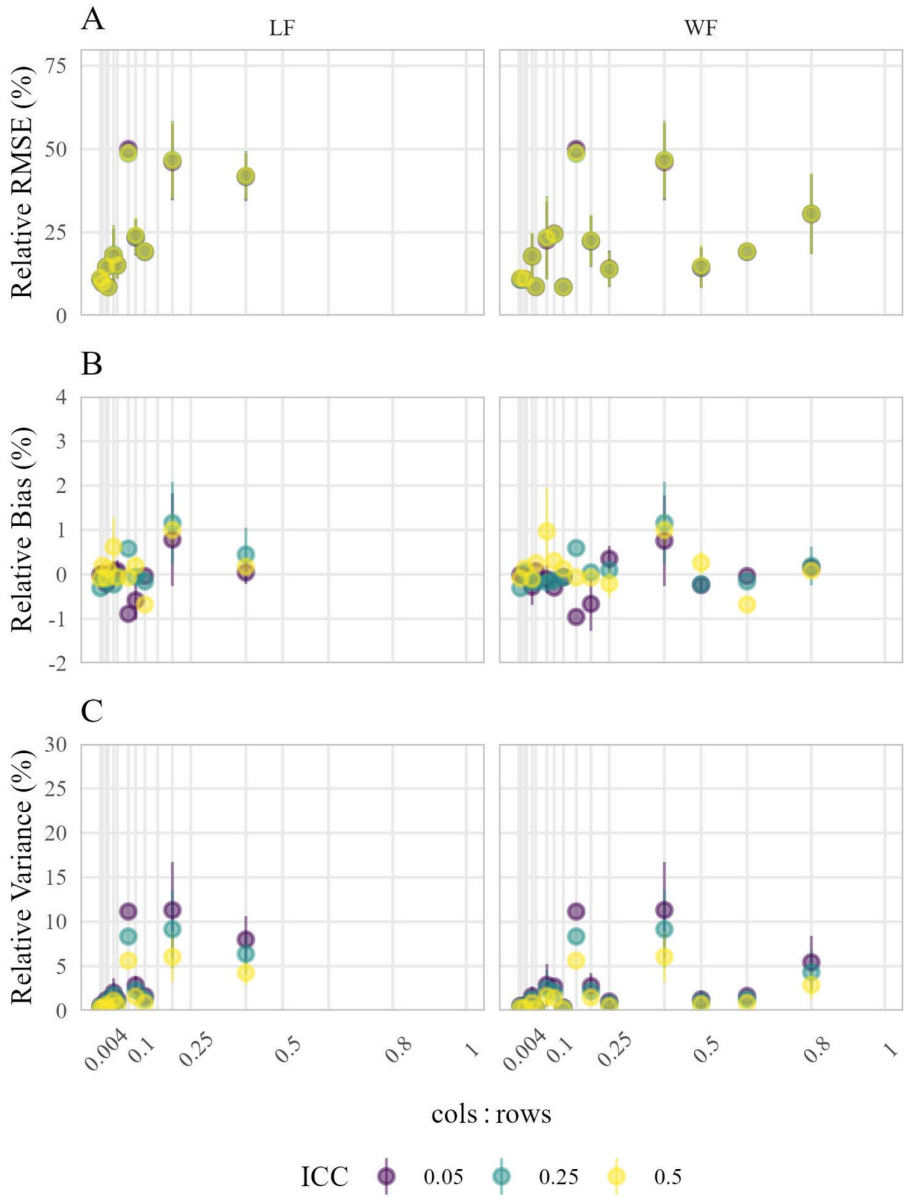


Figure 9. Estimation accuracy of within-group parameters aggregated by *Cols* : *Rows*. Points indicate means; lines indicate means ± standard errors (i.e., variability across simulation conditions). The *cols* : *rows* of LF-B and WF-T, $p : g$ and $(p \cdot n) : g$, are depicted.

not investigate the exact term of the eigenvalue biases. For example, as pointed out earlier, when level-2 variables are aggregates of level-1 variables, the matrix properties of $\hat{\boldsymbol{\Sigma}}_{LF-B}$ are not only influenced by p and g (which constitute *cols* : *rows*), but by n and the *ICC* (Bhargava & Disch, 1982; Hill & Thompson, 1978; Searle et al., 1992). This suggests that the bias term of its eigenvalues is influenced by these factors. Further, the effect of *cols* : *rows* on convergence and estimation accuracy was different for the LF and WF approach. This might be related to different eigenvalue biases. On the other hand, it could suggest that eigenvalue biases exert minor influence on convergence and estimation accuracy. Future research could investigate the exact term of the eigenvalue biases of LF and WF covariance matrices, and whether the difference in eigenvalue biases explains the difference in convergence and estimation accuracy.

Secondly, because our study was the first to investigate the $p : N$ (*cols* : *rows*) equivalent within multilevel SEM, we included only simple intercept-only models. The influence in more complex models remains to be investigated. For example, it has been shown that in measurement models, convergence rates decreases with smaller numbers of indicators per factor, and smaller magnitude of factor loadings and factor correlations (J. C. Anderson & Gerbing, 1984; Boomsma, 1985; Jöreskog & Sörbom, 1984). Further, factor correlations are depended on correlations of indicators (i.e., observed variables), which implies that larger correlations of observed variables are desirable. In our study, correlations

of all observed variables were fixed to 0.3, and covariances of observed variables were computed by the correlation and the respective variances. Differences in covariances did not show any influence on convergence. However, covariances (correlations) of observed variables have a bivariate nature, whereas factor correlations have a multivariate nature. It would be interesting to investigate how eigenvalue bias influences convergence and estimation accuracy in factor models. Further, level-2 predictors might be examined. We included only level-2 variables that are aggregates of level-1 variables. In other words, we investigated contextual analysis models. With level-2 predictors, the importance of larger n for more accurate between parameters could be smaller. The eigenvalue bias term of between-group covariance matrices of predictor variables is assumed to differ from those of between-group covariance matrices of level-1 aggregates. In sum, future research should investigate the influence of the eigenvalue bias in more complex multilevel SEM models with measurement models and level-2 predictors.

Thirdly, future research could investigate whether improving the eigenvalue structure of LF and WF covariance matrices can improve performance. In single level settings, it has been shown that improving the eigenvalues of the sample covariance matrix can result in increased convergence rates and more accurate model estimates (e.g., Kamada & Kano, 2012; Kamada et al., 2014; Yuan & Bentler, 2017; Yuan & Chan, 2008; Yuan et al., 2011). In multilevel settings, it has been shown that improving the eigenvalues of the model-implied between-group matrix results in more accurate between-group parameters (e.g., Chung et al., 2015; McNeish, 2016; Zitzmann, 2018). It seems promising to investigate improving the eigenvalues of the LF and WF covariance matrices, more specifically, $\hat{\boldsymbol{\Sigma}}_{LF-B}$ in the LF approach and $\hat{\boldsymbol{\Sigma}}_{WF-T}$ in the WF approach, and whether this could increase the accuracy of between-group parameters.

Fourthly, future studies could examine the effect of the *ICC* on the bias of the between-group parameters more closely. First, the evidence for its existence is mixed. Whereas some studies found no effect (e.g., Hox et al., 2010; McNeish & Stapleton, 2016; Meuleman & Billiet, 2009; Stegmueller, 2013), other studies, ours included, found that a smaller *ICC* leads to increased bias (e.g., Hox & Maas, 2001; Lüdtke et al., 2008, 2011; Muthen & Satorra, 1995; Zitzmann et al., 2015). Second, within the studies that found an effect, the direction of the bias differs. Most studies found a downward bias. However, the bias was mostly aggregated over parameters of different types. Hox and Maas (2001) classified them by type and found upward bias in variances and downward bias in factor loadings. Further, the type of parameter also depends on the modeling approach. Lüdtke et al. (2011) compared manifest and latent approaches and found that only the approach who had latent variables at both levels resulted in upward bias. In our study, both approaches incorporated latent variables at both levels, and both exhibited a downward bias. However, we did not distinguish between different types of parameters. Future research could explore how the presence and direction of an *ICC* effect on the bias of between-group parameters may vary depending on the parameter type.

4.2. Conclusion

Our study has demonstrated two important main results. First, data format influences convergence but not estimation accuracy. Second, the *rows* and *cols* : *rows* of data matrices, which varies when conducting multilevel analysis in long versus wide format, along with the *ICC* are critical factors influencing the convergence and estimation accuracy of multilevel SEM approaches. To conclude with some literary advice: We aim for convergence all along, with accuracy nothing going wrong, especially if the *ICC*s are not strong, our data matrices should be comparatively long.

Disclosure statement

The authors report there are no competing interests to declare.

ORCID

Julia-Kim Walther http://orcid.org/0000-0001-5758-1211
Martin Hecht http://orcid.org/0000-0002-5168-4911
Benjamin Nagengast http://orcid.org/0000-0001-9868-8322
Steffen Zitzmann http://orcid.org/0000-0002-7595-4736

References

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, *49*, 155–173. https://doi.org/10.1007/BF02294170

Anderson, T. W. (2003). *Introduction to multivariate statistical analysis* (3rd ed ed.). Wiley.

Arruda, E. H., & Bentler, P. M. (2017). A regularized GLS for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*, 657–665. https://doi.org/10.1080/10705511.2017.1318392

Auguie, B., & Antonov, A. (2017). *gridExtra: Miscellaneous functions for "grid" graphics* (2.3) [Computer software]. https://CRAN.R-project.org/package=gridExtra

Barendse, M., & Rosseel, Y. (2020). Multilevel modeling in the 'wide format' approach with discrete data: A solution for small cluster sizes. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 696–721. https://doi.org/10.1080/10705511.2019.1689366

Ben-Shachar, M. S., Makowski, D., Lüdecke, D., Patil, I., Wiernik, B. M., Thériault, R., Kelley, K., Stanley, D., Caldwell, A., Burnett, J., & Karreth, J. (2023). *effectsize: Indices of effect size* (0.8.3) [Computer software]. https://cran.r-project.org/web/packages/effectsize/index.html

Bhargava, A. K., & Disch, D. (1982). Exact probabilities of obtaining estimated non-positive definite between-group covariance matrices. *Journal of Statistical Computation and Simulation*, *15*, 27–32. https://doi.org/10.1080/00949658208810561

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in lisrel maximum likelihood estimation. *Psychometrika*, *50*, 229–242. https://doi.org/10.1007/BF02294248

Boyd, L. H., & Iversen, G. R. (1979). *Contextual analysis: Concepts and statistical techniques*. Wadsworth Publishing Company.

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., & Dorie, V. (2015). Weakly informative prior for point estimation of covariance matrices in hierarchical models. *Journal of Educational and Behavioral Statistics*, *40*, 136–157. https://doi.org/10.3102/1076998615570945

Dempster, A. P. (1972). Covariance selection. *Biometrics*, *28*, 157–175. https://doi.org/10.2307/2528966

Deng, L., Yang, M., & Marcoulides, K. M. (2018). Structural equation modeling with many variables: A systematic review of issues and developments. *Frontiers in Psychology*, *9*, 580. https://doi.org/10.3389/fpsyg.2018.00580

Dragulescu, A., & Arendt, C. (2020). *xlsx: Read, write, format excel 2007 and excel 97/2000/XP/2003 files* (0.6.5) [Computer software]. https://CRAN.R-project.org/package=xlsx

Duncan, T. E., Duncan, S. C., Alpert, A., Hops, H., Stoolmiller, M., & Muthen, B. (1997). Latent variable modeling of longitudinal and multilevel substance use data. *Multivariate Behavioral Research*, *32*, 275–318. https://doi.org/10.1207/s15327906mbr3203_3

Fox, J., Weisberg, S., Price, B., Adler, D., Bates, D., Baud-Bovy, G., Bolker, B., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., Graves, S., Heiberger, R., Krivitsky, P., Laboissiere, R., Maechler, M., Monette, G., Murdoch, D., Nilsson, H., … R-Core. (2023). *car: Companion to applied regression* (3.1-2) [Computer software]. https://cran.r-project.org/web/packages/car/index.html.

Golub, G. H., & Van Loan, C. F. (2013). *Matrix computations*. John Hopkins University Press.

Gorsuch, R. L. (1983). *Factor analysis*. Lawrence Earlbaum Associates.

Gulliford, M. C., Ukoumunne, O. C., & Chinn, S. (1999). Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: Data from the health survey for England 1994. *American Journal of Epidemiology*, *149*, 876–883. https://doi.org/10.1093/oxfordjournals.aje.a009904

Hayashi, K., Yuan, K. H., & Liang, L. (2018). On the bias in eigenvalues of sample covariance matrix. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology* (Vol. 233, pp. 221–233). Springer International Publishing.

Herzog, W., Boomsma, A., & Reinecke, S. (2007). The model-size effect on traditional and modified tests of covariance structures. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*, 361–390. https://doi.org/10.1080/10705510701301602

Hill, W. G., & Thompson, R. (1978). Probabilities of non-positive definite between-group or genetic covariance matrices. *Biometrics*, *34*, 429–439. https://doi.org/10.2307/2530605

Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 157–174. https://doi.org/10.1207/S15328007SEM0802_1

Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, *64*, 157–170. https://doi.org/10.1111/j.1467-9574.2009.00445.x

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Huang, Y., & Bentler, P. M. (2015). Behavior of asymptotically distribution free test statistics in covariance versus correlation structure analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 489–503. https://doi.org/10.1080/10705511.2014.954078

Hugh-Jones, D. (2022). *huxtable: Easily create and style tables for LaTeX, HTML and other formats* (5.5.2) [Computer software]. https://CRAN.R-project.org/package=huxtable

Jackson, D. L. (2001). Sample size and number of parameter estimates in maximum likelihood confirmatory factor analysis: A Monte Carlo investigation. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 205–223. https://doi.org/10.1207/S15328007SEM0802_3

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N: Q hypothesis. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*, 128–141. https://doi.org/10.1207/S15328007SEM1001_6

Jak, S., Jorgensen, T. D., & Rosseel, Y. (2021). Evaluating cluster-level factor models with lavaan and Mplus. *Psych*, *3*, 134–152. https://doi.org/10.3390/psych3020012

Jöreskog, K. G., & Sörbom, D. (1984). *LISREL VI, analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods*. Scientific Software, Inc.

Kamada, A., & Kano, Y. (2012). *Statistical inference in structural equation modeling with a near singular covariance matrix*. 2nd Institute of Mathematical Statistics Asia Pacific Rim Meeting Tsukuba, Japan.

Kamada, A., Yanagihara, H., Wakaki, H., & Fukui, K. (2014). Selecting a shrinkage parameter in structural equation modeling with a near singular covariance matrix by the GIC minimization method. *Hiroshima Mathematical Journal*, *44*, 315–326. https://doi.org/10.32917/hmj/1419619749

Kelley, C. T. (1995). *Iterative methods for linear and nonlinear equations*. Society for Industrial and Applied Mathematics.

Lange, K., Chambers, J., & Eddy, W. (1999). *Numerical analysis for statisticians* (Vol. 2). Springer.

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2×2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological Methods*, *16*, 444–467. https://doi.org/10.1037/a0024376

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). September) The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*, 203–229. https://doi.org/10.1037/a0012869

Marcoulides, K. M., Yuan, K. H., & Deng, L. (2023). Structural equation modeling with small samples and many variables. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (2nd ed., pp. 525–542). Guilford Press.

McNeish, D. M. (2016). Using data-dependent priors to mitigate small sample bias in latent growth models: A discussion and illustration using mplus. *Journal of Educational and Behavioral Statistics*, *41*, 27–56. https://doi.org/10.3102/1076998615621299

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, *28*, 295–314. https://doi.org/10.1007/s10648-014-9287-x

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*, 259–284. https://doi.org/10.1037/1082-989X.10.3.259

Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methods*, *3*, 45–58.

Muthén, B. O. (1990). *Mean and covariance structure analysis of hierarchical data*. (*Tech. Rep.*). Department of Statistics.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*, 376–398. https://doi.org/10.1177/0049124194022003006

Muthen, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 267. https://doi.org/10.2307/271070

Navarro, D. (2021). *lsr: Companion to "learning statistics with R"* (0.5.2) [Computer software]. https://CRAN.R-project.org/package=lsr

Pedersen, T. L. (2022). *patchwork: The composer of plots* (1.1.2) [Computer software]. https://CRAN.R-project.org/package=patchwork

R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). SAGE.

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., & Firth, D. (2023). *MASS: Support functions and datasets for venables and Ripley's MASS* (7.3-58.2) [Computer software]. https://CRAN.R-project.org/package=MASS

Robinson, D., Hayes, A., Patil, I., Chiu, D., Gomez, M., Demeshev, B., Menne, D., Nutter, B., Johnston, L., Bolker, B., Briatte, F., Arnold, J., Gabry, J., Selzer, L., Simpson, G., Preussner, J., Hesselberth, J., Wickham, H., Lincoln, M., … Reinhart, A. (2023). *broom: Convert statistical objects into tidy tibbles* (1.0.5) [Computer software]. https://cran.r-project.org/web/packages/broom/index.html

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

Rosseel, Y. (2018). *Multilevel structural equation modeling with lavaan*. https://docplayer.net/123371066-Multilevel-structural-equation-modeling-with-lavaan.html

Rosseel, Y., Jorgensen, T. D., Rockwood, N., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., Barendse, M., Scharf, F., & Du, H. (2023). *lavaan:*

Latent variable analysis (0.6-14) [Computer software]. https://CRAN.R-project.org/package=lavaan

Schäfer, J., & Strimmer, K. (2005). A Shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*, 32. https://doi.org/10.2202/1544-6115.1175

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. John Wiley & Sons, Inc.

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

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

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and bayesian approaches. *American Journal of Political Science*, *57*, 748–761. https://doi.org/10.1111/ajps.12001

Stein, C. (1956). Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. In J. Neyman (Ed.), *Proceedings of the third Berkeley symposium on mathematical statistics and probability: Contributions to the theory of statistics* (Vol. 1, pp. 197–206). University of California Press.

Stein, C. (1975). *Estimation of a covariance matrix*. 39th Annual Meeting IMS, Atlanta, GA.

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., & Studio, R. (2023). *ggplot2: Create elegant data visualisations using the grammar of graphics* (3.4.1) [Computer software]. https://CRAN.R-project.org/package=ggplot2

Wickham, H., François, R., Henry, L., Müller, K., Vaughan, D., Posit, & PBC. (2023). *dplyr: A grammar of data manipulation* (1.1.0) [Computer software]. https://CRAN.R-project.org/package=dplyr

Wickham, H., Girlich M., & R Studio. (2022). *tidyr: Tidy messy data* (1.2.0) [Computer software]. https://CRAN.R-project.org/package=tidyr

Wickham, H., & R Studio. (2022). *stringr: Simple, consistent wrappers for common string operations* (1.5.0) [Computer software]. https://cran.r-project.org/web/packages/stringr/index.html

Wilke, C. O. (2020). *cowplot: Streamlined plot theme and plot annotations for "ggplot2"* (1.1.1) [Computer software]. https://cran.r-project.org/web/packages/cowplot/index.html

Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 256–293). SAGE Publications.

Xing, L., & Yuan, K. (2017, July 20). *Model evaluation with small N and/or large p* [Paper presentation]. Annual Meeting of the Psychometric Society, Zurich, Switzerland.

Yu, G., & Xu, S. (2022). *ggbreak: Set axis break for "ggplot2"* (0.1.2) [Computer software]. https://cran.r-project.org/web/packages/ggbreak/index.html

Yuan, K. H., & Bentler, P. M. (2017). Improving the convergence rate and speed of Fisher-scoring algorithm: Ridge and anti-ridge methods in structural equation modeling. *Annals of the Institute of Statistical Mathematics*, *69*, 571–597. https://doi.org/10.1007/s10463-016-0552-2

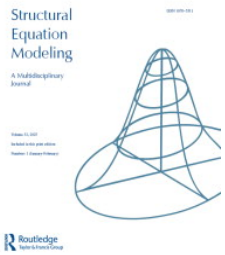
Yuan, K. H., & Chan, W. (2008). Structural equation modeling with near singular covariance matrices. *Computational Statistics & Data Analysis*, *52*, 4842–4858. https://doi.org/10.1016/j.csda.2008.03.030

Yuan, K. H., Jiang, G., & Yang, M. (2018). Mean and mean-and-variance corrections with big data. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 214–229. https://doi.org/10.1080/10705511.2017.1379012

Yuan, K. H., Wu, R., & Bentler, P. M. (2011). Ridge structural equation modelling with correlation matrices for ordinal and continuous data: Ridge SEM with correlation matrices. *The British Journal of Mathematical and Statistical Psychology*, *64*, 107–133. https://doi.org/10.1348/000711010X497442

Zitzmann, S. (2018). A computationally more efficient and more accurate stepwise approach for correcting for sampling error and measurement error. *Multivariate Behavioral Research*, *53*, 612–632. https://doi.org/10.1080/00273171.2018.1469086

Zitzmann, S., Lüdtke, O., & Robitzsch, A. (2015). A bayesian approach to more stable estimates of group-level effects in contextual studies. *Multivariate Behavioral Research*, *50*, 688–705. https://doi.org/10.1080/00273171.2015.1090899



Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hsem20

Shrinking Small Sample Problems in Multilevel Structural Equation Modeling via Regularization of the Sample Covariance Matrix

Julia-Kim Walther, Martin Hecht & Steffen Zitzmann

To cite this article: Julia-Kim Walther, Martin Hecht & Steffen Zitzmann (2025) Shrinking Small Sample Problems in Multilevel Structural Equation Modeling via Regularization of the Sample Covariance Matrix, Structural Equation Modeling: A Multidisciplinary Journal, 32:1, 46-65, DOI: 10.1080/10705511.2024.2380919

To link to this article: https://doi.org/10.1080/10705511.2024.2380919

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC

Published online: 09 Aug 2024.

Submit your article to this journal

Article views: 531

View related articles

View Crossmark data

Citing articles: 2 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=hsem20

STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
2025, VOL. 32, NO. 1, 46–65
https://doi.org/10.1080/10705511.2024.2380919

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Shrinking Small Sample Problems in Multilevel Structural Equation Modeling via Regularization of the Sample Covariance Matrix

Julia-Kim Walther[a], Martin Hecht[b] and Steffen Zitzmann[c]

[a]University of Tübingen; [b]Helmut Schmidt University; [c]Medical School Hamburg

ABSTRACT

Small sample sizes pose a severe threat to convergence and accuracy of between-group level parameter estimates in multilevel structural equation modeling (SEM). However, in certain situations, such as pilot studies or when populations are inherently small, increasing samples sizes is not feasible. As a remedy, we propose a two-stage regularized estimation approach designed for scenarios with both a small number of groups and small group sizes, and a low ICC. The method employs the wide format approach to multilevel SEM, where, at first, the sample covariance matrix is replaced by a shrinkage estimate, and then, this estimate is used to fit the SEM. By means of a simulation study, we evaluated the effectiveness of our two-stage approach. Our findings demonstrate that this method consistently ensures model convergence, provides more accurate between-level estimates, and even improves accuracy of within-level estimates in cases of very small group sizes.

KEYWORDS

ICC; multilevel SEM; regularization; small samples

In psychology and the education sciences, observational units are often nested within higher-level units, such as students (level-1 units) within classes (level-2 units). Multilevel structural equation modeling (SEM) is a powerful tool for estimating parameters across these different levels. In the within-between framework used by common statistical software (e.g., Mplus and *lavaan*), parameters are decomposed into within-group level (e.g., student) and between-group level (e.g., class) components. Challenges arise when sample sizes are small at any level, leading traditional maximum likelihood estimation (MLE) methods to either fail to converge or produce highly inaccurate estimates of between-group level parameters (e.g., Hox et al., 2010; Hox & Maas, 2001; Lüdtke et al., 2008, 2011; McNeish & Stapleton, 2016; Meuleman & Billiet, 2009; Shin & Raudenbush, 2010; Stegmueller, 2013; Zitzmann, 2018; Zitzmann et al., 2015). However, collecting larger samples can be costly, time-consuming, or impractical for specific study designs, such as pilot studies with limited classes and students, or for certain populations, such as school boards. Moreover, small variances at the class (between-group) level, often expressed in relation to large variances at the student (within-group) level as low Intra Class Correlation (ICC), further lower convergence rates (Lüdtke et al., 2011; Zitzmann, 2018), and accuracy of class (between-group) level parameter estimates (Hox & Maas, 2001; Lüdtke et al., 2011; Muthen & Satorra, 1995; Zitzmann et al., 2021). Therefore, in scenarios with small samples and low ICCs, there is a need for an alternative approach that is straightforward to implement and can mend both convergence and accuracy issues.[1]

A broad category of methods, known as *regularization* encompasses techniques aimed at enhancing convergence and accuracy in statistical analyses. Originally developed by Tikhonov (1943) to address stability issues in inverse matrix problems, the concept of regularization was swiftly adopted by the statistical community to adapt traditional MLE to produce "reasonable answers in unstable situations" (Bickel et al., 2006, p. 272). "Unstable situations" cover a variety of scenarios, among them small sample sizes, where the goal is typically to minimize the chances of encountering degenerate matrices, inadmissible solutions, and models that either do not converge or yield highly inaccurate outcomes. Techniques commonly used involve refining traditional maximum likelihood estimation (MLE) by incorporating approaches such as shrinkage, constraints, fixed parameters, or penalties. Overall, the goals and techniques of regularization approaches differ considerably fairly. For instance, shrinkage estimation of the covariance matrix (e.g., Touloumis, 2015) aims at obtaining a well-behaved eigenstructure and more accurate estimates, whereas penalizing the objective function of estimators (e.g., P.-H. Huang et al., 2017; Jacobucci et al., 2016) is motivated by the goal of achieving more parsimonious models. Despite

CONTACT Julia-Kim Walther julia-kim.walther@uni-tuebingen.de Hector Research Institute of Education Sciences and Psychology, University of Tübingen, 72072 Tübingen, Germany.

[1]Note that, when research questions concern only within-group variation, fixed effect approaches might be an alternative solution for handling small sample settings. For example, one could dummy-code the groups such that one obtains fixed effects estimates for each group (see e.g., Allison, 2009; Muthen & Satorra, 1995; Snijders & Bosker, 2012). However, if one is interested in the variation of the between-group parameters, or if the number of groups is too large (given the separate modeling of between-group parameters for any group), then mixed-effects modeling, such as multilevel SEM, might be a more sensible choice.

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

their differences, all approaches share a common feature: they introduce a moderate amount of bias into estimation. This is driven by the principle of a "bias-variance tradeoff," where the strategy is to reduce variance by accepting increased bias, thereby enhancing the overall accuracy of the estimation. We understand regularization as umbrella term for biased estimators that are employed in unstable situations.[2]

Within the general SEM framework, several regularization techniques have been employed to alleviate estimation problems encountered with small sample sizes. For instance, techniques such as constrained MLE and Bayesian estimation introduce bias into MLE by either limiting the range of possible parameter values, such as setting latent variances to one, or using weakly informative priors. These methods, which include contributions from Anderson and Gerbing (1984), Chen et al. (2001), and Zitzmann et al. (2022) for constrained MLE, and Depaoli and Clifton (2015), Zitzmann et al. (2016), and Ulitzsch et al. (2023) for Bayesian estimation, specifically target the calculation of model parameters, essentially the "output" of a structural equation model (SEM). However, in situations where non-convergence and poor estimation accuracy is contingent on a distorted eigenstructure of the sample covariance matrix[3], essentially the "input" of a SEM, simply adjusting model parameters through regularization will not suffice.

In such cases, regularizing the sample covariance matrix itself may prove to be a more effective solution. The ridge method, widely used for addressing eigenstructure issues in the sample covariance matrix (Kamada et al., 2014; Yuan et al., 2011; Yuan & Chan, 2008), involves a subtle yet impactful adjustment: it adds a small value to its diagonal elements (i.e., variances of the observed variables). This technique has been demonstrated to significantly improve both the rate of convergence and, potentially, also the accuracy of estimations (Kamada et al., 2014; Kamada & Kano, 2012; Yuan & Bentler, 2017). However, employing techniques beyond ridging, which primarily yields a well-behaved eigenstructure in the sample covariance matrix, may enhance the accuracy of estimation a fortiori.

A considerable number of methods has been developed to regularize the sample covariance matrix in statistics, and applied research fields such as portfolio selection in finance and estimation of large covariance matrices in genomics. *Shrinkage estimation*, a key approach among these, has its origin in the work of Stein (1956), who highlighted the bias in eigenvalues of the sample covariance matrix in small samples. The Steinian (or Stein-type) shrinkage technique creates a weighted average of the sample covariance matrix and a predetermined target matrix, which imposes a specific structure. For instance, using the identity matrix as the target suggests that variances are one, covariances are zero, and eigenvalues are one. The weighting shrinks the sample covariance matrix and their eigenvalues towards those of the target matrix. These approaches vary by the choice of target matrix and how the weighting (or shrinkage) parameter is calculated. In unstable scenarios with small sample size paired with a large number of observed variables ("small N, large p"), shrinkage estimation has been shown to surpass traditional MLE in maintaining eigenstructure and improving accuracy (e.g., Ledoit & Wolf, 2004; Touloumis, 2015). Ledoit and Wolf (2012) concluded that without additional information on the true covariance matrix's structure, shrinkage estimation has been arguably the most effective method so far (for an overview of shrinkage estimation see, e.g., Ledoit & Wolf, 2020).

Even though particularly promising, shrinkage estimation of the covariance matrix has been barely scrutinized in the context of SEM. Notable exceptions include studies by Arruda and Bentler (2017) and De Jonckere and Rosseel (2023), who explored its application in single-level SEM, and found it to enhance overall model evaluation, convergence and accuracy without significant computational costs. Despite these findings, evidence remains sparse, and in multilevel SEM, it is even more so. Here, Zitzmann et al. (2021) applied shrinkage (Bayesian) estimation to the between-group variance of the predictor in a bivariate two-level model which led to more accurate model parameters at the between-group level in small samples. The present article aims to examine the effectiveness of shrinkage estimation of the covariance matrix for handling small sample sizes and low ICCs in multilevel SEMs in a proof of concept manner. More specifically, it scrutinizes whether integrating shrinkage estimation into a two-stage SEM estimation approach improves convergence rates and the accuracy of between-group level parameter estimates. To explore this, we examine balanced, continuous two-level data using two-level intercept-only models by means of a simulation study. The article is structured as follows. Firstly, as we use the single-level CFA approach to multilevel SEM (Barendse & Rosseel, 2020; Mehta & Neale, 2005; Walther et al., 2024), which utilises the data in wide format (WF), we offer a concise overview of this approach. Secondly, we detail the shrinkage approach proposed by Touloumis (2015) that we have adopted in this study, and we elaborate on how it modifies the (co)variances at both levels when applied in multilevel SEM. Thirdly, we present the outcomes of our simulation study, discuss its implications, and suggest directions for future research.

1. Multilevel Structural Equation Modeling

Suppose, we observed four classes (number of groups $g = 4$) with two students within each class (balanced group size $n = 2$). This yields a total sample size of four students ($N = g \cdot n = 4$). We investigate two observed variables

[2]Note that in psychology and the educational sciences, we might be more familiar with terms other than regularization. "Stabilization" is often used in the context of accuracy and model selection (e.g., Breiman, 1996; Ulitzsch et al., 2023; Zitzmann, 2018). "Smoothing" is a prominent term in the context of improving the eigenstructure of covariance matrices (e.g., Lorenzo-Seva & Ferrando, 2021; Wothke, 1993). More recently, "regularization" found its way into the mainstream literature to denote matters related to improper solutions, model sparsity, and overfitting (e.g., Arruda & Bentler, 2017; Jacobucci et al., 2016; Jung & Takane, 2007; Liang & Jacobucci, 2020; Orzek & Voelkle, 2023; Williams & Rodriguez, 2022). However, there is no strict usage of the terms, and we are not aware of any consistent taxonomy.

[3]This means that the sample eigenvalues are more spread out compared to their population counterparts which makes non-invertible (i.e., singular, degenerate), non-positive definite matrices with large condition numbers more likely.

($p = 2$), namely, engagement during class (x_1), and performance in a test (x_2). The whole *data set* is depicted in Panel A in Figure 1. We are interested in whether students within the same class show more similar engagement during class (x_1) and performance in the test (x_2) than students across different classes. In other words, we scrutinize whether variance at the class (between-group) level is substantially large compared to the student (within-group) level; that is, whether $ICC > 0$. The population models are depicted to the right in Panel A. For both variables, $ICC = 0.05$.

To analyze the data, we use a two-level intercept-only model. This model can be estimated through two different multilevel SEM approaches that mainly differ by their required *data format*: the long format (LF) approach (Muthén, 1990, 1994), and the wide format (WF) approach (Barendse & Rosseel, 2020; Mehta & Neale, 2005). The analytical and empirical equivalence of both approaches with MLE in terms of estimation accuracy has been demonstrated (Barendse & Rosseel, 2020; Mehta & Neale, 2005; Walther et al., 2024), and both methods can be implemented using the SEM package *lavaan* in *R*. However, given that the lesser-known WF approach is critical for our two-stage approach, we will focus on this approach in the following, highlighting the differences of the data format in terms of *data matrix*, *sample covariance matrix*, and the *model specification*. Nevertheless, we will consider the unregularized, standard LF approach in the simulation study. Details on model estimation and fitting functions for continuous variables are available in existing literature (e.g., Mehta & Neale, 2005).

1.1. The Wide Format (WF) Approach

The wide format (WF) approach essentially uses a single-level restricted CFA which is fitted to the total (two-level) data matrix in WF (WF-T). In WF-T, every observed variable p is split into every n^{th} unit (see Panel B), which we call "specific-units" variables in contrast to the p "all-units" variables in the LF approach. The rationale underneath is that "people [n] are variables too" (Mehta & Neale, 2005, p. 1). For instance, $x_{1.2}$ is engagement during class (x_1) for every 2nd student in class. The sample covariance matrix is estimated by the MLE for single-level data. Thus, we obtain a single-level represented two-level sample covariance matrix $\mathbf{S}_{WF-T}$ from the $p \cdot n$ "specific-units" variables in WF-T (see Panel C).

In the model, class (between-group) level parameters are modelled by common factors, and student (within-group) level parameters are modelled by unique factors that are equality constrained. The means and (co)variances of the common factors are estimated freely to obtain the class (between-group) level parameters. Variances of the unique factors of each common factor are equality constrained to estimate student (within-group) level variances. Covariances among unique factor of every n-th observed variable of each p are equality constrained to estimate student (within-group) level covariances (see Panel D). The equality constraints represent the homoscedasticity assumption (of the specific-units variables). For our example, we would have two common factors (because of $p = 2$ observed variables) with two observed variables for each common factor (because of $p \cdot n$ specific-units variables in WF-T). Thus, two means, two variances, and one covariance of common factors for the class (between-group) level, and two variances, and one covariance of unique factors for the student (within-group) level are estimated freely.

The implementation of traditional MLE in SEM software such as *lavaan* requires a positive definite sample covariance matrix (Hamaker et al., 2003; Singer, 2010; Van Montfort et al., 2018; Voelkle et al., 2012). Amongst other things, this necessitates a data matrix whose number of columns is less than or equal to the number of rows because otherwise, at least one sample eigenvalue becomes zero and the sample covariance matrix turns non-positive definite (e.g., Duncan et al., 1997; Gorsuch, 1983; Wothke, 1993). In the WF approach, *cols* $\leq$ *rows* translates to $(p \cdot n) \leq g$. Alternatively, the raw data formulation of MLE, full information maximum likelihood (FIML), may be used, which circumvents the problem (Hamaker et al., 2003; Trendafilov & Unkel, 2011; Unkel & Trendafilov, 2010; Voelkle et al., 2012). In *lavaan*, FIML estimation could be applied by setting 'missing = "fiml"'. However, since we aim to replace the sample covariance matrix with a shrinkage estimate that has an improved eigenstructure, we must use traditional MLE instead of FIML. Moreover, we must use single-level SEM (i.e., the WF approach), because in multilevel SEM, such as implemented in *lavaan* version 0.6–15, we cannot provide a covariance matrix instead of a data matrix. We will turn towards shrinkage estimation in the subsequent section.

2. Shrinkage Estimation of the Covariance Matrix

In shrinkage estimation, the population covariance matrix $\mathbf{\Sigma}$ is estimated as a weighted average of the sample covariance matrix and a pre-specified target matrix. The amount of weighting is controlled by the shrinkage parameter $\lambda \in [0, 1]$. If $\lambda = 0$, no shrinkage is applied, and the sample covariance matrix will be kept. If $\lambda = 1$, we obtain the target matrix as the estimate of $\mathbf{\Sigma}$. In linear shrinkage, which we focus on, the same shrinkage intensity is applied to every element of the covariance matrix. To avoid misunderstanding: "shrinkage" does not necessarily mean that the elements get smaller, but they are shrunken towards a certain value (of the target matrix). For example, if $\sigma_S = 0.1$ and $\sigma_T = 1$, then 0.1 is "shrunken" towards 1. The target matrix is chosen for its well-behaved eigenstructure, making shrinkage estimates more likely to be positive definite, non-singular, and well-conditioned, often resulting in greater accuracy compared to the traditional ML sample covariance matrix, as demonstrated in studies such as Ledoit and Wolf (2004, 2020). Additionally, shrinkage estimation can be viewed as a form of Bayesian estimation with weakly informative priors, a perspective supported by Ledoit and Wolf (2004), and others. In the next section, we will briefly review the linear shrinkage estimator proposed by Touloumis (2015), which we term *covshrink* for convenience.

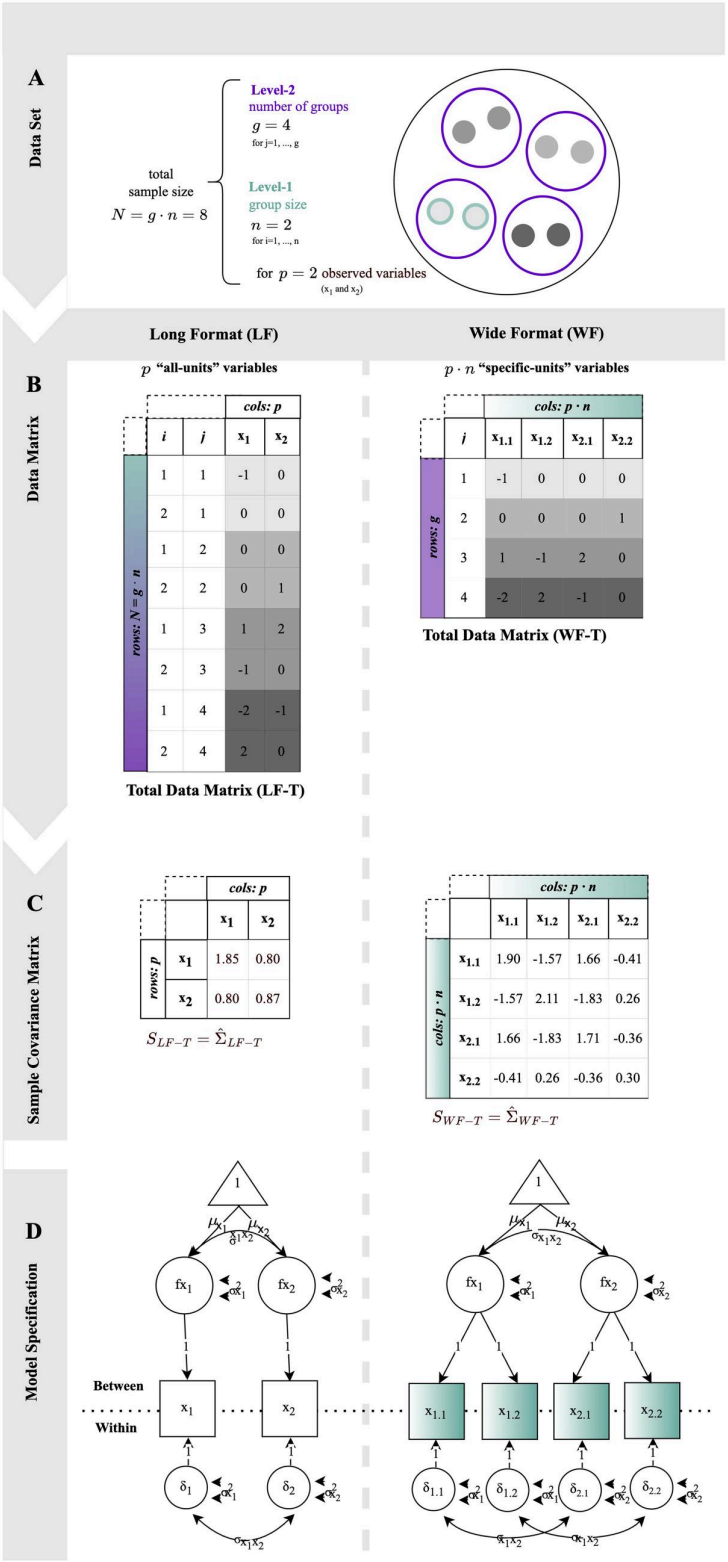


Figure 1. Data and model in the long format (LF) and wide format (WF) approach.
Note. Data set: the data collected in a given setting. *Data Matrix*: the data set in matrix form, where columns refer to observed variables and rows to observed units. *Data Format*: one of two possible formats of the data matrix, long format (LF) or wide format (WF). In WF, every observed variable p is split for every unit in the group n. For instance, $x_{1.2}$ is x_1 for every 2nd unit in the group. *Sample Covariance Matrix*: a symmetric matrix which contains (co)variances of the observed variables. *Model Specification*: representation of the model to be estimated, here, a two-level intercept-only model. Between-group parameters are located above the dashed line; within-group parameters below. At the within-group level, identical parameters indicate equality constrains. Data matrix or sample covariance matrix, and model specification are input to *lavaan*. Example data set with number of groups $g = 4$, group size $n = 2$, and number of observed variables $p = 2$. The R code to generate data and model is available in Appendix A.

2.1. Covshrink: A Linear Shrinkage Estimator of the Covariance Matrix

Touloumis (2015) refined the popular linear shrinkage estimator by Ledoit and Wolf (2004) through (1) extending the set of target matrices, and (2) deriving consistent closed form solutions of the shrinkage parameters in "small N, large p" settings. This new family of estimators has demonstrated improved estimation compared to preceding methods, indicated by the simulated percentage relative improvement in average loss (SPRIAL), which compares the MSE of the target estimator to that of a baseline estimator (e.g., the sample covariance matrix), in such settings (Touloumis, 2015). The general equation for the linear shrinkage estimator is expressed as:

$$\hat{\mathbf{S}}^{*} = (1 - \hat{\lambda})\mathbf{S} + \hat{\lambda}\mathbf{T}, \tag{1}$$

where S is the unbiased MLE of the (single-level) $p \times p$ population covariance matrix, T is the target matrix, and $\hat{\lambda}$ is the shrinkage parameter, which depends on the choice of the target matrix. The target matrix can be one of three diagonal matrices: the equal target matrix $\hat{v}\mathbf{I}_p$ with the mean of the sample variances in the diagonal (originally proposed by Ledoit & Wolf, 2004), the identity matrix $\mathbf{I}_p$ with ones in the diagonal, or the unequal target matrix $\mathbf{D}_S$ with the sample variances in the diagonal. Across all types of target matrices, off-diagonal elements (i.e., covariances) of the shrinkage estimate are systematically pulled towards zero. However, the specific non-zero value to which on-diagonal elements (i.e., variances) are pulled varies depending on the target matrix employed. When using the equal target matrix $\hat{v}\mathbf{I}_p$, variances are pulled towards the mean of the sample variances. Meanwhile, the identity matrix $\mathbf{I}_p$ pulls variances towards one, while the unequal target matrix $\mathbf{D}_S$ leaves the variances unchanged. The closed form solution of the shrinkage parameter of the equal matrix $\hat{v}\mathbf{I}_p$, where $\hat{v} = Y_{1N/p}$, is:

$$\hat{\lambda}_E = \frac{Y_{2N} + Y_{1N}^2}{NY_{2N} + Y_{1N}^2 + \frac{p-N+1}{p} Y_{1N}^2}, \tag{2}$$

for the shrinkage parameter of the identity matrix $\mathbf{I}_p$:

$$\hat{\lambda}_I = \frac{Y_{2N} + Y_{1N}^2}{NY_{2N} + Y_{1N}^2 - (N-1)(2Y_{1N} - p)}, \tag{3}$$

and for the shrinkage parameter of the unequal target matrix $\mathbf{D}_S$:

$$\hat{\lambda}_U = \frac{Y_{2N} + Y_{1N}^2 - 2Y_{3N}}{NY_{2N} + Y_{1N}^2 - (N-1)Y_{3N}}, \tag{4}$$

where Y_{1N}, Y_{2N}, and Y_{3N} are combinations of U-statistics (for their estimation, see Touloumis, 2015, pp. 5, 12). According to Touloumis (2015), the optimal shrinkage intensity, which minimizes the MSE between the population covariance matrix and the respective shrinkage estimator, is approximated by sample-based unbiased and ratio-consistent estimators. The resulting biased shrinkage estimators of $\boldsymbol{\Sigma}$ are $\hat{\mathbf{S}}_E^{*}$ (equal target matrix), $\hat{\mathbf{S}}_I^{*}$ (identity target matrix), and $\hat{\mathbf{S}}_U^{*}$ (unequal target matrix). Because the shrinkage parameters have a closed form, the approach is computationally fast, regardless of the number of observed variables p. Moreover, the obtained estimates are non-singular and well-conditioned. These are useful properties for convergence (e.g., *lavaan* requires a positive definite sample covariance matrix in single-level SEM), and estimation accuracy (e.g., large condition numbers have been linked to the less stable estimates; Y. Huang & Bentler, 2015; Kelley, 1995; Lange et al., 1999; Yuan & Bentler, 2017).

3. Shrinkage Estimation of the Covariance Matrix in Multilevel Structural Equation Modeling

Shrinkage estimation of the covariance matrix is part of our two-stage approach.[4] At the first stage, the sample covariance matrix is replaced by a shrinkage estimate of $\boldsymbol{\Sigma}$. At the second stage, the model is estimated based on this refined estimate. Touloumis (2015) shrinkage estimator was optimized for "small N, large p" scenarios which, can be translated to "small g, small n, large p" configurations in the context of multilevel analysis. While this two-stage approach appears to be a resource-efficient strategy for addressing issues such as non-convergence and inaccurate between-group level parameter estimates resulting from small samples or low ICCs, only a limited body of research has investigated the performance of such methods within the SEM framework (Arruda & Bentler, 2017; De Jonckere & Rosseel, 2023; Zitzmann et al., 2021). Existing evidence suggests that similar two-stage approaches can indeed enhance convergence and estimation accuracy. However, such an approach has not yet been proposed and investigated in the context of multilevel SEM. In the subsequent section, we will delve deeper into how the (co)variances in the shrinkage estimate differ from those in the sample covariance matrix, and elucidate the implications for model parameters.

3.1. WFcovshrink: Shrinkage Estimation of the Covariance Matrix in the WF Approach

Recall that the WF approach is a single-level SEM approach that utilises the single-level represented two-level sample covariance matrix $\mathbf{S}_{WF-T}$ where the (co)variances of $p \cdot n$ "specific-units" variables are contained (revisit Figure 1 for more details). The normal theory derived, biased MLE reads:

$$\mathbf{S}_{WF-T} = \frac{1}{g}\sum_{j=1}^{g}(\mathbf{X}.\mathbf{i}_j - \overline{\mathbf{X}.\mathbf{i}})(\mathbf{X}.\mathbf{i}_j - \overline{\mathbf{X}.\mathbf{i}})^{T}, \tag{5}$$

where X.i denotes the data matrix in WF (WF-T) and $\overline{\mathbf{X}.\mathbf{i}}$ denotes a row vector with grand mean estimates. The sample covariance matrix is the estimate of the population covariance matrix, $\mathbf{S}_{WF-T} = \hat{\boldsymbol{\Sigma}}_{WF-T}$. When shrinkage estimation of the covariance matrix is applied, the (single-level) $p \times p$ dimensioned S is replaced by the (single-level

[4]Note that in fact, every SEM is a two-stage approach as the sample covariance matrix has to be estimated in order to estimate the model parameters. Nonetheless, usually users supply the data matrix and the software estimates the sample covariance matrix automatically. Thus, from a user perspective, standard SEM can be considered a one-stage approach.

represented two-level) $(p \cdot n) \times (p \cdot n)$ dimensioned $\mathbf{S}_{WF-T}$[5] in Equation (1), and N by g, and p by $p \cdot n$ in Equations (2), (3), and (4). Within the present study, we scrutinize all three target matrices, resulting in the shrinkage estimators with the equal target matrix, $\hat{\mathbf{S}}_E^*$, the identity target matrix, $\hat{\mathbf{S}}_I^*$, and the unequal target matrix, $\hat{\mathbf{S}}_U^*$.

For an illustration of the effect of the shrinkage estimation by Touloumis (2015) in the WF approach (WFcovshrink) in the following, we focus on $\hat{\mathbf{S}}_E^*$, see Figure 2. In Panel A, it is highlighted how $\hat{\mathbf{\Sigma}}_{WF-T}$ is used to model $\hat{\theta}$. In Panel B, the principle of how the shrinkage estimate $\hat{\mathbf{S}}_E^*$ alters $\hat{\theta}$ is explained. In Panel C, a concrete example is presented.

In Panel A, leftmost, we see the (earlier introduced) model specification of the two-level intercept-only model in the WF approach. A restricted CFA is fitted to the $p \cdot n$ "specific-units" variables in the data matrix in WF. In the middle, the (co)variances of these $p \cdot n$ "specific-units" variables in $\mathbf{S}_{WF-T}$ are shown. To the right, these $p \cdot n$ "specific-units" (co)variances are reformulated as the p "all-units" (co)variances that are modelled thereof. (Co)variances of $x_{1.1}$ and $x_{1.2}$ (see upper left, green block) are used to model the variances of one common and two unique factors which correspond to the between-group and within-group level variances of x_1. Their variances contribute to the between-group and within-group level variances, whereas their covariance contributes only to the between-group level variance via the common factor. Similarly, (co)variances of $x_{2.1}$ and $x_{2.2}$ (see lower right, green block) are used to model between-group and within-group level variances of x_2. The covariances of $x_{1.1}$ and $x_{1.2}$ with $x_{2.1}$ and $x_{2.2}$, respectively (see lower left or upper right, orange block), are used to model the covariances of the two common factors, and every n-th unique factor of each common factor which correspond to between-group and within-group level covariances of x_1 and x_2.

This reformulation helps to understand the principle of how shrinkage estimation with the equal target matrix alters the estimates of the two-level intercept-only model $(\hat{\theta})$, which is illustrated in Panel B. To the left, the reformulated $\mathbf{S}_{WF-T}$ is shown again. The on-diagonal elements of $\mathbf{S}_{WF-T}$ (grey bar) are averaged $(\hat{v})$ and used as the on-diagonal elements ("equal variances") in the equal target matrix $\hat{v}\mathbf{I}_{p \cdot n}$. In reformulated terms, $\hat{v}$ is the grand mean of the total variances of both variables x_1 and x_2 $(\overline{\hat{\sigma}_B^2 + \hat{\sigma}_W^2})$. The off-diagonal elements of $\hat{v}\mathbf{I}_{p \cdot n}$ are zero. To the right, we see an overview of the directions in which the sample (co)variances in $\mathbf{S}_{WF-T}$ are pulled by shrinkage estimation. Generally, on-diagonal elements are pulled towards the mean of the diagonal elements, and off-diagonal elements are pulled towards zero. Using the reformulation, this means that total variances $(\hat{\sigma}_B^2 + \hat{\sigma}_W^2)$ are pulled towards the grand mean of the total variances $(\overline{\hat{\sigma}_B^2 + \hat{\sigma}_W^2})$, and between-group variances $(\hat{\sigma}_B^2)$ are pulled towards zero. Consequently, within-group variances $(\hat{\sigma}_W^2)$ are pulled towards the grand mean of the total variances $(\overline{\hat{\sigma}_B^2 + \hat{\sigma}_W^2})$, too. Between-group covariances $(\hat{\sigma}_B)$ and within-group covariances $(\hat{\sigma}_W)$ are pulled towards zero. The expected biases in $\hat{\theta}$ are depicted in the rightmost table. It is expected that between-group level variances, and between-group and within-group covariances, have downward biases, whereas within-group variances have an upward biases. Therefore, estimates of ICC will be more conservative than those derived by the unregularized WF approach.

Let us consider this more concretely. In Panel C, WFcovshrink is illustrated by means of an example data set (in which $g = 50$ in contrast to the earlier example data set). Leftmost, $\mathbf{S}_{WF-T}$ (estimated by the unbiased MLE) is depicted. The middle of the panel shows the equal target matrix $\hat{v}\mathbf{I}_{p \cdot n}$ where $\hat{v} = 0.99$ (mean of the sample variances in $\mathbf{S}_{WF-T}$, or reformulated, grand mean of the total variances $\overline{\hat{\sigma}_B^2 + \hat{\sigma}_W^2}$ of x_1 and x_2). In the present medium g, small n, large p setting, the shrinkage parameter is $\hat{\lambda} = 0.59$, and thus, the shrinkage estimate is to a large extent influenced by the target matrix. To the right, the resulting shrinkage estimate $\mathbf{S}_E^*$ is presented. To comprehend how $\mathbf{S}_E^*$ alters $\hat{\theta}$, we compare the model parameter estimates retrieved from the WF approach and the WFcovshrink approach (the R code is available in the Appendix A). In this example, population parameters are $\sigma_B^2 = 0.05$, $\sigma_W^2 = 0.95$, thus $ICC = 0.05$, and $\sigma_B = 0.015$, and $\sigma_W = 0.285$ for both variables x_1 and x_2.

It can be seen in Table 1 that for the between-group level, over- and underestimation was decreased (by pulling the estimates closer to zero). For the within-group level, underestimation of the variance of x_1 was decreased but overestimation slightly increased for x_1 (by pulling the estimates of the variances closer to their grand mean), and overestimation of their covariance was decreased (by pulling the estimate closer to zero).

In both approaches, one estimate of variances at the between-group level was negative.[6] Concerning the resulting ICCs, the estimates of the WFcovshrink approach $(0.06/(0.06 + 0.95) = 0.06$ and $-0.04/(-0.04 + 1.02) = -0.04)$ were more accurate than those of the unregularized

[5]Note that in the implementation of the shrinkage estimation in R (*ShrinkCovMat* package), only the unbiased MLE, which has g–1 in the denominator, can be used. In contrast, the default of single-level SEM in *lavaan* (i.e., WF approach) is the normal theory derived, biased MLE in Equation 5 (Rosseel et al., 2023, reference manual p.81 accessed on 16 September 2023, lav_matrix_cov function). We run the unregularized WF approach with both the unbiased and the biased MLE to check whether they differ substantially. In Figure B1 in the Appendix we see that for convergence there were no differences, and for estimation accuracy, there were negligible differences in using the unbiased or biased estimator of the sample covariance matrix.

[6]The unregularized WF approach that uses the ML sample covariance matrix has high variability and low bias in small samples, whereas the shrinkage estimate in the WFcovshrink approach reduces variability by means of bias. Thus, in other data sets, the unregularized WF approach may yield non-negative, overestimated between-group level variances. There are different procedures to deal with inadmissible, negative between-group level variances (so called "Heywood cases"), that we would expect (and empirically found in the present study) more often in the downwardly biased WFcovshrink approach, for instance, setting them to zero (see e.g., Zitzmann et al., 2022, who justify the procedure by the very definition of MLE). However, non-negative, overestimated between-group level variance, which we might expect more often in the unregularized WF approach, are taken at face value. Thus, the downward bias in the WFcovshrink approach might be dealt with better (in addition to its estimates being more accurate).

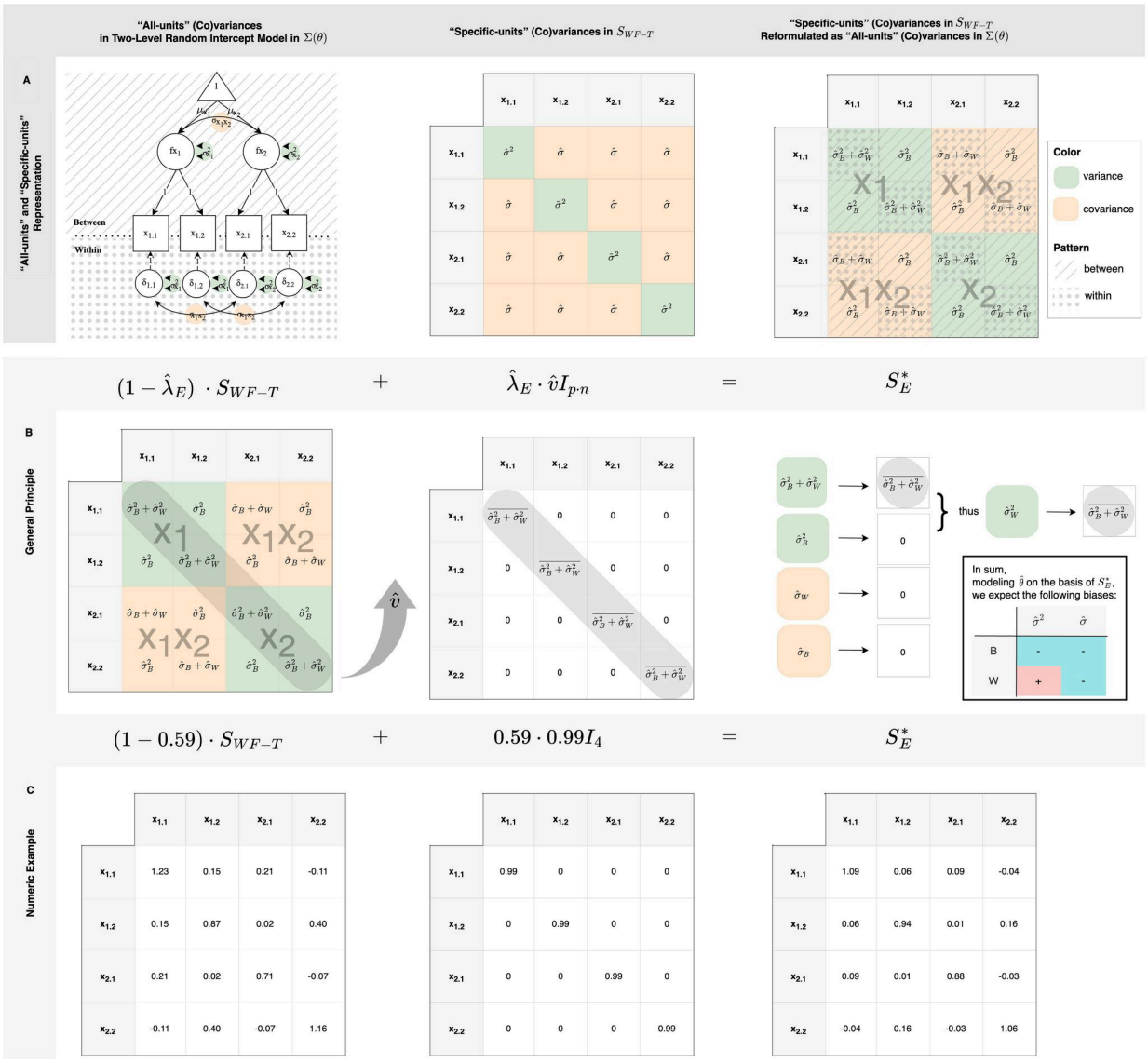


Figure 2. Shrinkage estimation of the covariance matrix in the WF approach.
Note. $\mathbf{S}_{WF-T}$ contains (co)variances of $p \cdot n$ "specific-units" variables. In Panel A, these are reformulated as (co)variances of p "all-units" variables modelled in the two-level intercept-only model. Panel B introduces the principle of how the shrinkage estimate with the equal target matrix alters estimates of the two-level intercept-only model. In Panel C, a numeric example with the earlier data set (number of groups $g = 50$, group size $n = 2$, and number of observed variables $p = 2$) is given. The R code to generate the (unbiased) sample covariance matrix and apply shrinkage estimation is available in Appendix A.

Table 1. Model parameter estimates of two-level intercept-only model for example data set.

Approach	Between			Within		
	$\sigma^2_{x1} = 0.05$	$\sigma^2_{x2} = 0.05$	$\sigma_{x1x2} = 0.015$	$\sigma^2_{x1} = 0.95$	$\sigma^2_{x2} = 0.95$	$\sigma_{x1x2} = 0.285$
$\hat{\theta}_{WF}$	0.15	−0.08	−0.05	0.88	1.01	0.35
$\hat{\theta}_{WFcovshrink_E}$	0.06	−0.04	−0.02	0.95	1.02	0.15

Note. Example data set with number of groups $g = 50$, group size $n = 2$, and number of observed variables $p = 2$. The R code to generate data and estimate the models is available in Appendix A.

WF approach $(0.15/(0.15 + 0.88) = 0.14$ and $-0.08/(-0.08 + 1.01) = -0.08)$. In sum, all but one estimate are closer to their population counterparts in WFcovshrink compared to the unregularized WF approach. Nevertheless, this was just one example data set. Whether WFcovshrink yields empirically reliable similar gains in performance in other settings, and by means of other target matrices than the equal target matrix $\hat{v}\mathbf{I}_{p\cdot n}$, remains to be put to test. We addressed these questions with a simulation study, which we will present next.

4. Simulation Study

With this simulation study, we aimed to investigate whether applying shrinkage estimation, as part of the two-stage SEM

estimation approach, would increase convergence and estimation accuracy in multilevel SEM when small samples at any level or small ICCs are present. The idea is to obtain a biased but more precise estimate of the covariance matrix $\boldsymbol{\Sigma}$ that yields more accurate model parameters $\hat{\theta}$ in turn. Specifically, we applied the shrinkage estimator by Touloumis (2015) to the WF approach in multilevel SEM. In the following, we outline the method of our study before presenting and discussing the main findings.

4.1. Method

The computations were conducted on an AMD Ryzen Threadripper PRO 3975WX 32-cores (3.50 GHz) CPU on a Windows 10 (Version 20H2) platform utilising R version 4.3.1 (R Core Team, 2023), along with several R packages: *cowplot* version 1.1.1 (Wilke, 2020), *DescTools* version 0.99.50 (Signorell et al., 2024), *dplyr* version 1.1.2 (Wickham et al., 2023), *ggplot2* version 3.4.2 (Wickham et al., 2023), *huxtable* version 5.5.6 (Hugh-Jones, 2022), *lavaan* version 0.6-15 (Rosseel et al., 2023), *patchwork* version 1.1.2 (Pedersen, 2022), *ShrinkCovMat* version 1.4.0 (Touloumis, 2019), *tidyr* version 1.3.0 (Wickham et al., 2022), and *xlsx* version 0.6.5 (Dragulescu & Arendt, 2020). The R code for data generation, analysis, table, and figures is available at https://github.com/demianJK/WFcovshrink.

4.1.1. Data Generation

We varied different factors that we allocate to either *sample characteristics* or *population characteristics* to facilitate interpretation. We make this distinction to emphasize what we can modify (by our study design) and what not. Sample characteristics comprise the number of groups g, the group size n, and the number of observed variables p. We included the following numbers of groups: 4, 10, 30, 50, and 100. The smallest number of groups is given by the minimum sample size that the R function for shrinkage estimation can deal with (which relates to g in the WF approach). The maximum number of groups was chosen to see how the WFcovshrink approaches perform in samples large enough to achieve good performance by the unregularized LF and WF approaches to multilevel SEM. The group size was varied between 2, 5, and 10. We restrained the upper group size to 10, because the WF approach is rather advised for smaller n scenarios (Barendse & Rosseel, 2020; Walther et al., 2024), because of larger computational costs, and preliminary simulations supported that this holds true for WFcovshrink as well. As numbers of observed variables p, we selected 2, 5, and 10. The population characteristics encompass the variances and covariances of the population covariance matrix at both the between- and within-group level. The variance at both levels was determined by the ICC, which is defined as the ratio of between-group variance to the total variance (Hox et al., 2017), $\sigma_B^2/(\sigma_B^2 + \sigma_W^2)$. Two levels of the ICC were included, 0.05 and 0.25, which represent the lower and upper levels of realistic ICCs in the social sciences (Adams et al., 2004; Gulliford et al., 1999). The total variances were set to 1, and thus, $ICC = \sigma_B^2$. The covariances were determined by the correlation at each level. Correlations of .10 and .30 were chosen, inspired by meta-analytically derived small and large correlations in the social sciences (Gignac & Szodorai, 2016). Covariances were calculated through the variance and the correlation. The combination of all factor levels in our simulation study resulted in a fully-crossed design with 360 conditions. For each condition, 1000 data sets were simulated.

4.1.2. Data Analysis

Two-Level Intercept-Only Model: As pointed out earlier, we considered only the two-level intercept-only model or, put differently, a model that estimates the (co)variances of the p all-units variables at the between-group and within-group levels, and the means of the between-group level, freely. We did so because various structured models (e.g., x_1 as predictor of x_2 or the other way around) have the same underlying covariance matrix, and we were primarily interested in examining the effects of shrinkage estimation of the covariance matrix on model performance.

Approaches: We compared the performance of the proposed two-stage estimation WFcovshrink to the unregularized WF approach, and the unregularized, standard LF approach. For WFcovshrink, we scrutinized a consistent usage of all target matrices: the equal target matrix in WFcovshrink(E), the identity matrix in WFcovshrink(I), and the unequal target matrix in WFcovshrink(U).

4.1.3. Evaluation Criteria

We conducted comparisons of model performance based on convergence and estimation accuracy. A model was deemed converged if the optimizer indicated that it had found a solution. The convergence rate represents the percentage of converged models out of the total number of estimated models per condition. Estimation accuracy was evaluated in terms of bias and overall accuracy (which incorporates both bias and variance of an estimator). We considered the relative bias, $\sum(\hat{\theta} - \theta)/\theta \cdot 100\%$, and the relative root mean squared error (RMSE), $\sqrt{\sum (\hat{\theta} - \theta)^2}/\theta \cdot 100\%$.

4.2. Results

Hereinafter, we will delve into the key findings of the simulation study. We will commence by examining convergence, followed by a discussion on estimation accuracy (bias and overall accuracy). Readers interested in further results are referred to the supplementary materials provided in the Appendix. To summarize, we found evidence that the input type (data or sample covariance matrix) and the type of MLE of the sample covariance matrix (biased or unbiased) did not influence the performance of the WF approach substantially (Figure B1), that the WFcovshrink approaches had no severely increased computation times in contrast to the WF approach (Figure B2), and that the WFcovshrink approaches yielded higher percentages of negatively estimated between-group level

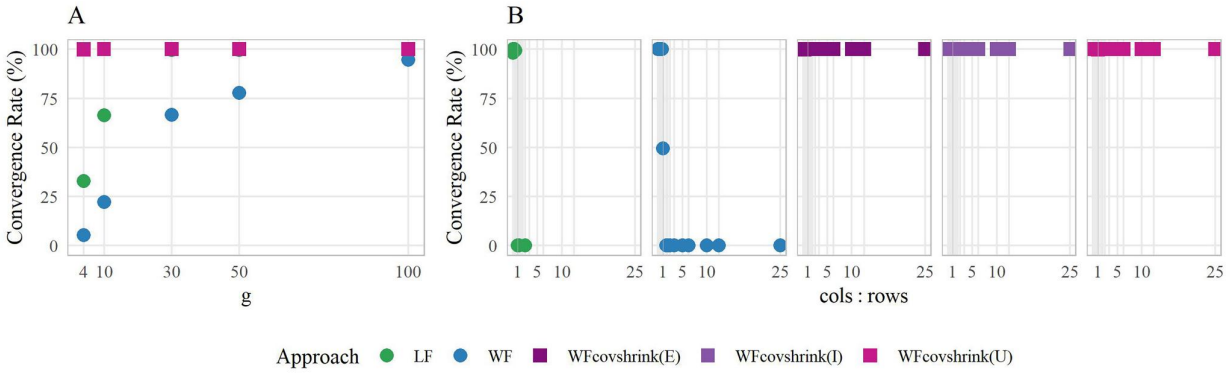


Figure 3. Convergence rates by sample characteristics.
Note. g = number of groups; WFcovshrink(E) = equal target matrix; WFcovshrink(I) = identity target matrix; WFcovshrink(U) = unequal target matrix. The *cols*:*rows* refer to those of long format between-group data matrix LF-B ($p : g$) and the wide format total data matrix WF-T ($(p \cdot n) : g$) for the LF and WF approaches, respectively.

variances and ICC (which we will link later to an increased downward bias).

4.2.1. Convergence

In terms of sample characteristics, the sample size at level-2, g, proved to be the most influential factor affecting convergence. As illustrated in Panel A of Figure 3, convergence rates aggregated by g revealed a typical observation: for the LF and WF approaches, convergence rates increased with increasing sample size. In contrast, the WFcovshrink approach consistently converged across all sample sizes. Previous research (Walther et al., 2024) has highlighted the significance of the relationship between the columns and rows of the data matrices in understanding convergence rates, as depicted in Panel B. Replicating earlier findings, we observed that $cols < rows$ and $cols \leq rows$ are required for converging models in the LF and WF approaches, respectively. Additionally, the LF approach tended to converge in more diverse settings because satisfying $p < g$ (in the long format between-group data matrix LF-B) is easier than satisfying $(p \cdot n) \leq g$ (in the wide format total data matrix WF-T) (Walther et al., 2024). Notably, this restriction did not apply to the WFcovshrink approaches, regardless of *cols*:*rows* of WF-T. It is interesting to note that convergence rates did neither significantly differ by the number of observed variables (p) nor the population characteristics.

4.2.2. Estimation Accuracy

In the following, we review the estimation accuracy of parameters derived from the LF, WF, and WFcovshrink approaches. Firstly, we examine the relative bias for the model parameters, and secondly, the overall estimation accuracy (relative RMSE) at the between- and within-group levels and the thereof estimated ICCs. Note that for all estimation accuracy parameters, we only considered settings resulting in convergence rates greater than zero across all approaches. Given the WF approach's tendency to exhibit the lowest convergence rates, this implies that we exclusively considered settings where $p \cdot n \leq g$. Otherwise, comparing estimation accuracy measures aggregated by different settings would lead to unfair comparisons, as the WFcovshrink approaches consistently converged, even in practically unrealistic, highly inaccurate settings (e.g., $g = 4$, $n = 2$, and $p = 10$).

Bias: We focused on four types of parameters of the random-intercept models, variances and covariances at the between- and within-group level, respectively, and the thereof estimated ICCs. These are depicted in Figure 4. Overall we see, as expected from the known bias-variance tradeoff, that the WFcovshrink approaches had increased biases in contrast to the unregularized approaches. Moreover, when comparing the hypothesized direction of biases in Panel B of Figure 2 with the actual empirical observations, we found a match between our hypotheses and the observed evidence. More specifically, at the between-group level, both variances and covariances exhibited a tendency towards underestimation. Conversely, at the within-group level, the unregularized approaches were unbiased regardless of the sample size (number of groups g), while the WFcovshrink approaches introduced an upward bias in variances, and a downward bias in covariances. Following from this, all approaches exhibited a downward bias in the estimates of the ICC, and the WFcovshrink approaches consistently yielded a more conservative underestimation. This downward bias trend in the between-group level variances and the estimates of the ICC was further evidenced by the significant proportion of negatively estimated variances and ICCs in the WFcovshrink approach, as illustrated in Figure B3 in the Appendix.

Overall Estimation Accuracy: In the upper panel of Figure 5, the relative RMSE of the between-group parameters aggregated by number of groups g and group size n is shown. Overall, smaller numbers of groups g and group sizes n resulted in less accurate estimates. However, the WFcovshrink approaches consistently yielded more accurate estimates, with the most significant improvements observed in settings with small g and especially small n samples. For example, in a more realistic scenario with a group size of 5 and 50 groups, the unregularized approaches produced relative RMSEs of approximately 200%, while the WFcovshrink approaches reduced it by half. In the middle panel, which

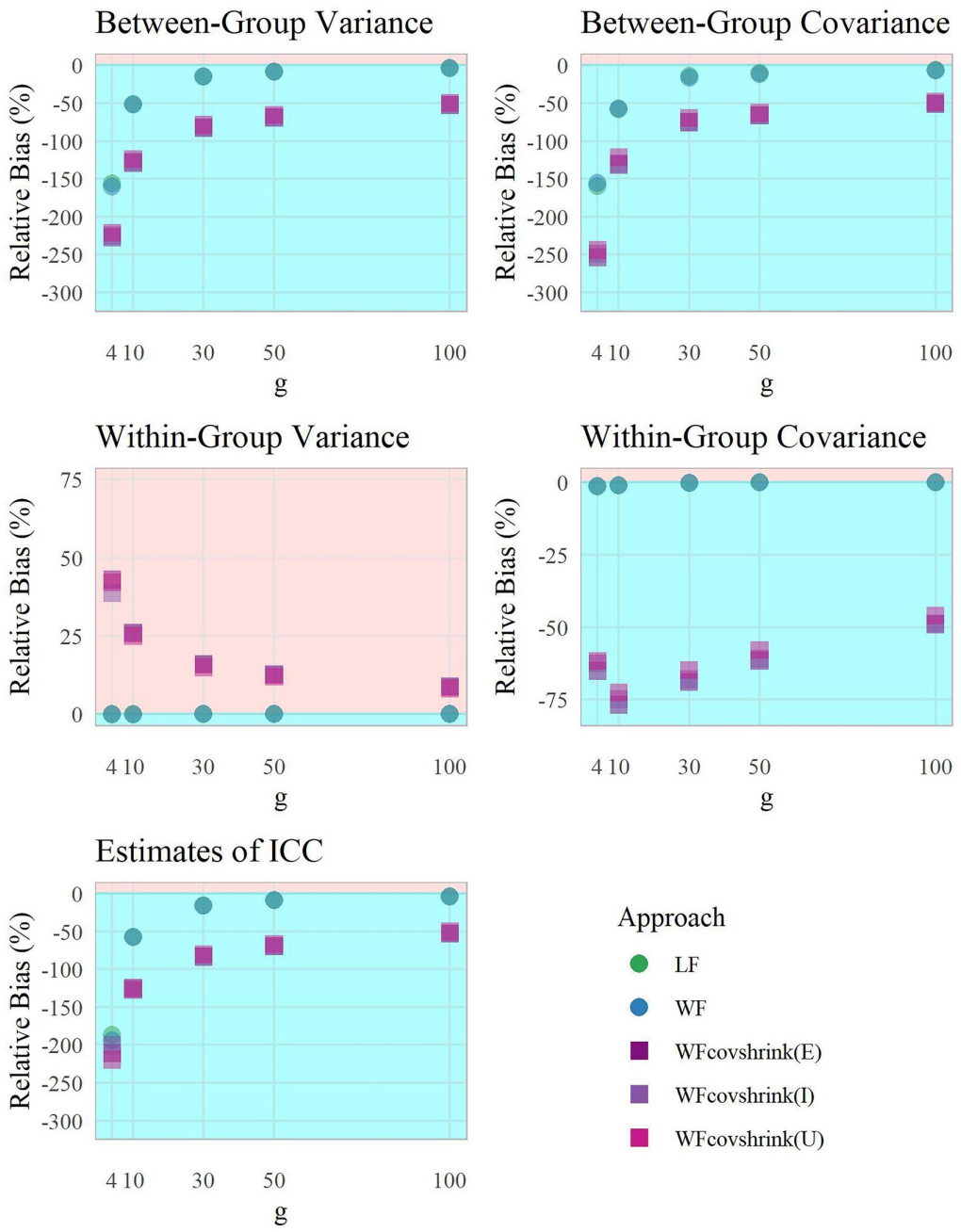


Figure 4. Relative bias of parameter estimates.
Note. g = number of groups; WFcovshrink(E) = equal target matrix; WFcovshrink(I) = identity target matrix; WFcovshrink(U) = unequal target matrix.

focuses on the within-group level, it can be seen that the WFcovshrink approaches were generally more accurate than the unregularized approaches only when the group size was very small ($n = 2$). However, as the group sizes increased, the estimates from the WFcovshrink approaches tended to be somewhat less accurate. Returning to the setting with a group size of 5 and 50 groups, the unregularized approaches exhibited an average relative RMSE of approximately 30%, whereas the WFcovshrink approaches showed an average relative RMSE of around 45%. The relative RMSE of the ICC estimates are shown in the lower panel. They combine the results of the between- and within-group level: when the group size was very small ($n = 2$), the WFcovshrink approaches were more accurate, particularly, the smaller the number of groups g were, but when the group size n became larger, the unregularized approaches became more accurate. In the setting with a group size of 5 and 50 groups, the estimated ICCs showed an average relative RMSE of approximately 75% in the WFcovshrink approaches, and approximately 65% in the unregularized approaches. In sum, the benefit of the WFcovshrink approaches appealed to the between-group parameters in all

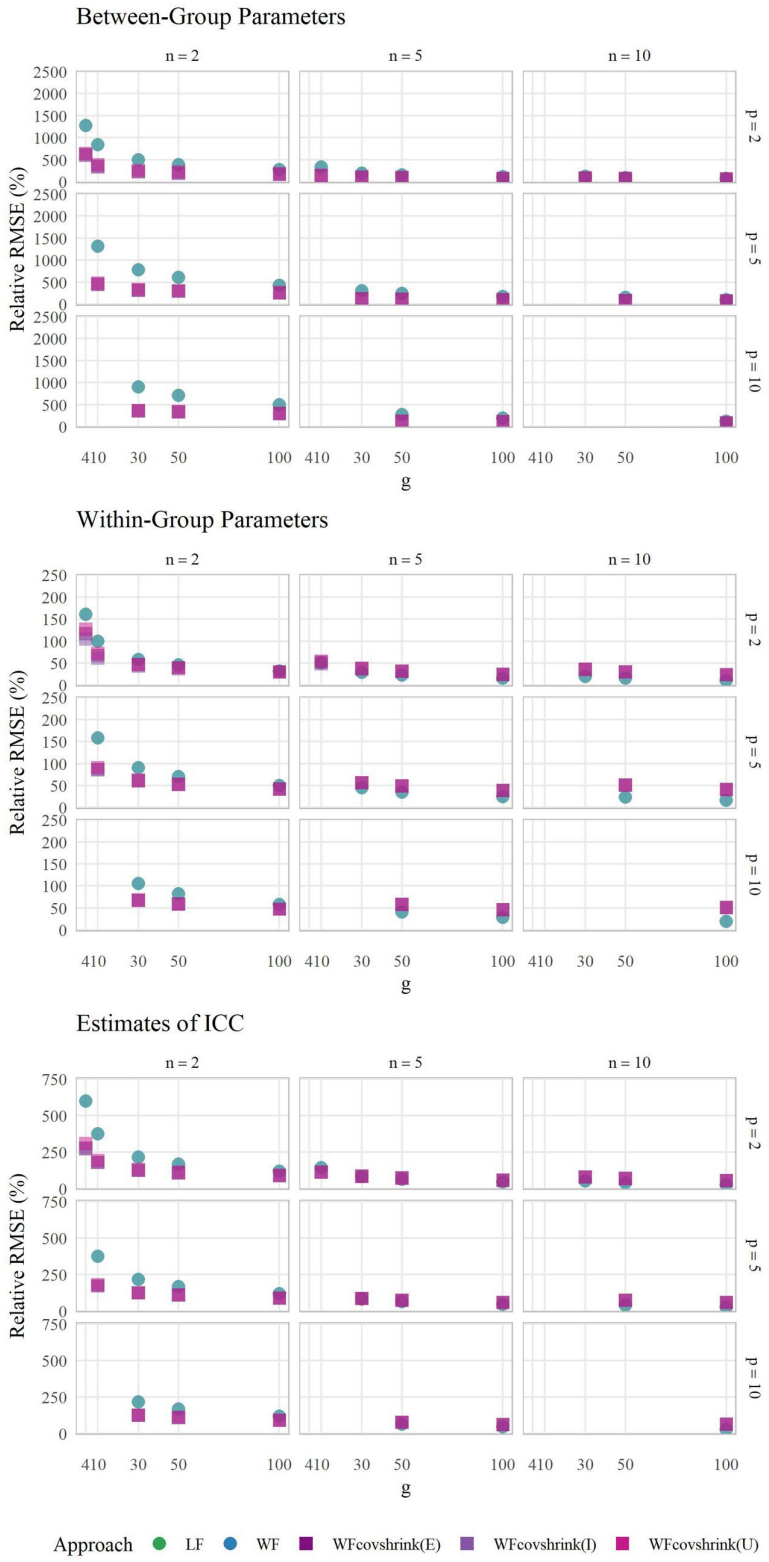


Figure 5. Overall estimation accuracy by sample characteristics.
Note. g = number of groups; n = group size; p = number of observed variables; WFcovshrink(E) = equal target matrix; WFcovshrink(I) = identity target matrix; WFcovshrink(U) = unequal target matrix.

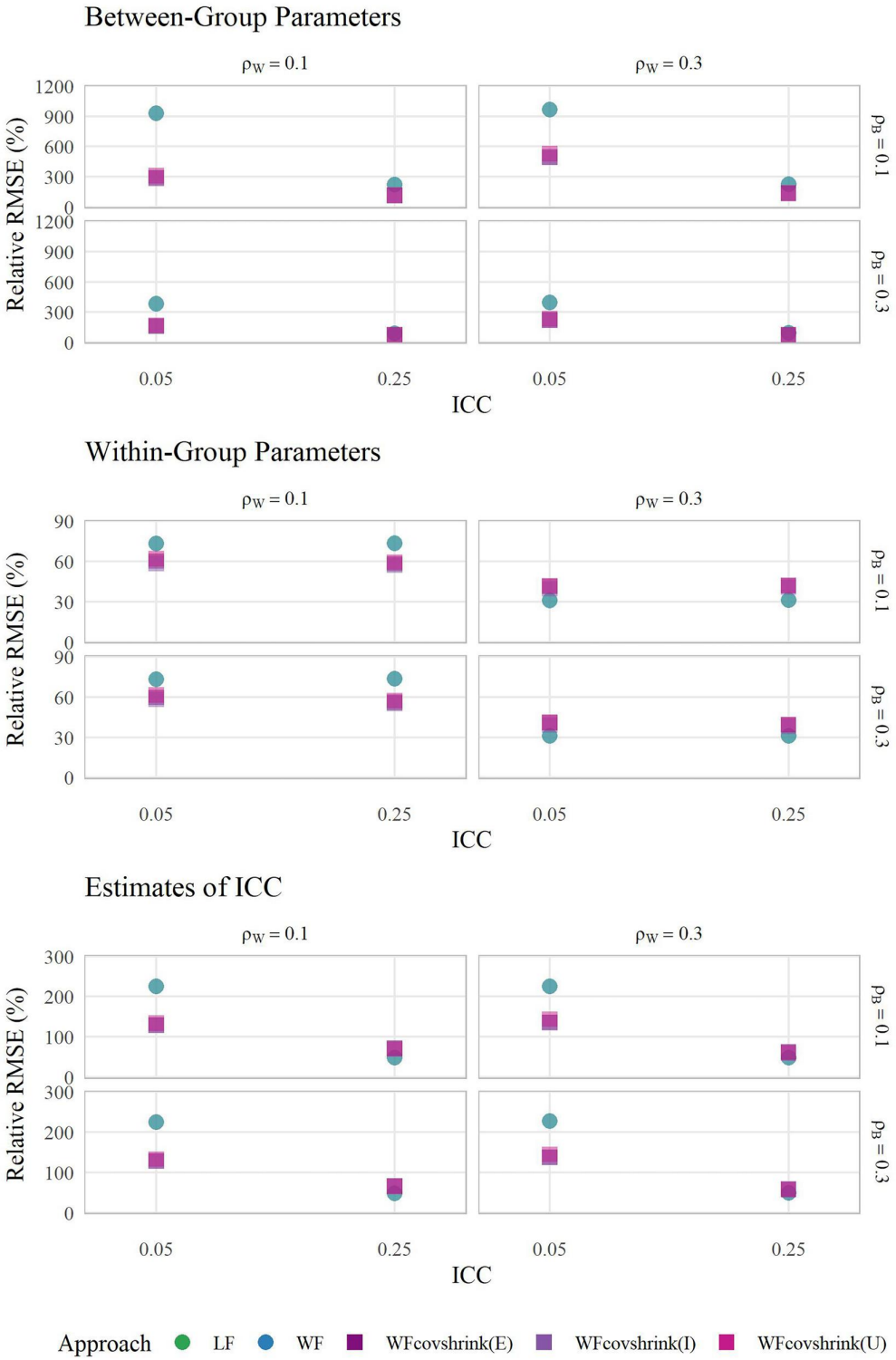


Figure 6. Overall estimation accuracy by population characteristics.
Note. ICC = Intraclass Correlation; ρ_B = correlation at between-group level; ρ_W = correlation at within-group level; WFcovshrink(E) = equal target matrix; WFcovshrink(I) = identity target matrix; WFcovshrink(U) = unequal target matrix.

settings, including small group sizes ($n \leq 10$) for small to moderate numbers of groups ($g \leq 100$), but to within-group and ICC parameters only in very small group sizes settings ($n = 2$) for small to moderate number of groups ($g \leq 100$).

Figure 6 illustrates the estimation accuracy influenced by the ICC and the correlations of variables in the population. The upper panel, which focuses on the between-group parameters, shows that smaller ICC values (indicating

smaller variances at the between-group level) led to less accurate estimates across all approaches. Notably, the WFcovshrink approaches consistently yielded more accurate estimates compared to the unregularized approaches here. Furthermore, an interaction effect was evident between ICC and correlation at both levels, impacting the overall estimation accuracy of the between-group level parameters. Settings with a low ICC and small correlations at the between-group level resulted in the least accurate estimates across all approaches (approximately 900% in the unregularized approaches and approximately 400% in the WFcovshrink approaches). Notably, in these settings, larger correlations at the within-group level additionally decreased the accuracy of the between-group level parameters. Conversely, settings with a high ICC and large correlations at the between-group level yielded the most accurate estimates across all approaches (approximately 6% in the unregularized approaches, slightly less in the WFcovshrink approaches). Here, the correlations at the within-group level had no substantial influence. Similar to scenarios with small sample sizes, the WFcovshrink approaches proved most effective in addressing the more challenging settings. As indicated in the middle panel, which shows the accuracy of the within-group level parameters, we observed that smaller correlations at the within-group level led to less accurate estimates across all approaches. It appeared that the WFcovshrink approaches resulted in more accurate estimates at the within-group level when these correlations at the within-group level were small but not when they were large. Thus, once again, we observed that the WFcovshrink approach was most effective in handling the more problematic settings. In the lower panel, showing the estimates of the ICCs, we found that the ICC in population had the strongest influence. Small ICCs resulted in the least accurate estimates throughout all approaches, but the WFcovshrink approaches were the most effective here again.

5. Discussion

Small sample sizes, such as small group sizes (level-1 units) and small numbers of groups (level-2 units), often pose challenges to multilevel SEM, including difficulties in achieving convergence and inaccuracies in estimating between-group level parameters. To tackle these issues, our research investigated the effectiveness of a two-stage estimation approach, WFcovshrink, which replaces the sample covariance matrix by an estimate of the linear shrinkage estimator introduced by Touloumis (2015). Unlike the traditional unregularized long format (LF) and wide format (WF) approaches, the WFcovshrink methods consistently achieved convergence, regardless of the sample size or the ratio of columns to rows in the data matrix. In terms of accuracy, WFcovshrink outperformed the other approaches in estimating between-group level parameters across all sample sizes tested. Regarding within-group level accuracy, WFcovshrink proved superior only in scenarios with extremely small group sizes ($n = 2$), but even when the number of groups reached up to 100. Our approach also delivered more accurate ICC estimates by exhibiting a conservative downward bias compared to the typically overestimated ICCs found in unregularized methods in cases with small ICCs (0.05) and very small group sizes ($n = 2$). Given that in psychology and the education sciences, the ICCs commonly encountered are usually at the lower end (Adams et al., 2004; Gulliford et al., 1999), the conservative nature of WFcovshrink's estimates might be preferred. WFcovshrink showed its greatest efficacy in the most challenging conditions: small samples at any level, low ICCs, and minor correlations at the between- or within-group level. The performance of the three target matrices within WFcovshrink was largely similar. In sum, incorporating shrinkage estimation of the sample covariance matrix into a two-stage approach for multilevel SEM significantly mitigated the issues of non-convergence and inaccurate parameter estimates at the between-group level, and for very small group sizes, it effectively shrunk the issue of imprecise within-group level parameter estimates.

However, we must acknowledge that the proposed two-stage approach is only a partial success at this time. While it proves the concept, it remains limited in practical application for two main reasons. Firstly, settings with very small group sizes ($n = 2$) combined with small ICCs are relatively rare. These may appear in pilot studies, but few other research areas consider such settings. Secondly, and closely related to the first point, more customized target matrices need to be considered. The employed target matrices were designed for single-level data, not for single-level representations of multilevel data. Thus, the multilevel nature of the data was not adequately accounted for. Future research calls for more customized target matrices, as without these, the approach is rarely applicable in any realistic setting.

Further points that limit the generalizability of our findings need to be addressed. Firstly, in each simulation scenario, the variances of all observed variables at each level, and consequently the ICCs, were identical. This uniformity might have led to overly optimistic results when using the equal target matrix for shrinkage estimation. Closely related, the total variance of each observed variable was set to 1 in the population. Thus, the results may be limited to situations with variables having unit variances and future research is needed to investigate observed variables with other metrics. However, when practitioners have data with other than unit variance, two pragmatic ways to use our approach may be (1) to first standardize the variances and then use any target matrix or (2) to use the equal or unequal target (but not the identity) target matrices. Secondly, we limited the simulation study to balanced data (i.e., equal group size), but unbalanced data is often the case in practice. How our approach might be used with missing data deserves further attention. Missing values can be imputed ad hoc, for example, with multiple imputation techniques such as MICE (Buuren & Groothuis-Oudshoorn, 2011), and the resulting complete data matrix can then be used for regularization of the covariance matrix. Though, in small sample scenarios, imputation ought to be carefully considered as it might introduce bias (Grund et al., 2018). Another idea

would be to use the pairwise complete data to estimate the regularized covariance matrix. In any case, standard errors are likely to be incorrect because of the varying sample sizes for each (co)variance and we would have to account for this fact. Moreover, these are just ideas that need to be empirically studied. Till then, the application of our approach is limited to balanced data (e.g., experimental data). Thirdly, we only investigated scenarios with small group sizes n because of larger computational costs of the $p \cdot n$ "specific-units" variables in the WF approach. The model size and syntax grows with both p and n as well. To formulate the constraints in the WF approach more efficiently one could use Kronecker product constraints as suggested by Oort (2001, 2009). In this regard, however, using (proprietary) software with more advanced support for matrix algebra in SEM, such as openMX, is suggested. Fourthly, our investigation focused exclusively on two-level intercept-only models. It remains to be tested how regularized estimators of unstructured covariance matrices perform with more structured models. Arruda and Bentler (2017) also used a regularized estimator of unstructured covariance matrices, but, unlike our approach, applied it to the weight matrix in generalized least squares (GLS) estimation – which is commonly the sample covariance matrix. They found that their approach improved overall model evaluation (test statistics, rejection rates) in small samples compared to standard GLS and MLE for a common CFA in simulation studies, which includes three latent factors, each with five manifest variable indicators and unique error variances. Although these results suggest that the benefits of regularized estimators of unstructured covariance matrices, such as the one we employed by Touloumis (2015), could extend to more structured models, a thorough exploration of this possibility would be a valuable avenue for future research. Lastly, the accuracy of standard errors produced by WFcovshrink and the development of potential corrections warrant further investigation in order to to ensure the approach's broader reliability and applicability.

In conclusion, the application of shrinkage estimation to the covariance matrix within multilevel structural equation modeling (SEM) is a relatively new and evolving field. Our study stands out as one of the pioneering efforts to integrate this shrinkage estimation of covariance matrices in the SEM framework, and, to the best of our knowledge, it is the first to examine this method specifically in the multilevel modeling context. However, before the approach can be applied broadly in practice, more research needs to be done. Still, we believe this method merits consideration by the research community, offering a valuable tool for enhancing the accuracy and convergence of multilevel SEM analyses in small sample size scenarios.

Disclosure statement

The authors report there are no competing interests to declare.

ORCID

Julia-Kim Walther http://orcid.org/0000-0001-5758-1211
Martin Hecht http://orcid.org/0000-0002-5168-4911
Steffen Zitzmann http://orcid.org/0000-0002-7595-4736

References

Adams, G., Gulliford, M. C., Ukoumunne, O. C., Eldridge, S., Chinn, S., & Campbell, M. J. (2004). Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *Journal of Clinical Epidemiology*, *57*, 785–794. https://doi.org/10.1016/j.jclinepi.2003.12.013

Allison, P. D. (2009). *Fixed Effects Regression Models*. SAGE publications.

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, *49*, 155–173. https://doi.org/10/cwnzr3

Arruda, E. H., & Bentler, P. M. (2017). A regularized GLS for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*, 657–665. https://doi.org/10/gcmfh5

Barendse, M., & Rosseel, Y. (2020). Multilevel modeling in the 'wide format' approach with discrete data: A solution for small cluster sizes. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 696–721. https://doi.org/10.1080/10705511.2019.1689366

Bickel, P. J., Li, B., Tsybakov, A. B., Van De Geer, S. A., Yu, B., Valdés, T., Rivero, C., Fan, J., & Van Der Vaart, A. (2006). Regularization in statistics. *Test*, *15*, 271–344. https://doi.org/10.1007/BF02607055

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, *24*, 2350–2383. https://doi.org/10.1214/aos/1032181158

Buuren, S. v., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67. https://doi.org/10.18637/jss.v045.i03

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research*, *29*, 468–508. https://doi.org/10/cs43xwZSCC:0000626

De Jonckere, J., & Rosseel, Y. (2023). A model-based shrinkage target to avoid non-convergence in small sample SEM. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*, 941–955. https://doi.org/10.1080/10705511.2023.2171420

Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 327–351. https://doi.org/10.1080/10705511.2014.937849

Dragulescu, A., & Arendt, C. (2020). November 10). *Xlsx: Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003 Files* (Version 0.6.5). Retrieved March 3, 2023, from https://CRAN.R-project.org/package=xlsx

Duncan, T. E., Duncan, S. C., Alpert, A., Hops, H., Stoolmiller, M., & Muthen, B. (1997). Latent variable modeling of longitudinal and multilevel substance use data. *Multivariate Behavioral Research*, *32*, 275–318. https://doi.org/10.1207/s15327906mbr3203_3

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Gorsuch, R. L. (1983). *Factor analysis*. Lawrence Earlbaum Associates.

Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: simulations and recommendations. *Organizational Research Methods*, *21*, 111–149. https://doi.org/10.1177/1094428117703686

Gulliford, M. C., Ukoumunne, O. C., & Chinn, S. (1999). Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: Data from the health survey for England 1994. *American Journal of Epidemiology*, *149*, 876–883. https://doi.org/10/gn2gxn

Hamaker, E. L., Dolan, C. V., & Molenaar, P. C. M. (2003). ARMA-based SEM when the number of time points T exceeds the number of cases N: Raw data maximum likelihood. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*, 352–379.ZSCC: 0000049. https://doi.org/10/dkmnbp

Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small

samples. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 157–174. https://doi.org/10.1207/S15328007SEM0802_1
Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, *64*, 157–170. https://doi.org/10.1111/j.1467-9574.2009.00445.x
Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.
Huang, P.-H., Chen, H., & Weng, L.-J. (2017). A penalized likelihood method for structural equation modeling. *Psychometrika*, *82*, 329–354. https://doi.org/10/gbm65j
Huang, Y., & Bentler, P. M. (2015). Behavior of asymptotically distribution free test statistics in covariance versus correlation structure analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 489–503. https://doi.org/10/gcz6zp
Hugh-Jones, D. (2022). *Huxtable: Easily Create and Style Tables for LaTeX, HTML and Other Formats* (Version 5.5.6). Retrieved March 3, 2023, from https://CRAN.R-project.org/package=huxtable
Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: a Multidisciplinary Journal*, *23*, 555–566. https://doi.org/10/gcmfh8
Jung, S., & Takane, Y. (2007). Regularized common factor analysis. *New Trends in Psychometrics*, 141–149. Retrieved February 14, 2022, from https://www.semanticscholar.org/paper/Regularized-Common-Factor-Analysis-Jung-Takane/2df36e88cacc510b1f900960b6e784df886f0fbb
Kamada, A., & Kano, Y. (2012, July 1–4). *Statistical inference in structural equation modeling with a near singular covariance matrix* [Paper presentation]. 2nd Institute of Mathematical Statistics Asia Pacific Rim Meeting, Tsukuba, Japan.
Kamada, A., Yanagihara, H., Wakaki, H., & Fukui, K. (2014). Selecting a shrinkage parameter in structural equation modeling with a near singular covariance matrix by the GIC minimization method. *Hiroshima Mathematical Journal*, *44*, 315–326. https://doi.org/10/gpgn8b
Kelley, C. T. (1995). *Iterative methods for linear and nonlinear equations*. Society for Industrial and Applied Mathematics. Retrieved June 9, 2022, from https://epubs.siam.org/doi/book/10.1137/1.9781611970944
Lange, K., Chambers, J., & Eddy, W. (1999). *Numerical analysis for statisticians* (Vol. 2). Springer.
Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*, 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4
Ledoit, O., & Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, *40*. https://doi.org/10.1214/12-AOS989
Ledoit, O., & Wolf, M. (2020). The power of (non-)linear shrinking: A review and guide to covariance matrix estimation. https://doi.org/10.5167/UZH-170642
Liang, X., & Jacobucci, R. (2020). Regularized structural equation modeling to detect measurement bias: Evaluation of lasso, adaptive lasso, and elastic net. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 722–734. ZSCC: 0000023. https://doi.org/10.1080/10705511.2019.1693273
Lorenzo-Seva, U., & Ferrando, P. J. (2021). Not positive definite correlation matrices in exploratory item factor analysis: Causes, consequences and a proposed solution. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*, 138–147. https://doi.org/10.1080/10705511.2020.1735393
Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2×2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological Methods*, *16*, 444–467. https://doi.org/10.1037/a0024376
Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*, 203–229. https://doi.org/10/c8qdh2
McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, *28*, 295–314. https://doi.org/10.1007/s10648-014-9287-x
Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*, 259–284. https://doi.org/10.1037/1082-989X.10.3.259
Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methods*, *3*, 45–58.
Muthén, B. O. (1990). *Mean and covariance structure analysis of hierarchical data*. Department of Statistics, UCLA.
Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*, 376–398. https://doi.org/10.1177/0049124194022003006
Muthen, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 267. https://doi.org/10.2307/271070
Oort, F. J. (2001). Three-mode models for multivariate longitudinal data. *The British Journal of Mathematical and Statistical Psychology*, *54*, 49–78. https://doi.org/10.1348/000711001159429
Oort, F. J. (2009). Three-mode models for multitrait-multimethod data. *Methodology*, *5*, 78–87. https://doi.org/10.1027/1614-2241.5.3.78
Orzek, J. H., & Voelkle, M. C. (2023). Regularized continuous time structural equation models: A network perspective. *Psychological Methods*, *28*, 1286–1320. https://doi.org/10.1037/met0000550
Pedersen, T. L. (2022). *Patchwork: The composer of plots* (Version 1.1.2). Retrieved March 3, 2023, from https://CRAN.R-project.org/package=patchwork
R Core Team. (2023). *R: A language and environment for statistical computing*. https://www.R-project.org/
Rosseel, Y., Jorgensen, T. D., Rockwood, N., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., Barendse, M., Scharf, F., & Du, H. (2023). *Lavaan: Latent Variable Analysis* (Version 0.6-15). Retrieved March 3, 2023, from https://CRAN.R-project.org/package=lavaan
Shin, Y., & Raudenbush, S. W. (2010). A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics*, *35*, 26–53. https://doi.org/10/c63vdm
Signorell, A., Aho, K., Alfons, A., Anderegg, N., Aragon, T., Arachchige, C., Arppe, A., Baddeley, A., Barton, K., Bolker, B., Borchers, H. W., Caeiro, F., Champely, S., Chessel, D., Chhay, L., Cooper, N., Cummins, C., Dewey, M., Doran, H. C., & Zeileis, A. (2024). *DescTools: Tools for Descriptive Statistics* (Version 0.99.50). Retrieved March 21, 2024, from https://cran.r-project.org/web/packages/DescTools/index.html
Singer, H. (2010). SEM modeling with singular moment matrices Part I: ML-estimation of time series. *The Journal of Mathematical Sociology*, *34*, 301–320. https://doi.org/10.1080/0022250X.2010.509524
Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2. ed). SAGE.
Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and bayesian approaches. *American Journal of Political Science*, *57*, 748–761. https://doi.org/10.1111/ajps.12001
Stein, C. (1956). *Some problems in multivariate analysis, Part I*. Stanford Univ CA.
Tikhonov, A. N. (1943). On the stability of inverse problems. *Doklady Akademii Nauk SSSR*, *39*, 195–198.
Touloumis, A. (2015). Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional settings. *Computational Statistics & Data Analysis*, *83*, 251–261. https://doi.org/10.1016/j.csda.2014.10.018
Touloumis, A. (2019). *ShrinkCovMat: Shrinkage Covariance Matrix Estimators* (Version 1.4.0). Retrieved March 21, 2024, from https://cran.r-project.org/web/packages/ShrinkCovMat/index.html
Trendafilov, N. T., & Unkel, S. (2011). Exploratory factor analysis of data matrices with more variables than observations. *Journal of Computational and Graphical Statistics*, *20*, 874–891. https://doi.org/10.1198/jcgs.2011.09211
Ulitzsch, E., Lüdtke, O., & Robitzsch, A. (2023). Alleviating estimation problems in small sample structural equation modeling—A comparison of constrained maximum likelihood, Bayesian estimation, and fixed reliability approaches. *Psychological Methods*, *28*, 527–557. https://doi.org/10.1037/met0000435.supp

Unkel, S., & Trendafilov, N. T. (2010). Simultaneous parameter estimation in exploratory factor analysis: An expository review. *International Statistical Review*, *78*, 363–382. https://doi.org/10.1111/j.1751-5823.2010.00120.x

Van Montfort, K., Oud, J. H., & Voelkle, M. C. (2018). *Continuous time modeling in the behavioral and related sciences*. Springer.

Voelkle, M. C., Oud, J. H. L., von Oertzen, T., & Lindenberger, U. (2012). Maximum likelihood dynamic factor modeling for arbitrary *N* and *T* using SEM. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*, 329–350. https://doi.org/10.1080/10705511.2012.687656

Walther, J.-K., Hecht, M., Nagengast, B., & Zitzmann, S. (2024). To be long or to be wide: How data format influences convergence and estimation accuracy in multilevel structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *0*, 1–16. https://doi.org/10.1080/10705511.2024.2320050

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., & Dunnington, D, RStudio. (2023). *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics* (Version 3.4.2). Retrieved March 3, 2023, from https://CRAN.R-project.org/package=ggplot2

Wickham, H., François, R., Henry, L., Müller, K., Vaughan, D., & Posit, P. B. C. (2023). *Dplyr: A Grammar of Data Manipulation* (Version 1.1.2). Retrieved March 3, 2023, from https://CRAN.R-project.org/package=dplyr

Wickham, H., Girlich, M., & RStudio. (2022). *Tidyr: Tidy Messy Data* (Version 1.3.0). Retrieved June 14, 2022, from https://CRAN.R-project.org/package=tidyr

Wilke, C. O. (2020). *Cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'* (Version 1.1.1). Retrieved May 19, 2023, from https://cran.r-project.org/web/packages/cowplot/index.html

Williams, D. R., & Rodriguez, J. E. (2022). Why overfitting is not (usually) a problem in partial correlation networks. *Psychological Methods*, *27*, 822–840. https://doi.org/10/gqb2r6

Wothke, W. (1993). Nonpositive Definite Matrices in Structural Modeling. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 256–293). Sage Publications.

Yuan, K.-H., & Bentler, P. M. (2017). Improving the convergence rate and speed of Fisher-scoring algorithm: Ridge and anti-ridge methods in structural equation modeling. *Annals of the Institute of Statistical Mathematics*, *69*, 571–597. https://doi.org/10/gpgn83

Yuan, K.-H., & Chan, W. (2008). Structural equation modeling with near singular covariance matrices. *Computational Statistics & Data Analysis*, *52*, 4842–4858. https://doi.org/10.1016/j.csda.2008.03.030

Yuan, K.-H., Wu, R., & Bentler, P. M. (2011). Ridge structural equation modelling with correlation matrices for ordinal and continuous data: Ridge SEM with correlation matrices. *The British Journal of Mathematical and Statistical Psychology*, *64*, 107–133. https://doi.org/10/cwd74t

Zitzmann, S. (2018). A computationally more efficient and more accurate stepwise approach for correcting for sampling error and measurement error. *Multivariate Behavioral Research*, *53*, 612–632. https://doi.org/10/gpgn86

Zitzmann, S., Lüdtke, O., & Robitzsch, A. (2015). A Bayesian approach to more stable estimates of group-level effects in contextual studies. *Multivariate Behavioral Research*, *50*, 688–705. https://doi.org/10/gg5fg2

Zitzmann, S., Lüdtke, O., Robitzsch, A., & Hecht, M. (2021). On the performance of bayesian approaches in small samples: A comment on Smid, McNeish, Miocevic, and van de Schoot (2020). *Structural Equation Modeling: A Multidisciplinary Journal*, *28*, 40–50. ZSCC: 0000027. https://doi.org/10.1080/10705511.2020.1752216

Zitzmann, S., Lüdtke, O., Robitzsch, A., & Marsh, H. W. (2016). A Bayesian approach for estimating multilevel latent contextual models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 661–679. https://doi.org/10.1080/10705511.2016.1207179

Zitzmann, S., Wagner, W., Hecht, M., Helm, C., Fischer, C., Bardach, L., & Göllner, R. (2021). How many classes and students should ideally be sampled when assessing the role of classroom climate via student ratings on a limited budget? An optimal design perspective. *Educational Psychology Review*, *34*, 511–536. ZSCC: NoCitationData[s0]. https://doi.org/10.1007/s10648-021-09635-4

Zitzmann, S., Walther, J.-K., Hecht, M., & Nagengast, B. (2022). What is the maximum likelihood estimate when the initial solution to the optimization problem is inadmissible? The case of negatively estimated variances. *Psych*, *4*, 343–356. https://doi.org/10.3390/psych4030029

Appendices

Appendix A

R Code

```
### Example Code
# - for data, sample covariance matrix, and two-level random-intercept model in Figure 1 (with g = 4)
# - for shrinkage estimate of the sample covariance matrix, and two-level random-intercept model in Figure 2,
  and estimates in Table 1 (with g = 50)

# Note that the code only works for p = 2 and n = 2.
# If you want to examine other settings, check out the
# the code for the simulation study on Github:
# https://github.com/demianJK/WFcovshrink

## (0) preparation ##########################################################

# load required packages
library(lavaan) # for model estimation
library(tidyr) # for reformating LF to WF with pivot_wider()
library(ShrinkCovMat) # for shrinkage estimation
# (code runs in ShrinkCovMat 1.4.0 which is the latest on CRAN, Feb 21st 2024)

# set random number seed to obtain example data

set.seed(4395)
```

```
## (1) Population Characteristics ########################################

# We use the lavaan syntax to set the population models.
popModel_B <- "x1~~0.05*x1; x2~~0.05*x2; x1~~0.015*x2" # between level
popModel_W <- "x1~~0.95*x1; x2~~0.95*x2; x1~~0.285*x2" # within level
# means are zero by default

# We have two variables x1 and x2.
p <- 2
# The variances for both variables are the same at each level.
# The variance at the between level is 0.05.
# The variance at the within level is 0.95.
# Thus, the ICC = 0.25.
# The correlation of the two variables is the same at both levels (.3).
# The covariances differ.
# Transform the correlation formula to get the covariances.
# corr_x1x2 = cov_x1x2 / (sd_x1 * var_x2)
# | * (sd_x1 * sd_x2) and sd_x1 = sd_x2 thus | * var_x1
# corr_x1x2 * var_x1 = cov_x1x2

## (2) Sample Characteristics ############################################

g <- 50 # number of groups (you may change this)
n <- 2 # group size (balanced data)
N <- g * n # total sample size

# the data sampling is done in long format (LF)
sample_B <- simulateData(popModel_B, sample.nobs = g,
     model.type = "lavaan") # between level
sample_W <- simulateData(popModel_W, sample.nobs = N, # within level
     model.type = "lavaan")
groups <- rep(1:g, each = n) # group numbers ("j" in Figure 1)
LF_T <- sample_W # create data frame with the same dimensions
LF_T[,] <- 0 #. and clear all entries
  for (j in unique(groups)) { # merge the sampled data from both levels
   for (i in min(which(groups == j)):max(which(groups == j)))
   LF_T[i,] <- sample_W[i,] + sample_B[j,]
}
LF_T$persons <- rep(1:n, g) # unit numbers ("i" in Figure 1)
LF_T$groups <- as.factor(groups)
LF_T <- cbind(LF_T[,(p+1):(p+2)], LF_T[,1:p]) # rearrange columns
# LF-T is the total data matrix in long format (LF)..
round(LF_T[,3:(3+p-1)], 0) # note that we round for Figure 1 and 2
#.. and the total covariance matrix is estimated by the unbiased estimator (see Muthén, 1994)
Sigma_LF_T <- cov(LF_T[,3:4])
round(Sigma_LF_T, 2)

# Now we reformat to wide format (WF).
WF_T <- pivot_wider(LF_T, names_from = "persons", values_from = 3:4, names_sep = ".")
round(WF_T[,2:(2+(p*n)-1)], 0)
varnames <- colnames(WF_T[,2:(2+(p*n)-1)])

# shrinkage estimate S*_E with equal target Matrix vI_p
# note that unbiased S is employed

WF_T_trans <- t(WF_T[,-1]) # transpose because ShrinkCovMat(data, .) expects
# that rows correspond to variables and columns to observations

# estimate S*_E (note that the approach uses unbiased S_WF-T)
WFcovshrink_E <- ShrinkCovMat::shrinkcovmat.equal(data = WF_T_trans,
    centered = FALSE)

round(WFcovshrink_E$Sigmasample, 2) # unbiased S, round for Figures 1 and 2
round(WFcovshrink_E$Target, 2) # vI_p, round for Figure 2
round(WFcovshrink_E$Sigmahat, 2) # S*_E, round for Figure 2
round(WFcovshrink_E$lambdahat, 2) # lambda_E, round for Figure 2
# names of covariance matrix required for lavaan
colnames(WFcovshrink_E$Sigmahat) <- varnames
rownames(WFcovshrink_E$Sigmahat) <- varnames
```

```
## (3) estimate models ###################################################

model_WF <- paste0(# Level: 1 (unique factors)
      "x1.1~~Vx1_w*x1.1; x1.2~~Vx1_w*x1.2; x2.1~~Vx2_w*x2.1;
      x2.2~~Vx2_w*x2.2; x1.1~~Cx12_w*x2.1; x1.2~~Cx12_w*x2.2;",
      # these are the desired within variances and covariances
      # Vx1_w, Vx2_w, and Vx12_w are equality constraints
      # Level: 2 (common factors)
      "x1.1 ~ 0*1; x1.2 ~ 0*1; ; x2.1 ~ 0*1; x2.2 ~ 0*1;",
      # if level-2 variables are aggregates of level-1 variables,
      # intercepts at level-1 have to be fixed to 0
      "fx1=~1*x1.1 + 1*x1.2; fx2=~1*x2.1 + 1*x2.2;",
      # measurement model with factor loadings set to 1
      "fx1~~fx1; fx2~~fx2; fx1~~fx2;",
      # these are the desired between variances and covariances
      "fx1 ~ 1; fx2 ~ 1") # between means
fit_WF <- sem(model = model_WF,
      data = WF_T)
summary(fit_WF)

fit_WFcovshrink_E <- sem(model_WF,
      sample.cov = WFcovshrink_E$Sigmahat,
      sample.cov.rescale = FALSE,
      # rescale sample.cov with (g-1/g)?
      sample.nobs = g,
      sample.mean = colMeans(WF_T[,-1]))
summary(fit_WFcovshrink_E)

## (4) estimate ICCs #####################################################
# ICC in population: 0.05 (see "Population Characteristics")
# the ICCs are estimated by the parameters of the model-implied matrices

## WF
fit_WF@Fit@x[7]/(fit_WF@Fit@x[7]+fit_WF@Fit@x[1]) # x1
fit_WF@Fit@x[8]/(fit_WF@Fit@x[8]+fit_WF@Fit@x[3]) # x2

## WFcovshrink(E)
fit_WFcovshrink_E@Fit@x[7]/(fit_WFcovshrink_E@Fit@x[7]+fit_WFcovshrink_E@Fit@x[1]) # x1
fit_WFcovshrink_E@Fit@x[8]/(fit_WFcovshrink_E@Fit@x[8]+fit_WFcovshrink_E@Fit@x[3]) # x2
```

Appendix B

Supplemental Analysis

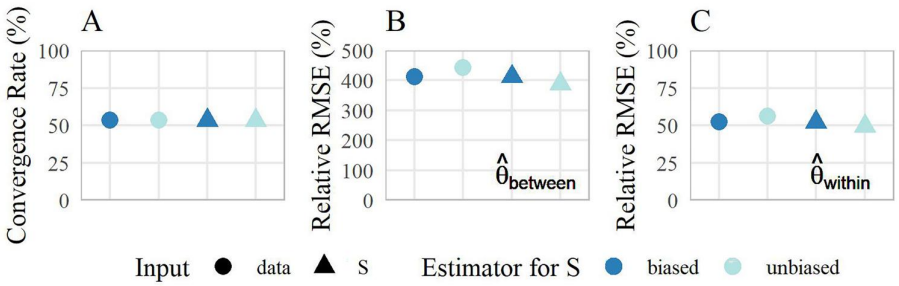


Figure B1. The WF approach and its different input and sample covariance matrix estimator possibilities.
Note. Data = input was data matrix; S = input was sample covariance matrix; biased = normal theory derived ML of sample covariance matrix used (default); unbiased sample covariance matrix was used ("sample.cov.rescale = TRUE").

We wanted to control for two differences in the WF and WFcovshrink approaches to check whether performance gains are solely attributable to the proposed two-stage approach. Firstly, we wanted to make sure that convergence gains in the WFcovshrink approach are not due to the input type, or more specifically, supplying a sample covariance matrix instead of a data matrix. The data matrix has to satisfy that the number of columns (number of observed variables) is smaller than the number of rows (number of observations), which is rooted in the implementation of traditional MLE in *lavaan* (see 1.1. The Wide Format (WF) Approach), and we wanted to check whether *lavaan* has similar constrains when supplying a sample covariance matrix (without sweeping the whole source code). Secondly, we wanted to ensure that accuracy gains in the WFcovshrink approach are not due to using different MLE of the sample covariance matrix; in particular, the unbiased one in WFcovshrink in contrast to the biased one that is the default in the WF approach. Figure B1 depicts the results of the conjugate analysis. With regard to convergence rate (Panel A), there were no differences. For overall estimation accuracy of between-group level (Panel B) and within-group level parameter estimates (Panel C), we found marginal differences. There seemed to have been some kind of interaction between the input and estimator. More specifically, the most accurate estimations were derived by supplying the sample covariance estimated by unbiased MLE. In contrast, supplying data and using the unbiased MLE as well yielded the least accurate estimations. However, overall these differences might be negligible. Thus, convergence and estimation accuracy gains can be mostly attributed to replacing the sample covariance matrix by a shrinkage estimate in the WFcovshrink approaches.

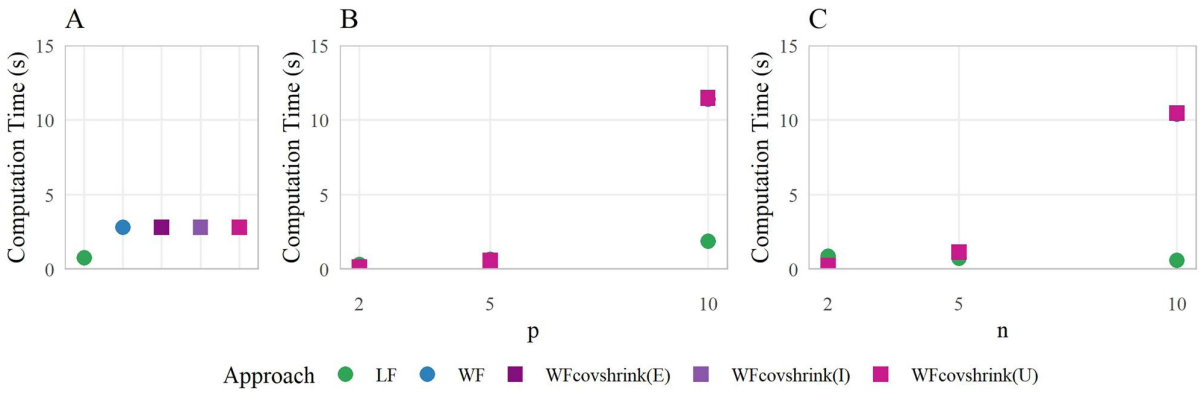


Figure B2. Computation time by sample characteristics.
Note. n = group size; p = number of observed variables; WFcovshrink(E) = equal target matrix; WFcovshrink(I) = identity target matrix; WFcovshrink(U) = unequal target matrix.

The computation time of a model is the time the optimizer needed to find a solution. In the WFcovshrink approaches, the time for the shrinkage estimation, which was marginally small for all three target matrices, was added. Figure B2 shows computation times for different aggregation levels. Panel A depicts the overall average computation times, which were smallest in the LF approach ($\approx 1s$), but not substantially different in the WF and WFcovshrink approaches ($\approx 3s$). In greater detail, Panel B and C depict that the computation time in all WF approaches magnified fairly by the number of observed variables p and the group size n. Recall that both quantities determined the number of model parameters that are freely estimated (p) and equality constrained (n). Again, there was no substantial difference in the WFcovshrink approaches compared to the unregularized WF approach. In the LF approach, the number of freely estimated in the LF approach was only determined by the number of observed variables p. Thus, the computation time of the LF approach was not influenced by the group size n, and its computation times were on average smaller. This could be explained by smaller dimensions of the covariance matrix in LF ($p \cdot p$) compared to the WF approaches ($(p \cdot n) \cdot (p \cdot n)$). Note that the population characteristics did not result in substantial differences in computation times. To put this finding into practical context: the larger computation times of any WF approach might be of little consequence if we only estimate a small number of models.

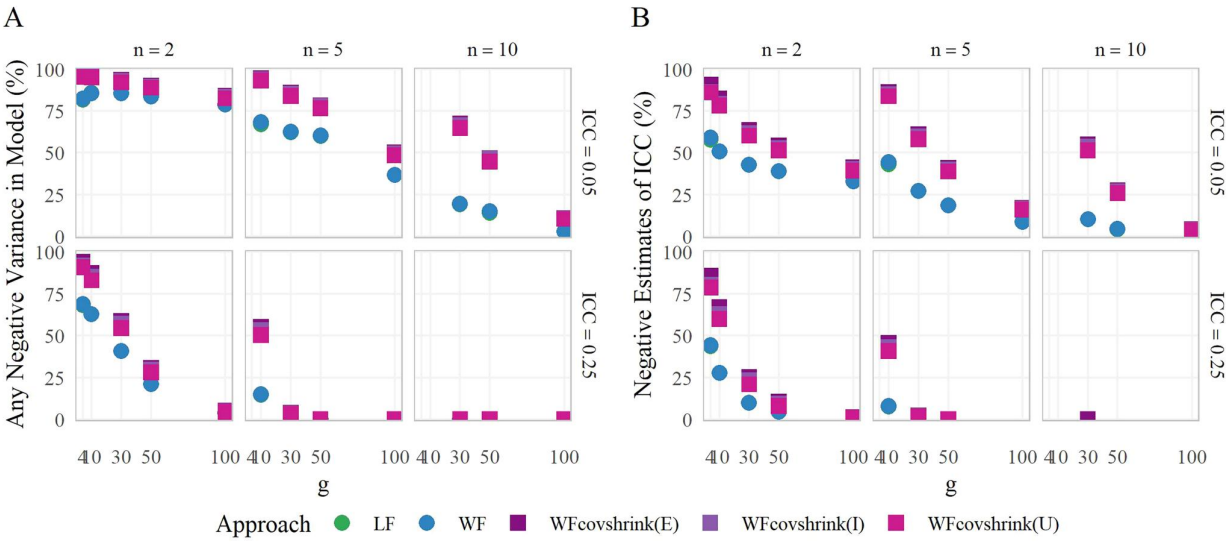
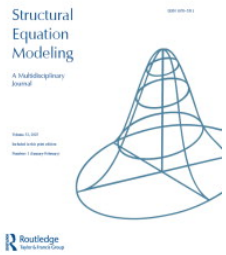


Figure B3. Negatively estimated variances at the between-group level and ICC.
Note. g = number of groups; n = group size; ICC = Intraclass Correlation; WFcovshrink(E) = equal target matrix; WFcovshrink(I) = identity target matrix; WFcovshrink(U) = unequal target matrix. In Panel A, percentages of negative variance in *any model* are shown. Note that negative variances were only present at the between-group level. In Panel B, percentages of negative estimates of the ICC of *any observed variable* are depicted. Thus, percentages depicted in Panel A are larger than those in Panel B.

Two types of inadmissibly negative estimates are depicted in Figure B3: between-group level variances and ICCs (i.e., quotient of between-group and total variances). Note that at the within-group level, no negative variances were encountered. The percentages of *at least one* negative variances at the between-group level in a model are shown in Panel A. Across all approaches, the percentage was larger when the number of groups g, the group size n, or the ICC was smaller. Overall, the WFcovshrink approaches yielded higher percentages of models with negative variances at the between-group level than the unregularized approaches. In Panel B, percentages of negatively estimated ICC for *every* observed variable in a model are shown. A similar picture emerged: percentages soared when the number of groups g, the group size n, or the ICC was smaller, and the percentages of the WFcovshrink approaches were higher. The increase in these negative estimates in the WFcovshrink approaches is probably related to the amplification of downward bias of between-group level estimates (see Figure 4).



Routledge
Taylor & Francis Group

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hsem20

Multilevel Multigroup Structural Equation Modeling In A Single-Level Framework

Julia-Kim Walther, Martin Hecht, Benjamin Nagengast & Steffen Zitzmann

To cite this article: Julia-Kim Walther, Martin Hecht, Benjamin Nagengast & Steffen Zitzmann (06 Jan 2025): Multilevel Multigroup Structural Equation Modeling In A Single-Level Framework, Structural Equation Modeling: A Multidisciplinary Journal, DOI: 10.1080/10705511.2024.2434596

To link to this article: https://doi.org/10.1080/10705511.2024.2434596

© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC.

Published online: 06 Jan 2025.

Submit your article to this journal

Article views: 175

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=hsem20

STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
https://doi.org/10.1080/10705511.2024.2434596

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Multilevel Multigroup Structural Equation Modeling In A Single-Level Framework

Julia-Kim Walther[a], Martin Hecht[b], Benjamin Nagengast[a,c], and Steffen Zitzmann[d]

[a]University of Tübingen; [b]Helmut Schmidt University; [c]Korea University; [d]Medical School Hamburg

ABSTRACT

Heterogeneity of variance is more than a statistical nuisance when variance parameters are of substantial interest. In multilevel modeling (e.g. students within classes), for instance, the inclusion of discrete variables at the between-cluster level (e.g. school type) may lead to the detection of differences between variances at the within-cluster level (e.g. students' performance in a test). The resulting heterogeneous variances (e.g. lower variance for students at high schools compared to grammar schools) have the potential to inform research and practice (e.g. on educational effectiveness). Along the lines of 'people are variables too', we demonstrate how the single-level formulation of multilevel structural equation models, the wide format approach (Barendse & Rosseel, 2020; Mehta & Neale, 2005), can be used in combination with multigroup modeling in order to obtain heterogeneous variance estimates. We provide evidence for the proposed WFmultigroup approaches' accuracy by means of a simulation study and showcase its application with an empirical illustration with the *lavaan* package in R.

KEYWORDS

Groups; heterogeneity; heteroscedasticity; multilevel; variance

Homogeneity of variances is a standard assumption in multilevel analysis. When disentangling within-cluster (e.g., student) and between-cluster (e.g., class) effects, it is assumed that within-cluster (residual) variances are equal across clusters, for instance, that variability of students' performance in a test is equal across classes. However, we may think of multiple scenarios where the homogeneity assumption is likely to be violated. For example, the variability of student's performance in a test might be contingent on the type of school they attend. The performance of students from high schools might be less variable than that of students from grammar schools. Indeed, empirical evidence suggests that heterogeneity of variance is a frequently observed phenomenon (Goldstein, 2005). Keselman et al. (1998) reviewed articles from prominent educational and behavioral science journals and reported a median variance ratio (VR) of 2.25. In other words, the group with the largest variance (e.g., grammar schools) showed variability more than twice the size of the group with the smallest variance (e.g., high schools). Nevertheless, a recent evaluation of reporting practice in multilevel research (Luo et al., 2021) showed that only 4.5% of studies checked the homogeneity assumption. The heterogeneity of variances appears to be less methodologically considered than empirically observed.

Whether heterogeneity of variances is considered a nuisance or an avenue depends on the research focus. Evidence suggests that unaccounted heterogeneity biases standard errors but not point estimates (Huang et al., 2023; Korendijk et al., 2008; Rosopa et al., 2019). Thus, if one is merely interested in means (e.g., of heterogeneous variances), then the standard post-hoc procedure is to correct the standard errors. This can be done, for example, by using robust standard errors (see Maas & Hox, 2004), resampling techniques (e.g., Zitzmann et al., 2023; see also Zitzmann et al., 2024), or by applying a non-linear transformation to the dependent variable (e.g., Hodges, 1998). If one is planning a study where one expects variances to be heterogeneous, calculating adequate sample sizes for the heterogeneous populations a priori is suggested (Candel & van Breukelen, 2015).

On the other hand, heterogeneous variance components might be of substantive interest. Analysing heterogeneous within-cluster (co)variances in students' performance can reveal differences in teaching effectiveness or curriculum impact within schools. These differences in variability might offer a valuable increment to mean tendencies alone (i.e., the mean performance of students from high schools and grammar schools). For instance, Raudenbush and Bryk (1987) found that catholic schools had somewhat smaller variability than public schools in math achievement. This finding may help limit potential variables that give rise to differential variances in math achievement by exploring in which variables the two school types differ. To quantify the heterogeneous within-cluster variances within the within-between variance decomposition that takes place in multilevel modeling in common statistical software, for instance,

CONTACT Julia-Kim Walther julia-kim.walther@gmx.de Hector Research Institute of Education Sciences and Psychology, University of Tübingen, 72072 Tübingen, Germany.

© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

Mplus (L. Muthén & Muthén, 1998–2017) and *lavaan* (Rosseel, 2012), Hedeker and Mermelstein (2007) and West et al. (2022) suggested to calculate group-specific Intraclass Correlations (ICCs are defined as the proportion of between-cluster variance out of the sum of the between- and within-cluster variances, i.e., the total variance; Hox et al., 2017), for instance, one ICC for high schools and one for grammar schools. In *Mplus*, for instance, these are given in the summary of the data. These may facilitate to decide whether certain between-cluster variables (e.g., school type) are relevant for the variability of a given outcome (e.g., students' test performance) or not.

To model heterogeneous variances, advanced statistical techniques have to be employed. Broadly speaking, there are two main frameworks that are suited to model heterogeneous variances for multilevel data: hierarchical models with heterogeneous variances and multilevel multigroup SEM. Hierarchical models with heterogeneous variances (also known as HET or dispersion models; e.g., Raudenbush & Bryk, 1987) are prominent in longitudinal research where inter-individual differences in intra-individual change is the subject of investigation. They are available in the *nlme* package in R. However, their main disadvantage is that one can neither model more than one dependent variable simultaneously nor measurement error. Multilevel multigroup SEMs (ML MG SEM; e.g., B. Muthén, 1997), however, are able to do so. Generally, multigroup models are frequently employed to test for measurement invariance in confirmatory factor analysis (CFA) across groups (e.g., school type, countries, measurement occasions), which is a prerequisite for cross-group comparisons such as group mean differences. When the data is hierarchical (e.g., schools in different countries, classes on multiple measurement occasions in a cohort study), then ML MG SEM allows to account for both the multigroup and multilevel nature. While these modeling approaches are available in common statistical software, we demonstrate along the lines of 'people are variables too' how they can be estimated in a single-level framework using the wide format approach (Barendse & Rosseel, 2020; Mehta & Neale, 2005; Walther, Hecht, Nagengast, et al., 2024). First, one needs to reformulate the multilevel SEM as single-level restricted confirmatory factor analysis (CFA) in the wide format (WF) approach. Then, one applies the multigroup feature to estimate group-specific (within-cluster) variances.

The present article has two objectives. Firstly, we will introduce our proposed WFmultigroup approach, which develops the notion of multilevel multigroup SEM as a single-level restricted CFA for multiple groups, and illustrates how to implement it in the *lavaan* package in R. Secondly, we will make the point that multilevel multigroup SEMs, which are usually used for testing for measurement invariance across groups, can also be used to model heterogeneous within-cluster (co)variances of manifest variables that are stratified by discrete between-cluster variables. The proposed WFmultigroup approach is supported by a simulation study and its application is demonstrated through an empirical example. The restrictions and limitations of the method will be addressed in the discussion.

1. The WFmultigroup Approach

1.1. Background

By the beginning of the century, hierarchical modeling and structural equation modeling, which have been thought of as two non-overlapping traditions for a considerable time, have been shown to be equivalent (e.g., Bauer, 2003; Rovine & Molenaar, 2000). Subsequently, Barendse and Rosseel (2020) and Mehta and Neale (2005) demonstrated that a multilevel structural equation can be fit by means of a single-level measurement model (CFA). A crucial feature of this reformulation is the data format. In the standard multilevel SEM, the data matrix is used in long format (LF), whereas in the single-level approach, the wide format (WF) data matrix is subjected. These LF and WF approaches to multilevel SEM have been shown to be empirically equivalent under various conditions in terms of estimation accuracy (Barendse & Rosseel, 2020; Mehta & Neale, 2005; Walther, Hecht, Nagengast, et al., 2024).

We were motivated by similar considerations about equality: when a multilevel SEM can be estimated as a single-level CFA, then a multilevel multigroup SEM may be estimated as a single-level multigroup CFA. Therefore, we suggest extending the WF approach by multigroup modeling and altering the model specification to allow for group-specific variances. In the remainder of this article, we will illustrate how a model with heterogeneous within-cluster (co)variances stratified by a between-cluster predictor can be fitted. However, models with different assumptions on heterogeneity at both levels as stratified by a between-cluster variable can be estimated with the proposed approach as well (see the complete code of the empirical illustration in Appendix B).

1.2. How It Works

Figure 1 illustrates the differences of the standard LF, the WF, and the proposed WFmultigroup approach to multilevel SEM. The depicted minimal example data set consists of ten clusters ($g = 10$) with two units in each cluster ($n = 2$). For every unit we observe two continuous variables ($p = 2$), x_1 and x_2, which are aggregated in order to obtain between-cluster variables. There is one further discrete between-cluster variable with two levels ($k = 2$) that serves as the grouping variable.

In Panel A, it can be seen that the WF approaches, in contrast to the standard LF approach, split the p observed variables into $p \cdot n$ variables in the data frame ("people are variables too", Mehta & Neale, 2005, p. 1). For instance, $x_{1.1}$ is the observed variable x_1 for every 1st unit in the cluster ($i = 1$). Thus, rows in the WF data matrix correspond to the numbers of clusters ($g = 10$; level-2 units) whereas in the LF data matrix, they correspond to the total number of units in all clusters ($g \cdot n = N = 20$; level-1 units).

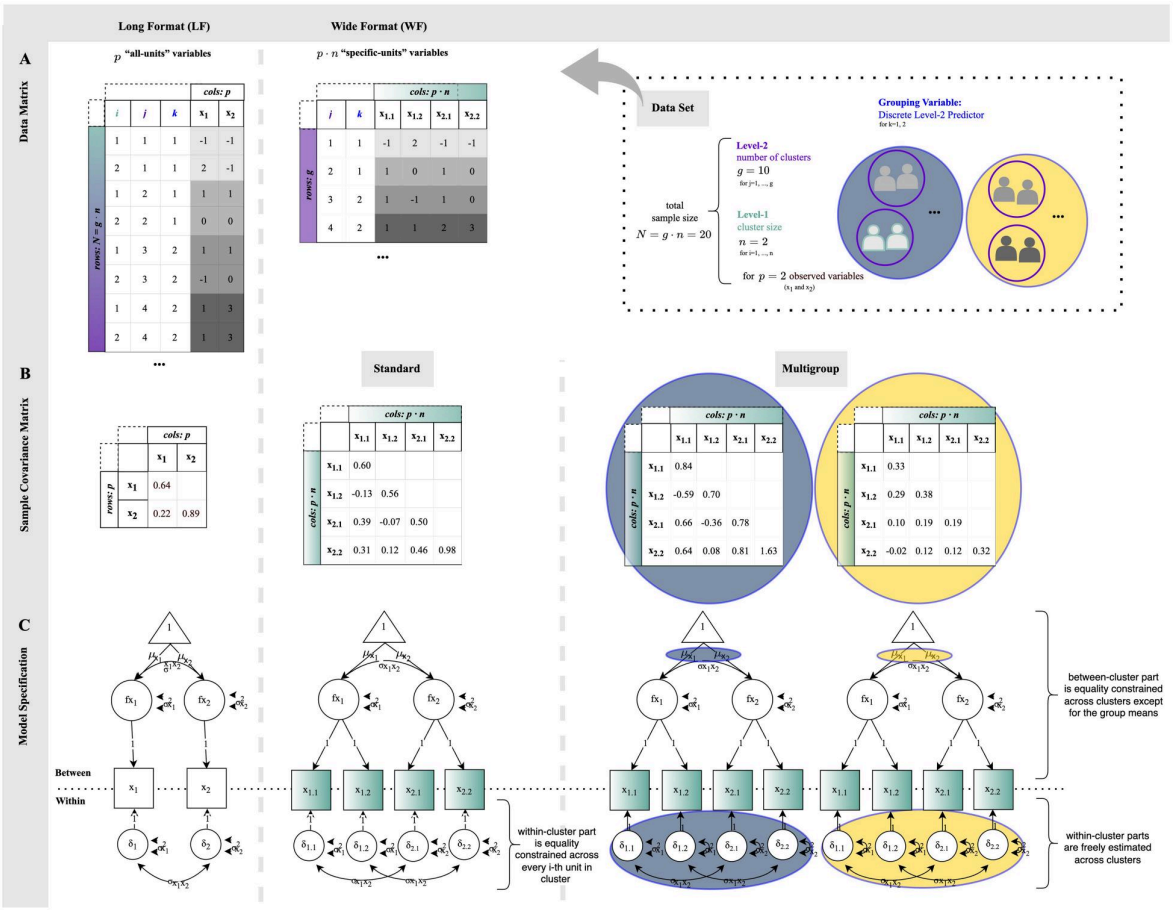


Figure 1. The LF, WF, and WF multigroup approaches. *Data set*: the data collected in a given setting. *Data Matrix*: the data set in matrix form, where columns refer to observed variables and rows to observed units. *Data Format*: one of two possible formats of the data matrix, long format (LF) or wide format (WF). In WF, every observed variable p is split for every unit in the cluster (n). For instance, $x_{1.1}$ is x_1 for every first unit in each cluster. *Sample Covariance Matrix*: a symmetric matrix that contains (co)variances of the observed variables. *Model Specification*: representation of the model to be estimated, here, this is a bivariate two-level intercept-only model. Between-cluster parameter estimates are located above the dashed line; within-cluster parameter estimates are located below. At each level, identical parameter estimates indicate equality constrains. The example data set has $g = 10$ clusters á $n = 2$ units, and $p = 2$ observed variables. Note that only the first four clusters are depicted. The R code to generate the data and models is available on Github (https://github.com/demianJK/WFmultigroup). The figure is adapted from "Shrinking Small Sample Problems in Multilevel Structural Equation Modeling via Regularization of the Sample Covariance Matrix" by J.-K. Walther, M. Hecht, and S. Zitzmann, 2024, *Structural Equation Modeling Journal*, 1–20. https://doi.org/10.1080/10705511.2024.2380919.

From the data matrices, the respective sample covariance matrices are estimated (see Panel B). Their dimensions are obtained from the number of respective "observed" variables (i.e., columns of the data matrix): $p \times p$ in the LF approach, and $(p \cdot n) \times (p \cdot n)$ in the WF approaches. The standard WF approach has one sample covariance matrix, whereas the WFmultigroup approach has two (i.e., one per group). Hence, the sample size for each sample covariance matrix depends on the number of clusters and cluster sizes in each group. In our example data set, there are balanced numbers of clusters and cluster sizes. Thus, each matrix is estimated by five clusters with two units each ($g = 5$ and $n = 2$) whereas the one WF sample covariance matrix is estimated by the full ten clusters with two units each ($g = 10$ and $n = 2$).

Regarding the model specification in Panel C, the WF approaches in contrast to the standard LF approach set equality constrains across the n splits of each observed variable p. Therewith, the within-cluster (co)variances of all i units within a cluster are set to be homogeneous. The WFmultigroup approach relaxes these equality constrains by applying constrains only for *each of the k groups*. Thereby, within-cluster (co)variances of all i units within a cluster are set to be homogeneous for each observed variable only per group. Put differently, within-cluster (co)variances are heterogeneous by group. The between-cluster means, which are modelled as latent factor intercepts, are also allowed to differ by group. In contrast, between-cluster (co)variances are set to be equal across groups, because we only assume the within-cluster (co)variances to be heterogeneous (though we could model the between-cluster (co)variances to be heterogeneous as well with this approach). Thus, one simply fits a multilevel SEM for each group with certain equality constrains across groups, which can be conceived as a multilevel multigroup SEM.

1.3. Sample Size Requirements

Whilst the WFmultigroup approach offers multiple possibilities for estimating parameters constrained and freely across

groups and levels, it has one noteworthy limitation due to its data format, which concerns sample size and convergence. The way the traditional maximum likelihood estimator (MLE) is implemented in *lavaan* requires a positive definite sample covariance matrix (Hamaker et al., 2003; Singer, 2010; Van Montfort et al., 2018; Voelkle et al., 2012; Walther, Hecht, Nagengast, et al., 2024), which, amongst others, necessitates that the supplied data matrix has just as many or less columns than rows. In the standard WF approach, $cols \leq rows$ translates to $(p \cdot n) \leq g$ (Walther, Hecht, Nagengast, et al., 2024). However, as multiple sample covariance matrices are estimated in the WFmultigroup approach (i.e., one per group), $(p \cdot n_k) \leq g_k$ has to hold for each group. When the number of clusters and cluster sizes differ substantially across groups, traditional MLE, which is based on the sample covariance matrix, might not be able to fit the model. However, one might use full information maximum likelihood (FIML) estimation, which uses the raw data instead and, hence, circumvents the problem (Hamaker et al., 2003; Trendafilov & Unkel, 2011; Unkel & Trendafilov, 2010; Voelkle et al., 2012). However, when the amount of missing data is too large, estimation might fail as well. One way to deal with both problems is multiple imputation, which we apply in the empirical example. However, before that, we will describe results from a small simulation study (without missing values) in which the performance of the proposed WFmultigroup approach was examined.

2. Simulation Study

We conducted a simulation study to investigate whether the proposed WFmultigroup approach is accurate and unbiased in estimating heterogeneous within-cluster (co)variance structures which are grouped by discrete between-cluster variables. Empirical equivalence of WFmultigroup with the "genuine" ML MG SEM for all homogeneous, heterogeneous between-cluster (co)variances and heterogeneous within- and between-cluster (co)variances models is demonstrated in the complete code for the empirical illustration in Appendix B.

2.1. Method

The computations were conducted on an AMD Ryzen Threadripper PRO 3975WX 32-cores (3.50 GHz) CPU on a Windows 10 (Version 20H2) platform utilising R version 4.4.0 (R Core Team, 2024), along with several R packages: *DescTools* version 0.99.50 (Signorell et al., 2024), *dplyr* version 1.1.4 (Wickham et al., 2023), *ggplot2* version 3.5.1 (Wickham, Chang, et al., 2024), *lavaan* version 0.6-17 (Rosseel et al., 2024), *patchwork* version 1.2.0 (Pedersen, 2024), *tidyr* version 1.3.1 (Wickham, Vaughan et al., 2024). The R code for data generation, analysis, and figures is available at https://github.com/demianJK/WFmultigroup.

2.1.1. Data Generation

We varied the number of clusters ($g = 200, 500, 1000$), the cluster size ($n = 2, 10, 30$), the variance ratio ($VR = 2, 5$), and the variance at the between-cluster level ($\sigma^2_B = 0.05, 0.25$). This resulted in $2 \times 2 \times 3 \times 3 = 36$ simulation conditions overall. The number of observed variables was fixed to $p = 2$, and two groups, as indicated by a discrete between-cluster variable ($k = 2$), were considered. The magnitudes of the between-cluster variances were informed by the lower and upper limits of frequently observed ICCs in the educational and behavioral sciences (Adams et al., 2004; Gulliford et al., 1999). In the first group, the total variance was set to 1, and the within-cluster variance was computed by $\sigma^2_{W1} = 1 - \sigma^2_B$ (and thus, $\sigma^2_B = ICC_1$). The within-cluster variance in the second group was computed by dividing through the VR. Note that the between-cluster (co)variances were equal across both groups as we only assumed the within-cluster (co)variances to be heterogeneous. The covariances at each level were determined by multiplying the variance with the fixed correlation of $\rho = 0.3$ which reflects a large correlation (Gignac & Szodorai, 2016).

2.1.2 Data Analysis

We considered only one model, a bivariate two-level intercept-only model with heterogeneous within-cluster (co)variances, which we estimated as a multigroup single-level CFA with *lavaan*. As Hedeker and Mermelstein (2007) and West et al. (2022) suggested, we computed group-specific ICCs by $ICC_1 = \sigma^2_B/(\sigma^2_B + \sigma^2_{W1})$ and $ICC_2 = \sigma^2_B/(\sigma^2_B + \sigma^2_{W2})$ for each variable.

2.1.3. Evaluation Criteria

We thoroughly investigated the estimation accuracy of the (co)variance structure in terms of the relative root mean squared error (RMSE), $\sqrt{\sum(\hat{\theta} - \theta)^2}/\theta \cdot 100\%$, which is a measure that combines both bias and variance of an estimator, and the relative bias, $\sum(\hat{\theta} - \theta)/\theta \cdot 100\%$. Convergence and coverage rates were also reported briefly. A model was considered converged if the optimizer indicated that it had found a solution. Convergence rates represent the percentage of converged models out of all estimated models. Coverage rates indicate the percentage of confidence intervals that encompass the population parameter. Note that for estimation accuracy and coverage rates, we considered only (co)variances (but not means) of the intercept-only models.

2.2. Results

Under every simulation condition, all models converged. Moreover, all coverage rates fell between the acceptable range of 91% to 98% (L. K. Muthén & Muthén, 2002). The more interesting results for relative RMSE and bias are depicted in Figure 2.

At the between-cluster level, previous findings could be replicated: smaller numbers of clusters, smaller cluster sizes, and smaller between-cluster variances (and thus, smaller ICCs as well) were detrimental for overall accuracy (see also Lüdtke et al., 2011; Meuleman & Billiet, 2009;

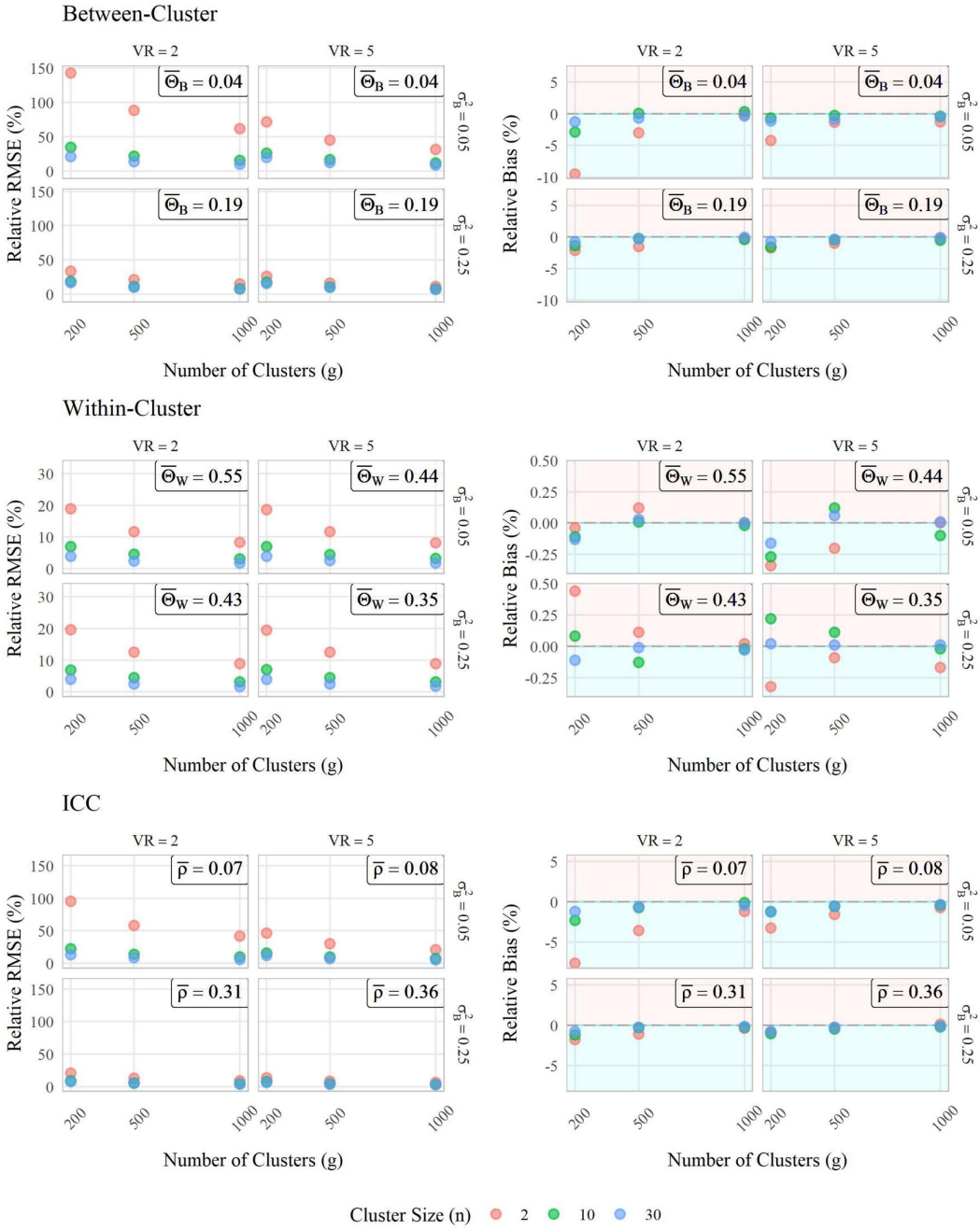


Figure 2. Estimation accuracy of between-cluster, within-cluster, and ICC parameter estimates. VR = variance ratio. Only (co)variance parameter estimates are considered. In a bivariate two-level intercept-only model with heterogeneous within-cluster (co)variances for two groups, this comprises three parameter estimates at the between-cluster level (i.e. two variances and one covariance, $\bar{\Theta}_B$), six parameter estimates at the within-cluster level (i.e. two variances and one covariance for both groups, $\bar{\Theta}_W$), and four ICC parameter estimates (i.e. one for each group per variable, $\bar{\rho}$).

Stegmueller, 2013; Walther, Hecht, Nagengast, et al., 2024; Zitzmann, 2018; Zitzmann et al., 2016). Combined, these lead to a relative RMSE of up to 150%, even when the minimum number of clusters was moderately large ($g = 200$). Increasing the cluster size moderately (from $n = 2$ to $n = 10$) reduced the relative RMSE by up to 40%. Smaller cluster sizes and smaller between-cluster variances were associated with larger negative biases. However, all sample sizes resulted in biases within the acceptable limit of |10%| (L. K. Muthén & Muthén, 2002). It is interesting to note that larger VRs led to more accurate and less biased between-cluster parameter estimates, especially when the cluster size was small. Drawing on the earlier example setting, when $g = 200$ and $n = 2$, when $VR = 2$, the relative RMSE was 150%, whereas when $VR = 5$, it dropped to half. We hypothesize that this might be related to the factor analytic modeling: In the single-level multigroup CFA framework, the between-cluster (co)variances are estimated as a common factor (co)variances that are equality constrained across groups. When

the VR was larger, the ratio of common to unique variance of the indicators (i.e., the $p \cdot n$ "observed" variables), which might be though of as their ICCs (common as between-cluster and unique as within-cluster variances), got larger by design in the second group. Thus, the amount of communality of the indicators across both groups increased. Especially when the number of indicators was small (i.e., small cluster sizes n), a larger VR could have compensated for its negative effect. This argumentation is in line with evidence suggesting that smaller common factor variances (i.e., commonalities) are more strongly influenced by sample size when it comes to factor recovery (MacCallum et al., 1999).

At the within-cluster level, smaller numbers of clusters and smaller cluster sizes were related to less accurate estimates as well, but the relative RMSE was only up to 20% at worst. Bias was close to zero. This replicates earlier findings suggesting that parameter estimates of between-cluster variables are less accurate and more biased than those of within-cluster parameter estimates (e.g., Depaoli & Clifton, 2015; Finch & French, 2011; Hox & Maas, 2001; Hox et al., 2010; Lüdtke et al., 2011; Muthen & Satorra, 1995; Zitzmann et al., 2016). There was no effect of the VR on the accuracy of the within-cluster parameter estimates.

The ICC estimates, as derived from the between- and within-cluster variance estimates, inherited both their strengths and weaknesses: smaller numbers of clusters, smaller cluster sizes, smaller between-cluster variances, and smaller VRs led to less accurate and more negatively biased estimates (as the between-cluster parameter estimates) but the magnitude of inaccuracy and bias was less strong (as for the within-cluster parameter estimates).

Overall, the proposed WFmultigroup approach lead to accurate and almost unbiased estimates and converging models with accurate standard errors. We recommend using at least a moderate number of clusters and cluster sizes to guarantee good accuracy and unbiasedness. In the case of a bivariate intercept-only model with two groups with balanced numbers of clusters and cluster sizes, a sample of $g = 200$ and $n = 10$, or more precisely, $g = 100$ and $n = 10$ for every group, satisfies this requirement.

3. An Empirical Illustration

In the following, we will work through a step-by-step guide on how to estimate a multilevel multigroup SEM as a single-level restricted multigroup CFA in *lavaan* using an empirical illustration. Specifically, we will investigate the heterogeneity of (co)variances of two observed variables, creative activities at school and growth mindset, in Albania and Ireland (i.e., the between-cluster variable is country). The analysis of their (co)variance structures can inform us about differences in the countries which one could subsequently explore to gain insight into variables that influence the variability of these outcomes. We will fit a model which assumes heterogeneity of within-cluster (co)variances (and homogeneity of between-cluster (co)variances) across groups in the single-level multigroup framework (WFmultigroup). In the main body of this article, only the code for the model specification is presented. The code for all other prior steps, such as data subsetting, inspection of missing data, and multiple imputation, as well as model specifications of models with homogeneous within- and between-cluster (co)variances, heterogeneous between-cluster (co)variances, and heterogeneous within- and between-cluster (co)variances with the WFmultigroup approach and the "genuine" ML MG SEM approach in *lavaan* can be found in the complete code in Appendix B. We draw on an open access data set of the Programme for International Assessment of Student Assessment (PISA) from 2022 which can be downloaded from https://www.oecd.org/pisa/data/2022database/. Note that the data set and variables were chosen by convenience to provide readers with a reproducible example and illustrate the WFmultigroup approach and thus, the investigated research question is not of substantive interest.

All computations of the empirical illustration were run on a Macbook Pro (2021) with an M1 Pro CPU on the Sonoma 14.5 platform utilising R version 4.4.0 (R Core Team, 2024) with the following packages: *dplyr* version 1.1.4 (Wickham et al., 2023), *foreign* version 0.8-87 (R Core Team et al., 2024), *ggplot2* version 3.5.1 (Wickham, Chang, et al., 2024), *huxtable* version 5.5.6 (Hugh-Jones, 2022), *lavaan* version 0.6-18 (Rosseel et al., 2024), *lme4* version 1.1-35.5 (Bates et al., 2024), *MICE* version 3.16.0 (Buuren et al., 2023), *naniar* version 1.1.0 (Tierney et al., 2024), *patchwork* version 1.2.0 (Pedersen, 2024), *psych* version 2.4.6.26 (Revelle, 2024), and *tidyr* version 1.3.1 (Wickham, Vaughan et al., 2024).

3.1. Data Set

3.1.1. The Sample

The complete PISA data set was collected within a stratified two-stage sampling process. Firstly, schools in which 15-year-old students (i.e., the target level-1 units) may be enrolled, were sampled. The minimum number of schools (i.e., level-2 units; clusters) for each country were 150. Secondly, students within these schools were sampled. The two observed variables that we consider are not part of the PISA test but the background information.

For our empirical illustration, we selected two countries from the pool of included countries: Albania and Ireland. The choice fell on them because both variables had a large VR in these countries and where thus well suited for the kind of analysis we want to illustrate. The total subsample consists of $g = 444$ schools with a total of $N = 11,698$ students. The number of schools (i.e., clusters) and students in each school (i.e., cluster sizes) for both countries are depicted in Figure 3. As can be seen in panel A, 274 schools are from Albania and 170 schools from Ireland, with a total of $N_{Albania} = 5,569$ and $N_{Ireland} = 6,129$ students. Unfortunately, however, the school sizes differ substantially from $n_{\min} = 1$ to $n_{\max} = 45$ with stark differences across countries (see Panel B). This will introduce a considerable amount of missing values later on when reformatting LF to

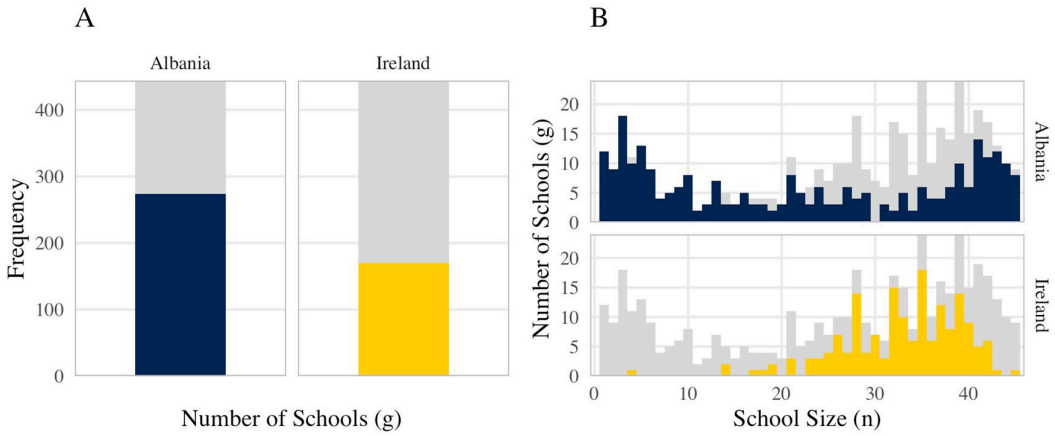


Figure 3. Number of schools and school sizes by country. Number of Schools = Clusters (i.e. Level-2 units); school size = cluster size (i.e. level-1 units students).

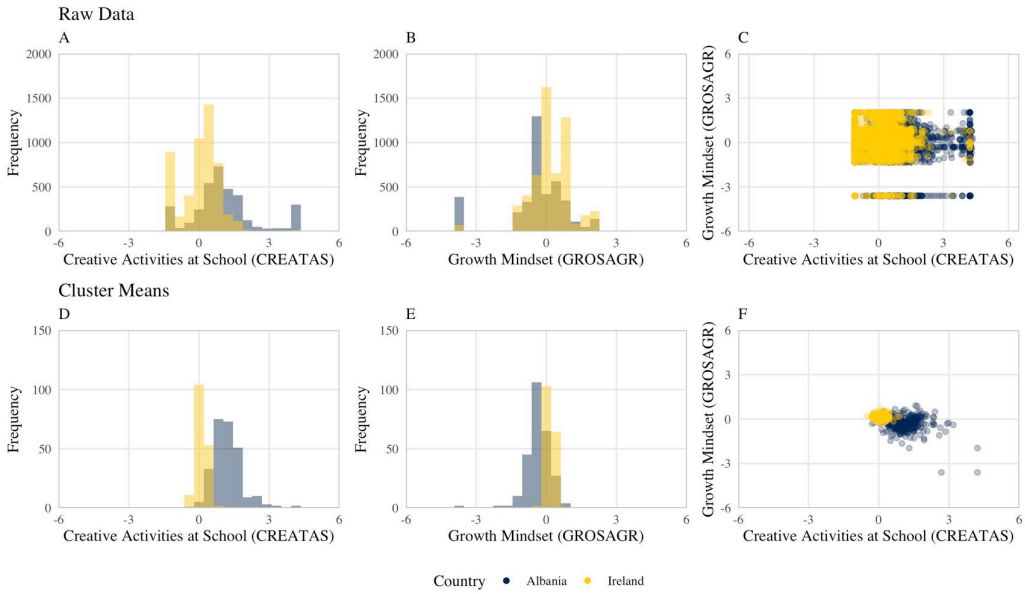


Figure 4. The distributions of raw data and cluster means. $N_{CREATAS(All)} = 8,449$ (28% missings) with $N_{CREATAS(Albania)} = 3,398$ (23.5% of all missings and 44.5% of missings in Albania) and $N_{CREATAS(Ireland)} = 5,051$ (4.5% of overall missings and 10% of missings in Ireland); $N_{GROSAGR(All)} = 9,319$ (20% missings) with $N_{GROSAGR(Albania)} = 3,870$ (19% of all missings and 58% of missings in Albania) and $N_{GROSAGR(Ireland)} = 5,449$ (1% of all missings and 2% of missings in Ireland); numbers refer to the LF data matrix with unbalanced cluster sizes (see Figure 3).

WF, where balanced cluster sizes are required, and thus, columns change from p to $p \cdot n_{\max}$.

3.1.2. The Observed Variables

The two variables that we included in our analysis are creative activities at school (CREATAS) and growth mindset (GROSAGR). According to the codebook and the plotted data (see Figure 4), they are continuous, and even if their distributions deviate from normality, see Panel A and B, the large sample sizes should warrant inferential conclusions, even in the presence of relatively large amounts of missing data (28% and 20%).

By plotting the raw data (Panel A to C) and the cluster means (Panel D to F) per group, one gets valuable information on potential heterogeneity of (co)variances. In Panel A and B, the univariate distributions of creative activities at school and growth mindset are depicted. The variability of each variable differs group-wise. The same holds true for the coherence of both variables in Panel C. This suggests that (at least) the within-cluster (co)variances are heterogeneous. When inspecting the distributions of the cluster means, the univariate distributions in Panel D and E and the bivariate distribution in Panel F, one sees that they differ group-wise as well. Taken together, this suggests that both the within- and the between-cluster (co)variances are heterogeneous. We simulated data under differing homogeneous and heterogeneous conditions at both levels and examined the variability of raw data and cluster means to support this claim (see Figure A1 in Appendix A). When both the within- and between-cluster levels in both groups were from different populations, then a pattern of group-wise differing raw data and cluster means appeared. Note, nevertheless, that in the main body of the article, only the

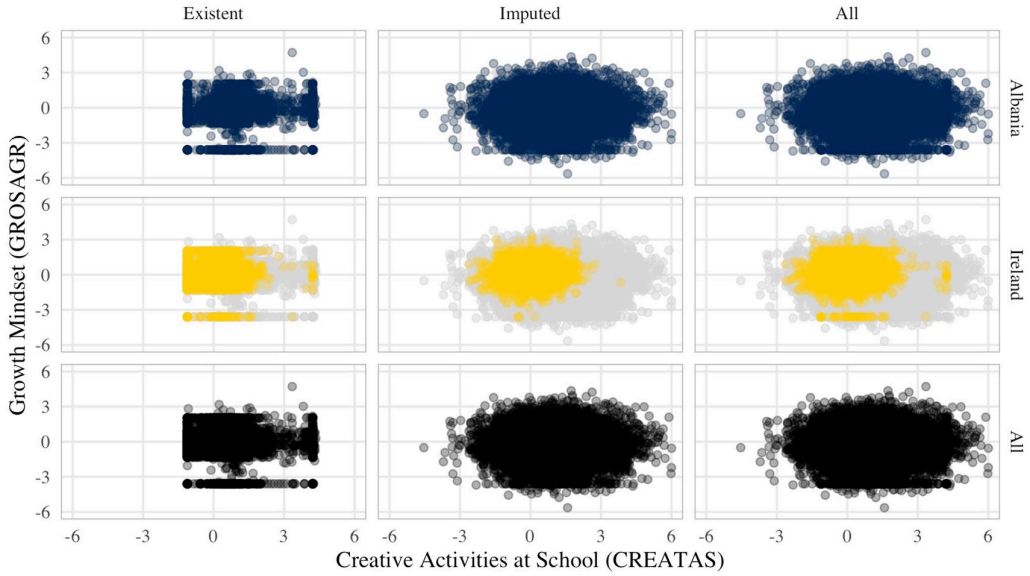


Figure 5. Existent and imputed data. "All (data)" refers to $N = 19,980$ for each variable where $N_{Albania} = 12,330$ and $N_{Ireland} = 7,650$; "Existent (data)": $N_{CREATAS(Albania)} = 3,398$ and $N_{CREATAS(Ireland)} = 5,051$, $N_{GROSAGR(Albania)} = 3,870$ and $N_{GROSAGR(Ireland)} = 5,449$ (but only complete case-wise existent cases are depicted); "Imputed (data)": $N_{CREATAS(Albania)} = 8,932$ and $N_{CREATAS(Ireland)} = 2,599$, $N_{GROSAGR(Albania)} = 8,460$ and $N_{GROSAGR(Ireland)} = 2,201$.

model specification of the model with heterogeneous within-cluster (co)variances is included. For the model specification of the other models, see the complete code in Appendix B.

We investigated the missing patterns of the data in multiple ways: by plots, inferential statistics with Little (1988)'s test of MCAR[1] for multivariate data, correlation tables, and with logistic multilevel models that predicted missingness. In sum, we found evidence that they are not MCAR but MAR. Missing values could be predicted by the value or missingness of the other variable and the country. Thus, missing patterns seem to be largely contingent on the data collection in the schools in both countries. Moreover, a considerable amount of missing values for each variable, given the stark differences in school sizes, is introduced when reformatting to WF (where the data matrix is $g \times p \cdot n_{\max}$) as balanced school sizes are necessary. As Schafer (1997) argued, an unbalanced design can be considered a missing data problem. Multiple imputation has been applied to deal with unbalanced designs in ANOVA before (Ginkel & Kroonenberg, 2021). Thus, we imputed not only the "genuine" missing values but the missing values that had to be introduced by the balanced cluster sizes required for reformatting. We used multiple imputation by chained equations (MICE; Buuren & Groothuis-Oudshoorn, 2011) in the LF data matrix. For each variable, we specified an imputation model containing the other variable as predictor and accounting for the clustering. Imputation was done separately for each country, such that we assumed homogeneous variances within each country. In total, for Albania, 72% of values of creative activities at school and 69% of values of growth mindset, and for Ireland, 34% and 29% of these values were imputed. Admittedly, these quantities are very large but the data sets used for imputation were considerably large as well: for Albania, $N_{CREATAS(Albania)} = 3,398$ and $N_{GROSAGR(Albania)} = 3,870$, and for Ireland, $N_{CREATAS(Ireland9} = 5,051$ and $N_{GROSAGR(Ireland)} = 2,201$. The existent and imputed data is depicted in Figure 5. Moreover, sensitivity analysis revealed that the means and standard deviations of the existent and imputed data sets were very close (see Table 1). Note that we combined the imputed data sets and run the model estimation on this complete data set instead of running separate models for each imputed data set and pooling the results, as suggested by Rubin (2004) and Schafer and Olsen (1998), because our kind of analysis was not supported in the multiple imputation package *MICE*. After multiple imputation, the total sample consisted of $g = 444$ schools with $n = 45$ students, which results in a total of $N = 19,980$ students where $N_{Albania} = 12,330$ and $N_{Ireland} = 7,650$.

Note that because of the nature of the data – a large sample of heterogeneous, clustered data with unbalanced numbers of clusters, highly differing cluster sizes and large amounts of missings – empirical evidence on ways to deal with the missings was sparse. While there was literature on large data sets with missing cases up to 99% per variable (Stuart et al., 2009), moderate sized clustered data ($g = 300$, $n = 2 - 25$; Huque et al., 2020), multigroup data (of randomized control trials; Jakobsen et al., 2017), unbalanced group sizes (Schafer, 1997), heterogeneous variances (with k-nearest neighbours imputation; Santos et al., 2022), and unbalanced group sizes (Schafer, 1997), no study considered all these together. Thus, we combined tested and untested advice in the reported way of dealing with the missing values. Note further that we tried several alternatives. Imputation in the WF data matrix did not

[1]There are different kinds of missing patterns. Missing Completely at Random (MCAR): missings are completely independent of other variables and the missing value itself. Missing at random (MAR): missings are dependent on other variables but not on the missing itself. Missing Not at Random (MNAR): missings are independent of the other variables but they are not random.

Table 1. Mean and standard deviation of existent and imputed data by country.

	Creative activities at school				Growth mindset			
	Albania		Ireland		Albania		Ireland	
Data	M	SD	M	SD	M	SD	M	SD
Existent	1.08	1.34	0.09	0.77	−0.36	1.29	0.16	0.86
Imputed	1.13	1.36	0.07	0.77	−0.33	1.31	0.16	0.87

For sample sizes, see note under Figure 6.

work. A joint imputation model for both countries did not yield plausible results. FIML estimation, doing nothing about the missings, or only imputing the "genuine" missing values (while still introducing a considerable amount of missings by reformatting) did not result in converging models either. In other contexts, however, these might be viable alternatives.

3.2. Model Specification

In the following, we will illustrate how to specify a model with heterogeneous (co)variances at the within-cluster level in the WFmultigroup approach in *lavaan*. There are $p \cdot n = 2 \cdot 45 = 90$ "observed" variables in the WF data matrix which are related mostly by equality constrains. Writing the *lavaan* model syntax manually would take an unnecessary long time. Instead, we use loops for recurring relations. For this, we need to create a vector with the names of the observed variables ('varName'), and one object that contains the number of observed variables 'p'.

```
varNames <- c("CREATAS", "GROSAGR")
p <- length(varNames)
```

We will first create the model syntax for the within-cluster part of the model. The within-cluster variances are estimated as residual variances in a single-level CFA. Thus, we need to specify $p_n \sim\sim p_n$ for all 90 "observed" variables. The n splits of each observed variable p have to be equality constrained in the WF approach in order to estimate the within-cluster parameters. This is achieved by using the same label for the variance parameters. Because we want the within-cluster variances to differ by group, we have to use different labels for the parameters in both groups. In sum, the variances are specified in the following form: 'CREATAS.1 ∼ ∼ c(CREATAS_albania, CREATAS_ireland)*CREATAS.1′ where, for instance, 'CREATAS_albania' denotes the equality constrained variance parameter across all n students in a school of group 1 (i.e., Albania). The whole set of specifications can be done with the following loop:

```
tmp2 <- c()
tmp3 <- c()
resid_var_w_hetero <- c()
for (j in 1:p){
  for (i in 1:n_max){
    tmp2[i] <- paste0(varNames[j], ".", i)
    tmp3[i] <- paste0(tmp2[i], "~~c(", varNames[j], "_albania, ",
                      varNames[j], "_ireland)*", tmp2[i])
  }
  resid_var_w_hetero[j] <- paste(tmp3, collapse="; ")
}
resid_var_w_hetero <- paste(resid_var_w_hetero, collapse="; ")
```

A similar proceeding is required for the group-specific covariances, for instance, 'CREATAS.1 ∼ c(CREATAS_GROSAGR_albania, CREATAS_GROSAGR_ireland) *GROSAGR.1′, where, for instance, 'CREATAS_GROSAGR_albania' is the within-cluster covariance of Albania, which can be created by another loop:

```
resid_cov_w_hetero <- c()
count <- 0
for (i in 1:n_max){
  for(j in 1:p){
    for(m in 1:p){
      if(j != m & m > j){
        count <- count + 1
        resid_cov_w_hetero[count] <-
          paste0(varNames[j], ".", i, "~~c(", varNames[j], "_",
                 varNames[m], "_albania, ", varNames[j], "_",
                 varNames[m], "_ireland)*", varNames[m], ".", i)
      }
    }
  }
}
resid_cov_w_hetero <- paste(resid_cov_w_hetero, collapse="; ")
```

Next we have to set the means of the $p \cdot n$ "observed" variables to zero, as these are aggregated within-cluster variables whose group-specific mean-structure is specified at the between-cluster level (which we will turn to later). We do this in the form 'CREATAS_1 ∼ 0*1′.

```
means_w <- c()
tmp <- c()
count <- 0
for (j in 1:p){
  for (i in 1:n_max){
    count <- count + 1
    tmp[count] <- paste0(varNames[j], ".", i, "~0*1")
  }
}
means_w <- paste(tmp, collapse = "; ")
```

Now that the model syntax for the (heterogeneous) within-cluster parameters is complete, we can move on to those of the (homogeneous) between-cluster parameters. Between-cluster variables are modelled as latent factors by the $p \cdot n$ "observed" variables. Firstly, we have to fix the factor loadings to 1 as all "observed" variables contribute equally to the factor, 'fCREATAS = ∼ 1*CREATAS_1 + 1*CREATAS_2+ … '.

```
fac_load_b <- c()
tmp <- c()
for (j in 1:p){
  for (i in 1:n_max){
    tmp[i] <- paste0("1*", varNames[j], ".", i)
  }
  fac_load_b[j] <- paste0("f", varNames[j], " =~", paste(tmp, collapse="+")
    )
}
fac_load_b <- paste(fac_load_b, collapse="; ")
```

Following, we will specify the factor variances and intercepts, which constitute the between-cluster variances and means, in the forms of 'fCREATAS ∼ 1′ and 'fCREATAS ∼

~fCREATAS', by way of example for the observed variable creative activities at school ('CREATAS'). Since both parameters make use of the same loop, we create them in the same run.

```
fac_var_b <- c()
fac_int_b <- c()
for (j in 1:p){
  fac_var_b[j] <- paste0("f", varNames[j], "~~f", varNames[j])
  fac_int_b[j] <- paste0("f", varNames[j], "~1")
}
fac_var_b <- paste(fac_var_b, collapse="; ")
fac_int_b <- paste(fac_int_b, collapse="; ")
```

Finally, the between-cluster covariance is set as 'fCREATAS ~ ~fGROSAGR' in the following way:

```
fac_cov_b <- c()
count <- 0
for(j in 1:p){
  for(m in 1:p){
    if(j != m & m > j){
      count <- count + 1
      fac_cov_b[count] <- paste0("f", varNames[j], "~~", "f", varNames[m])
    }
  }
}
fac_cov_b <- paste(fac_cov_b, collapse = "; ")
```

Because the factor (co)variances and means (i.e., between-cluster (co)variances and means) require relatively sparse code, we may set them manually in models with sparse observed variables. Now that we finished the model syntax, we can estimate the model by:

```
model_WF_W_homo <- paste(resid_var_w_homo, resid_cov_w_homo, means_w,
                         sep = "; ")
model_WF_B <- paste(fac_load_b, fac_var_b, fac_cov_b, fac_int_b, sep="; ")
model_WFmultigroup_homo <- paste(model_WF_W_homo, model_WF_B, sep="; ")

fit_WFmultigroup <- sem(model = model_WFmultigroup_hetero_B,
                        data = PISA_short_balanced_imp_WF,
                        group="CNT",
                        group.equal = c("lv.variances", "lv.covariances"))
```

where we combined all prior code snippets to our complete model specification 'model_WFmultigroup' and apply it to the imputed data set 'PISA_short_balanced_imp_WF'. The grouping variable country is handed over to 'group="CNT"'. We are able to set the between-cluster (co)variance structure to be equal across groups by group.equal = c("lv.variances", "lv.covariances"), and thus, we do not have to use labels for the (co)variance as for the within-cluster (co)variances. Unfortunately, there is no appropriate shorthand function parameter for equality constraining the manifest variables *n*-wise (i.e., the standard WF approach) per group. Thus, the within-cluster part of the model has to be specified in the model syntax (manually or by the loops we presented).

3.3. Model Parameter Estimates

In Figure 6, the model parameter estimates of the heterogeneous within-cluster (co)variances model are depicted. The within-cluster variances of creative activities at school were 1.73 in Albania and 0.57 in Ireland, and those of growth mindset were 1.68 in Albania and 0.74 in Ireland. In contrast their covariances were quite similarly close to zero: −0.02 in Albania and 0.04 in Ireland. Thus, overall, Albania had larger within-cluster variances than Ireland. These stark differences in variances in the heterogeneous model, $VR_{CREATAS} = 3.04$ and $VR_{GROSAGR} = 2.27$, also had an impact on the group-specific ICC parameter estimates. Albania with its larger within-cluster variances had smaller ICCs. Regarding creative activities at school, the ICC was 0.04 in Albania and 0.11 in Ireland. For growth mindset, estimates were 0.02 for Albania and 0.04 for Ireland. The differences in within-cluster (co)variances in the heterogeneous model, in combination with the differences in between-cluster means, inform us about the substantial differences in the distributions of the observed variables between both countries. Building on this, one might scrutinize differences in both countries in contextual variables such as educational policies, socio-economic status, and cultural programme in order to explain these distributional differences. This might be especially helpful when considering models in which school

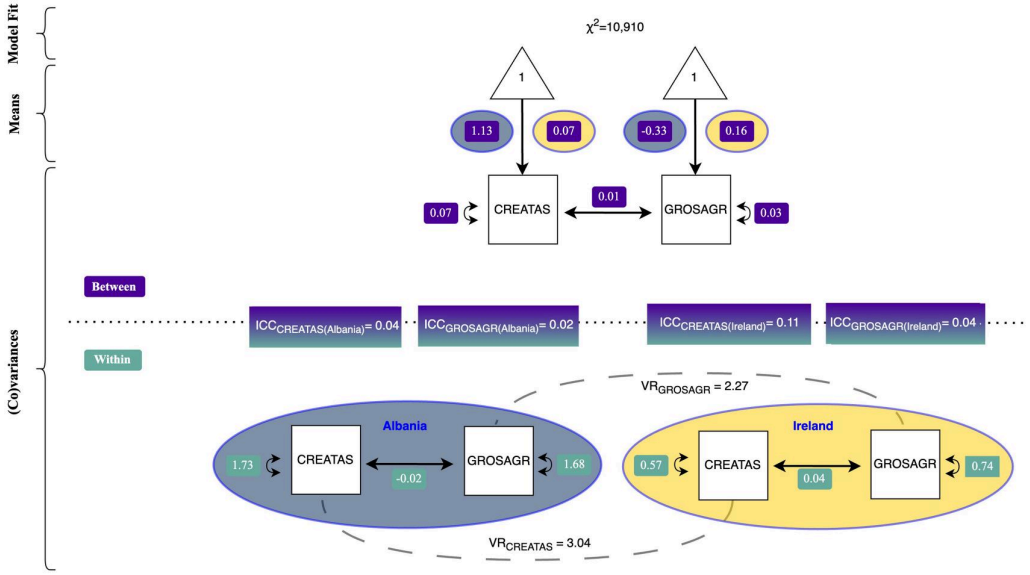


Figure 6. Models with heterogeneous and homogeneous variances. The figure was created manually with the free software draw.io (https://www.drawio.com/).

success is predicted. For subsequent analysis, one could include PISA test results as outcomes that are predicted by both creative activities at school and growth mindset.

4. General Discussion

Modeling heterogeneous within-cluster (co)variances extends traditional within-between variance decomposition and offers the potential to inform further research and educational policy making. The present article has empirically evaluated and illustrated how multilevel multigroup (ML MG) SEMs can be estimated as single-level multigroup restricted CFAs in which grouping is brought about by a discrete between-cluster variable. Within the small simulation study, we found evidence that the proposed WFmultigroup approach can result in accurate and unbiased estimates of a bivariate intercept-only model in settings with moderately large numbers of clusters and cluster sizes ($g > 100$ and $n > 10$ per group). Moreover, results suggest that larger between-cluster variances σ_B^2 and larger VRs (i.e., when variance heterogeneity was larger) can lower the required sample sizes for accurate between-cluster and ICC parameter estimates (and vice versa, that smaller between-cluster variances and smaller VRs require larger sample sizes). With the empirical illustration, we demonstrated the WFmultigroup approach's implementation in R with the package *lavaan*.

Some limitations of the WFmultigroup approach should, however, be noted. Firstly, the WFmultigroup approach might be inadequate when large cluster sizes and/or large numbers of groups are concerned. With the WF data matrix, $(p \cdot n_k) \leq g_k$ is the minimum requirement for convergence due to the implementation of MLE in *lavaan*. If this requirement is not fulfilled, one may need to revert to M*plus* or the "genuine" ML MG SEM in *lavaan*, where the LF data matrix is subjected, which imposes a less restrictive requirement, $p \leq (g_k \cdot n_k)$. Alternatively, full information maximum likelihood (FIML) estimation, which uses the raw data instead of the sample covariance matrices, or Bayesian estimation, which treats each missing value as random variable such that each missing value's uncertainty is accounted for by the uncertainty in the other parameters, might be applied. Note, however, that FIML might result in non-convergence when the amount of missings is too large (as in the empirical data set used in the present article) and that software options for Bayesian estimation in ML MG SEM might be limited. Secondly, when the amount of missing values is substantial and/or when the cluster sizes are highly unbalanced while the number of groups is small, then multiple imputation of the data might be questionable. In our empirical example data set, up to 72% of missing values of a variable in one group were imputed, and we justified the procedure by the large existent sample ($N = 3,398$ and $g = 274$), evidence for the data being MAR, and the results of the sensitivity analysis. However, in other settings, this procedure may not be warranted. Then, one might again resort to the alternatives discussed above. In any case, future research could investigate multiple imputation in the context of large sample, heterogeneous, clustered data with unbalanced numbers of clusters, highly differing cluster sizes and large amounts of missings. Lastly, to apply the WFmultigroup approach, one has to be aware of the grouping variables that give rise to heterogeneous variances. When there is a large quantity of possible between-cluster variables, manual exploration might take a considerable amount of time. An alternative strategy to identify heterogeneous within-cluster (co)variances might be to use classification algorithms such as SEM trees (e.g., Brandmaier et al., 2013). For instance, after estimating a multilevel multigroup model in which each cluster is considered a separate group, SEM trees might help find similarities between clusters that lead to broader groups. However, keep in mind that, depending on the number of observed variables, this approach may require a large amount of computational resources.

Next, possible extensions and applications of the proposed approach are discussed. Firstly, when the data contains a third level (e.g., schools, where level-1 units are students, and level-2 units are classes), but its sample size is scarce (e.g., less than ten units, see Asparouhov & Muthen, 2012), which reduces the chances of a converging model (see e.g., Lüdtke et al., 2011, who found this for level-2), then our WFmultigroup approach might be an appropriate alternative. This scenario is similar to our empirical illustration, where level-1 units were students, level-2 units were schools, and level-3 units, or rather the grouping variable, were countries (though we deliberately selected only two level-3 units). However, notice that cross-level interactions with level-3 variables cannot be modelled this way. Secondly, in contrast to the "genuine" ML MG SEM the WFmultigroup approach allows to free the equality constrains across units within a cluster (i.e., the equality constrains across the $p \cdot n$ "observed" variables of the data matrix in WF can be relaxed). When longitudinal data is concerned, this enables heterogeneous variances at different measurement occasions. For example, in a pre-post-test scenario, one might assume the variances to be smaller in the post condition. Thus, one could have a model which allows for group-specific (i.e., experimental and control condition) as well as time-specific (i.e., pre and post measurements) heterogeneous within- and between-cluster (co)variances. With hierarchical modeling, such a model might be estimated as well but here we could not fit measurement models and multiple outcomes. Thirdly, it would be interesting to explore more complex models that use heterogeneous within-cluster (co)variances as predictors or outcomes. Past research explored these possibilities. For instance, Gröhlich et al. (2009) examined whether homogeneous or heterogeneous ability groups are more suited for predicting learning and students' achievements and McNeish (2021) demonstrated how to estimate location scale models in general form as a multilevel SEM in M*plus*. In the latter, different models for both mean (location) and variance (scale) of outcomes can be specified. Our WFmultigroup approach could extend the scale location models by modeling heterogeneous variances.

Another avenue for future research may be to investigate the effect of the VR more thoroughly. Within our simulation study, we found that the accuracy of between-cluster parameter estimates was larger when the VR was increased. We

suggested that this would be related to the factor analytic modeling within the WF approach. Specifically, between-cluster (co)variances are estimated as common factor (co)variances that are equality constrained across groups. When the VR increased, the ratio of common (i.e., between-cluster) to unique (i.e., within-cluster) variances of the indicators (i.e., the $p \cdot n$ "observed" variables in the WF data matrix) in the second group increased as well, and thereby, the amount of communality of the indicators across both groups increased. Prior research showed that larger commonalities required smaller sample sizes for factor recovery (MacCallum et al., 1999). Future research could scrutinize this hypothesis and validate whether this effect is unique to the WFmultigroup or present in the "genuine" ML MG SEM as well.

The present article proposed a way to estimate heterogeneous within-cluster (co)variances, which are stratified by a discrete between-cluster variable, as multilevel multigroup SEMs in a single-level framework where a restricted CFA for multiple groups is fitted. Moreover, we demonstrated the application in detail with the *lavaan* package in R. We hope that the proposed approach facilitates research and teaching, and inspires new research endeavours that consider and explore heterogeneity of variances.

Disclosure statement

The authors report there are no competing interests to declare.

ORCID

Julia-Kim Walther http://orcid.org/0000-0001-5758-1211
Martin Hecht http://orcid.org/0000-0002-5168-4911
Benjamin Nagengast http://orcid.org/0000-0001-9868-8322
Steffen Zitzmann http://orcid.org/0000-0002-7595-4736

References

Adams, G., Gulliford, M. C., Ukoumunne, O. C., Eldridge, S., Chinn, S., & Campbell, M. J. (2004). Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *Journal of Clinical Epidemiology*, *57*, 785–794. https://doi.org/10.1016/j.jclinepi.2003.12.013

Asparouhov, T., & Muthen, B. (2012). *Multiple Group Multilevel Analysis*. http://www.statmodel.com/examples/webnotes/webnote16.pdf [Technical Report].

Barendse, M. T., & Rosseel, Y. (2020). Multilevel modeling in the 'Wide Format' approach with discrete data: A solution for small cluster sizes. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 696–721. https://doi.org/10.1080/10705511.2019.1689366

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., Krivitsky, P. N., Tanaka, E., & Jagan, M. (2024). *lme4: Linear mixed-effects models using "Eigen" and S4* (Version 1.1-35.5) [Computer software]. https://cran.r-project.org/web/packages/lme4/index.html

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, *28*, 135–167. https://doi.org/10.3102/10769986028002135

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

Buuren, S. v., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67. https://doi.org/10.18637/jss.v045.i03

Buuren, S. v., Groothuis-Oudshoorn, K., Vink, G., Schouten, R., Robitzsch, A., Rockenschaub, P., Doove, L., Jolani, S., Moreno-Betancur, M., White, I., Gaffert, P., Meinfelder, F., Gray, B., Arel-Bundock, V., Cai, M., Volker, T., Costantini, E., Lissa, C. v., & Oberman, H. (2023). *mice: Multivariate Imputation by chained equations* (Version 3.16.0) [Computer software]. https://cran.r-project.org/web/packages/mice/index.html

Candel, M. J., & van Breukelen, G. J. (2015). Sample size calculation for treatment effects in randomized trials with fixed cluster sizes and heterogeneous intraclass correlations and variances. *Statistical Methods in Medical Research*, *24*, 557–573. https://doi.org/10.1177/0962280214563100

Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 327–351. https://doi.org/10.1080/10705511.2014.937849

Finch, W. H., & French, B. F. (2011). Estimation of MIMIC model parameters with multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*, 229–252. https://doi.org/10.1080/10705511.2011.557338

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Ginkel, J. R. v., & Kroonenberg, P. M. (2021). Multiple imputation to balance unbalanced designs for two-way analysis of variance. *Methodology*, *17*, 39–57. https://doi.org/10.5964/meth.6085

Goldstein, H. (2005). Heteroscedasticity and complex variation. *Encyclopedia of Statistics in Behavioral Science*, *2*, 790–795.

Gröhlich, C., Scharenberg, K., & Bos, W. (2009). Wirkt sich Leistungsheterogenität in Schulklassen auf den individuellen Lernerfolg in der Sekundarstufe aus? *Journal for Educational Research Online*, *1*, 86–105. https://doi.org/10.25656/01:4557

Gulliford, M. C., Ukoumunne, O. C., & Chinn, S. (1999). Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: Data from the Health Survey for England 1994. *American Journal of Epidemiology*, *149*, 876–883. https://doi.org/10/gn2gxn

Hamaker, E. L., Dolan, C. V., & Molenaar, P. C. M. (2003). ARMA-based SEM when the number of time points T exceeds the number of cases N: Raw data maximum likelihood. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*, 352–379. https://doi.org/10/dkmnbp

Hedeker, D., & Mermelstein, R. J. (2007). Mixed-effects regression models with heterogeneous variance: Analyzing ecological momentary assessment (EMA) data of smoking. In T. D. Little, J. A. Bovaird & N. A. Card (Eds.), *Modeling ecological and contextual effects in longitudinal studies of human development* (pp. 183–206). Erlbaum.

Hodges, J. S. (1998). Some algebra and geometry for hierarchical models, applied to diagnostics. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *60*, 497–536. https://doi.org/10.1111/1467-9868.00137

Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 157–174. https://doi.org/10.1207/S15328007SEM0802_1

Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, *64*, 157–170. https://doi.org/10.1111/j.1467-9574.2009.00445.x

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Huang, F. L., Wiedermann, W., & Zhang, B. (2023). Accounting for heteroskedasticity resulting from between-group differences in multilevel models. *Multivariate Behavioral Research*, *58*, 637–657. https://doi.org/10.1080/00273171.2022.2077290

Hugh-Jones, D. (2022). *huxtable: Easily create and style tables for LaTeX, HTML and other formats* (Version 5.5.6) [Computer software]. https://CRAN.R-project.org/package=huxtable

Huque, M. H., Moreno-Betancur, M., Quartagno, M., Simpson, J. A., Carlin, J. B., & Lee, K. J. (2020). Multiple imputation methods for handling incomplete longitudinal and clustered data where the target analysis is a linear mixed effects model. *Biometrical Journal. Biometrische Zeitschrift*, *62*, 444–466. https://doi.org/10.1002/bimj.201900051

Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials – A practical guide with flowcharts. *BMC Medical Research Methodology*, *17*, 162. https://doi.org/10.1186/s12874-017-0442-1

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*, 350–386. https://doi.org/10.3102/00346543068003350

Korendijk, E. J. H., Maas, C. J. M., Moerbeek, M., & Van der Heijden, P. G. M. (2008). The influence of misspecification of the heteroscedasticity on multilevel regression parameter and standard error estimates. *Methodology*, *4*, 67–72. https://doi.org/10.1027/1614-2241.4.2.67

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*, 1198–1202. https://doi.org/10.1080/01621459.1988.10478722

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2×2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological Methods*, *16*, 444–467. https://doi.org/10.1037/a0024376

Luo, W., Li, H., Baek, E., Chen, S., Lam, K. H., & Semma, B. (2021). Reporting practice in multilevel modeling: A revisit after 10 years. *Review of Educational Research*, *91*, 311–355. https://doi.org/10.3102/0034654321991229

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*, 127–137. https://doi.org/10.1046/j.0039-0402.2003.00252.x

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*, 84–99. https://doi.org/10.1037/1082-989X.4.1.84

McNeish, D. (2021). Specifying location-scale models for heterogeneous variances as multilevel SEMs. *Organizational Research Methods*, *24*, 630–653. https://doi.org/10.1177/1094428120913083

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*, 259–284. https://doi.org/10.1037/1082-989X.10.3.259

Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methods*, *3*, 45–58.

Muthen, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 267. https://doi.org/10.2307/271070

Muthén, B. (1997). Latent variable modeling of longitudinal and multilevel data. *Sociological Methodology*, *27*, 453–480. https://doi.org/10.1111/1467-9531.271034

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 599–620. https://doi.org/10.1207/S15328007SEM0904_8

Muthén, L., & Muthén, B. (1998–2017). *Mplus user's guide.* (8th Ed.). Muthén & Muthén.

Pedersen, T. L. (2024). *patchwork: The composer of plots* (Version 1.2.0) [Computer software]. https://cran.r-project.org/web/packages/patchwork/index.html

R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.4.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

R Core Team, Bivand, R., Carey, V. J., DebRoy, S., Eglen, S., Guha, R., Herbrandt, S., Lewin-Koh, N., Myatt, M., Nelson, M., Pfaff, B., Quistorff, B., Warmerdam, F., Weigand, S., Foundation, F. S., & Inc. (2024). f*oreign: Read data stored by "Minitab", "S", "SAS", "SPSS", "Stata", "Systat", "Weka", "dBase"*, … (Version 0.8-87) [Computer software]. https://cran.r-project.org/web/packages/foreign/index.html

Raudenbush, S. W., & Bryk, A. S. (1987). Examining correlates of diversity. *Journal of Educational Statistics*, *12*, 241–269. https://doi.org/10.3102/10769986012003241

Revelle, W. (2024). *psych: Procedures for psychological, psychometric, and personality research* (Version 2.4.6.26) [Computer software]. https://cran.r-project.org/web/packages/psych/index.html

Rosopa, P., Brawley, A., Atkinson, T., & Robertson, S. (2019). On the conditional and unconditional type I error rates and power of tests in linear models with heteroscedastic errors. *Journal of Modern Applied Statistical Methods*, *17*. https://doi.org/10.22237/jmasm/1551966828

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

Rosseel, Y., Jorgensen, T. D., Wilde, L. D., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., Barendse, M., Rockwood, N., Scharf, F., Du, H., Jamil, H., & Classe, F. (2024). *lavaan: Latent variable analysis* (Version 0.6-18) [Computer software]. https://cran.r-project.org/web/packages/lavaan/index.html

Rovine, M. J., & Molenaar, P. C. (2000). A structural modeling approach to a multilevel random coefficients model. *Multivariate Behavioral Research*, *35*, 51–88. https://doi.org/10.1207/S15327906MBR3501_3

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys.* (Vol. 81). John Wiley & Sons.

Santos, M. S., Abreu, P. H., Fernández, A., Luengo, J., & Santos, J. (2022). The impact of heterogeneous distance functions on missing data imputation and classification performance. *Engineering Applications of Artificial Intelligence*, *111*, 104791. https://doi.org/10.1016/j.engappai.2022.104791

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, *33*, 545–571. https://doi.org/10.1207/s15327906mbr3304_5

Signorell, A., Aho, K., Alfons, A., Anderegg, N., Aragon, T., Arachchige, C., Arppe, A., Baddeley, A., Barton, K., Bolker, B., Borchers, H. W., Caeiro, F., Champely, S., Chessel, D., Chhay, L., Cooper, N., Cummins, C., Dewey, M., Doran, H. C., … Zeileis, A. (2024). *DescTools: Tools for descriptive statistics* (Version 0.99.50) [Computer software]. https://cran.r-project.org/web/packages/DescTools/index.html

Singer, H. (2010). SEM modeling with singular moment matrices Part I: ML-estimation of time series. *The Journal of Mathematical Sociology*, *34*, 301–320. https://doi.org/10.1080/0022250X.2010.509524

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science*, *57*, 748–761. https://doi.org/10.1111/ajps.12001

Stuart, E. A., Azur, M., Frangakis, C., & Leaf, P. (2009). Multiple imputation with large data sets: A case study of the Children's Mental Health Initiative. *American Journal of Epidemiology*, *169*, 1133–1139. https://doi.org/10.1093/aje/kwp026

Tierney, N., Cook, D., McBain, M., Fay, C., O'Hara-Wild, M., Hester, J., Smith, L., & Heiss, A. (2024). *naniar: Data structures, summaries, and visualisations for missing data* (Version 1.1.0) [Computer software]. https://cran.r-project.org/web/packages/naniar/index.html

Trendafilov, N. T., & Unkel, S. (2011). Exploratory factor analysis of data matrices with more variables than observations. *Journal of Computational and Graphical Statistics*, *20*, 874–891. https://doi.org/10.1198/jcgs.2011.09211

Unkel, S., & Trendafilov, N. T. (2010). Simultaneous parameter estimation in exploratory factor analysis: An expository review. *International Statistical Review*, *78*, 363–382. https://doi.org/10.1111/j.1751-5823.2010.00120.x

Van Montfort, K., Oud, J. H., & Voelkle, M. C. (2018). *Continuous time modeling in the behavioral and related sciences*. Springer.

Voelkle, M. C., Oud, J. H. L., von Oertzen, T., & Lindenberger, U. (2012). Maximum likelihood dynamic factor modeling for arbitrary *N* and *T* Using SEM. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*, 329–350. https://doi.org/10.1080/10705511.2012.687656

Walther, J.-K., Hecht, M., Nagengast, B., & Zitzmann, S. (2024). To be long or to be wide: How data format influences convergence and estimation accuracy in multilevel structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*, 759–774. https://doi.org/10.1080/10705511.2024.2320050

Walther, J.-K., Hecht, M., & Zitzmann, S. (2024). Shrinking small sample problems in multilevel structural equation modeling via regularization of the sample covariance matrix. *Structural Equation Modeling: A Multidisciplinary Journal*. Advance online publication. https://doi.org/10.1080/10705511.2024.2380919

West, B. T., Welch, K. B., & Galecki, A. T. (2022). *Linear mixed models: A practical guide using statistical software*. CRC Press.

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., Brand, T. v. d., Posit, & PBC. (2024). *ggplot2: Create elegant data visualisations using the grammar of graphics* (Version 3.5.1) [Computer software]. https://cran.r-project.org/web/packages/ggplot2/index.html

Wickham, H., François, R., Henry, L., Müller, K., Vaughan, D., Software, P., & PBC. (2023). *dplyr: A grammar of data manipulation* (Version 1.1.4) [Computer software]. https://cran.r-project.org/web/packages/dplyr/index.html

Wickham, H., Vaughan, D., Girlich, M., Ushey, K., Software, P., & PBC. (2024). *tidyr: Tidy messy data* (Version 1.3.1) [Computer software]. https://cran.r-project.org/web/packages/tidyr/index.html

Zitzmann, S. (2018). A computationally more efficient and more accurate stepwise approach for correcting for sampling error and measurement error. *Multivariate Behavioral Research*, *53*, 612–632. https://doi.org/10/gpgn86

Zitzmann, S., Lüdtke, O., Robitzsch, A., & Marsh, H. W. (2016). A Bayesian approach for estimating multilevel latent contextual models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 661–679. https://doi.org/10.1080/10705511.2016.1207179

Zitzmann, S., Nagengast, B., Hübner, N., & Hecht, M. (2024). A simple solution to heteroscedasticity in multilevel nonlinear structural equation modeling [Manuscript submitted for publication]. Department of Psychology, Medical School Hamburg.

Zitzmann, S., Weirich, S., & Hecht, M. (2023). Accurate standard errors in multilevel modeling with heteroscedasticity: A computationally more efficient jackknife technique. *Psych*, *5*, 757–769. https://doi.org/10.3390/psych5030049

Appendix A

Additional figures

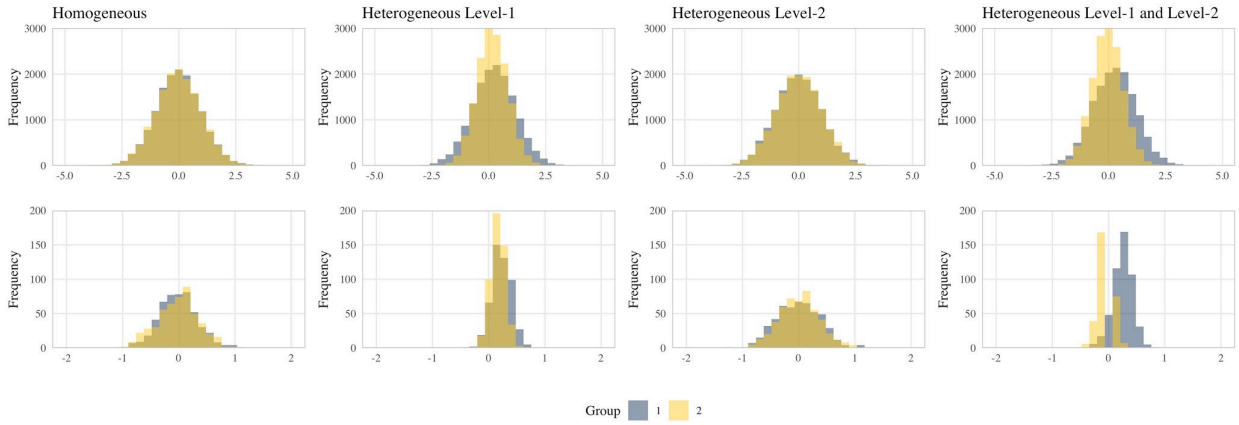


Figure A1. The distributions of raw data and cluster means under homogeneous and heterogeneous conditions. The upper row shows raw data and the lower row cluster means of one observed variable. The simulated heterogeneous conditions have been adapted from the PISA data from the empirical illustration where larger between- and within-cluster variances have been observed in the first group. Accordingly, in the heterogeneous conditions $\sigma^2_{B1} = 0.10$ and $\sigma^2_{W1} = 0.90$ (Group 1), and $\sigma^2_{B2} = 0.05$ and $\sigma^2_{W1} = 0.45$ (Group 2). For homogeneous conditions, both groups have the same variances as the first group. For example, for heterogeneous level-1: $\sigma^2_{B1} = \sigma^2_{B2} = 0.10$, $\sigma^2_{W1} = 0.90$, and $\sigma^2_{W1} = 0.45$. The number of clusters ($g = 3,000$), cluster sizes ($n = 30$), and VRs at both levels ($VR_{between-cluster} = VR_{within-cluster} = 2$) have been simplified. The code to generate the data and the figure can be found on Github (https://github.com/demianJK/WFmultigroup)

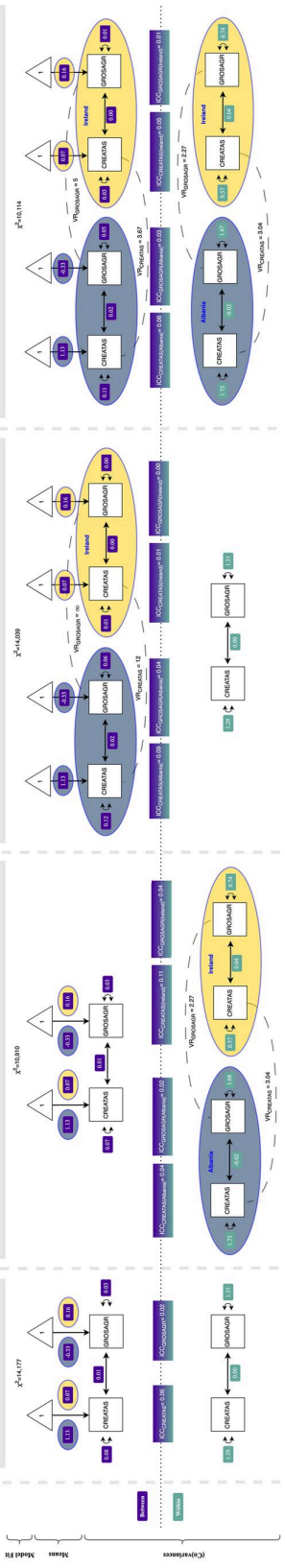


Figure A2. Four types of models with homogeneity and heterogeneity at the between- and within-cluster levels. The code to estimate all parameters for all models can be found in Appendix B. These are the estimates from the WFmultigroup approach, but those of the "genuine" ML MG SEM are very similar.

Appendix B

Complete R Code for empirical illustration

```r
###### (0) Prerequisites

## load required packages
library("dplyr") # select and filter data (version 1.1.4)
library("foreign") # read SPSS (version 0.8-87)
library("ggplot2") # figures (version 3.5.1)
library("huxtable") # APA table (version 5.5.6)
library("lavaan") # ML MG SEM (version 0.6-18)
# Note that this CRAN version of lavaan does not yield the same results in the homogeneous model in the "genuine" ML MG SEM approach
# as the WFmultigroup approach does. However, the most recent version on Github (0.6-19.2187) does so.
# install.packages("devtools")
# library("devtools")
# install_github("yrosseel/lavaan")
library("lme4") # logistic regression of missingness (version 1.1-35.5)
library("mice") # multiple imputation (version 3.16.0)
library("naniar") # MCAR test (version 1.1.0)
library("patchwork") # combining ggplots by + (version 1.2.0)
library("psych") # descriptive stats (version 2.4.6.26)
library("tidyr") # reformating (version 1.3.1)

## load data

# Go to https://www.oecd.org/pisa/data/2022database/
# Navigate to SPSS (TM) Data Files (compressed) >>> Student Questionnaire data file and download the file
PISA <- read.spss("../CY08MSP_STU_QQQ.SAV", to.data.frame=TRUE, use.value.labels = FALSE) # otherwise numerical vectors might be handled as factors
# the data frame is in LF (i.e., each row corresponds to a student)

# If you don't want to run the multiple imputation, simply load the final data frame and continue in line 409.
PISA_short_balanced_imp <- read.csv(file = "/Users/julia/Documents/Arbeit/Promotion/Forschung/Projects/03_WFmultigroup/numerical_ex/PISA_short_balanced_imp.csv")



##### (1) Data Subsetting

## select relevant variables
PISA_short <- select(PISA,
                     CNTSTUID, # unique student ID (level-1)
```

```
                      CNTSCHID, # school (level-2)
                      CNT, # CNT (group)
                      CREATAS, # Creative Activities at school
                      GROSAGR # Growth Mindset
)
# PISA_short is "LF unbalanced"

## select relevant cases (Albania and Ireland) of between-cluster variable
    country
PISA_short <- filter(PISA_short, CNT == "ALB" | CNT == "IRL")



##### (2) Inspecting the Data I: Data Structure and Data Types

## inspect data structure and data types
str(PISA_short)

# is it not necessary to factorise the discrete ID indicators CNTSTUID and
    CNTSCHID...

## ... but we recode the grouping variable for the figures
PISA_short$CNT <- ifelse(PISA_short$CNT == "IRL", yes="Ireland", no="
    Albania")
# (we do not factorise bc otherwise we would introduce problems with data
    subsetting and multiple imputation later on)



##### (3) Inspecting the Data II: Unbalanced Cluster Sizes

## get information on the selected subsample
N <- nrow(PISA_short)
schools <- unique(PISA_short$CNTSCHID)
g <- length(schools)
n <- as.vector(table(PISA_short$CNTSCHID))
n_mean <- mean(n)
n_min <- min(n)
n_max <- max(n)

country <- c()
for (j in 1:g){
  country[j] <-  unique(PISA_short$CNT[PISA_short$CNTSCHID == schools[j]])
}

nData <- data.frame(country = country,
                    school = schools,
                    n = n)
```

```
82
83 a <- ggplot(nData, aes(x=g, fill=country)) +
84   geom_bar(data = transform(nData, country = NULL), fill = "grey85") +
85   geom_bar(show.legend = FALSE) + facet_grid(. ~ country) +
86   scale_y_continuous(name="Frequency", expand=c(0,0)) +
87   scale_x_discrete(name="Number of Schools (g)",) +
88   scale_fill_manual(values=c("#002654", "#ffce00")) +
89   theme_minimal() +
90   theme(text = element_text(family="serif"), panel.grid.minor = element_
     blank(),
91         panel.border = element_rect(color = "grey",  fill = NA, linewidth =
      0.5)) +
92   labs(title="A")
93 # table(country)
94
95 b <- ggplot(nData, aes(x=n, fill=country)) +
96   geom_histogram(data = transform(nData, country = NULL), fill = "grey85",
     binwidth=1) +
97   geom_histogram(binwidth=1, show.legend = FALSE) + facet_grid(country ~ .)
       +
98   scale_y_continuous(name="Number of Schools (g)", expand=c(0,0)) +
99   scale_x_continuous(name="School Size (n)", expand=c(0.01,0.01), limits=c
     (0, NA)) +
100  scale_fill_manual(values=c("#002654", "#ffce00")) +
101  theme_minimal() +
102  theme(text = element_text(family="serif"), panel.grid.minor = element_
     blank(),
103        panel.border = element_rect(color = "grey",  fill = NA, linewidth =
      0.5)) +
104  labs(title="B")
105
106 a + b # Fig.3
107
108 # N=11.698 with g=444 and the distribution of cluster sizes (n) differs
     fairly.
109 # country-wise:
110 table(PISA_short$CNT) # N
111 table(nData$country) # g
112
113
114
115 ##### (4) Inspecting the Data III: Distribution of Variables
116
117 ## Raw Data
118
119 # univariate
120 a <- ggplot(PISA_short, aes(x=CREATAS, fill=CNT)) +
121   geom_histogram(show.legend = FALSE, position = "identity", alpha=0.5) +
```

```
  scale_x_continuous(name="Creative Activities at School (CREATAS)", expand
    =c(0,0), limits=c(-6, 6)) +
  scale_y_continuous(name="Frequency", expand=c(0, 0), limits=c(0, 2000)) +
  scale_fill_manual(values=c("#002654", "#ffce00")) +
  theme_minimal() +
  theme(text = element_text(family="serif"), panel.grid.minor = element_
    blank(),
        panel.border = element_rect(color = "grey",  fill = NA, linewidth =
     0.5)) +
  labs(title="Raw Data", subtitle="A")

b <- ggplot(PISA_short, aes(x=GROSAGR, fill=CNT)) +
  geom_histogram(show.legend = FALSE, position = "identity", alpha=0.5) +
  scale_x_continuous(name="Growth Mindset (GROSAGR)", expand=c(0,0), limits
    =c(-6, 6)) +
  scale_y_continuous(name="Frequency", expand=c(0, 0), limits=c(0, 2000)) +
  scale_fill_manual(values=c("#002654", "#ffce00")) +
  theme_minimal() +
  theme(text = element_text(family="serif"), panel.grid.minor = element_
    blank(),
        panel.border = element_rect(color = "grey",  fill = NA, linewidth =
     0.5)) +
  labs(subtitle="B")

# bivariate
c <- ggplot(PISA_short, aes(x=CREATAS, y=GROSAGR, col=CNT)) +
  geom_point(show.legend = FALSE, alpha=0.3) +
  scale_x_continuous(name="Creative Activities at School (CREATAS)", expand
    =c(0, 0), limits=c(-6.5, 6.5)) +
  scale_y_continuous(name="Growth Mindset (GROSAGR)", expand=c(0,0), limits
    =c(-6, 6)) +
  scale_color_manual(values=c("#002654", "#ffce00")) +
  theme_minimal() +
  theme(text = element_text(family="serif"), panel.grid.minor = element_
    blank(),
        panel.border = element_rect(color = "grey",  fill = NA, linewidth =
     0.5)) +
  labs(subtitle="C")

## Cluster means

# estimate cluster means and create data frame
CREATAS_cluster_means <- aggregate(PISA_short$CREATAS, list(PISA_short$
    CNTSCHID), FUN=mean, na.rm=TRUE, na.action=NULL)
GROSAGR_cluster_means <- aggregate(PISA_short$GROSAGR, list(PISA_short$
    CNTSCHID), FUN=mean, na.rm=TRUE, na.action=NULL)

PISA_short <- PISA_short[order(PISA_short$CNTSCHID),]
```

```

j <- c()
country <- c()
for (i in 1:nrow(PISA_short)){
  tmp_j <- PISA_short$CNTSCHID[i]
  if (i==1){
    country <- append(country, PISA_short$CNT[i])
    j <- append(j, tmp_j)
  } else {
    if (tmp_j > tail(j, n=1)){
      country <- append(country, PISA_short$CNT[i])
      j <- append(j, tmp_j )
    }
  }
}

PISA_short_cluster_means <- data.frame(j=1:444, country=country, CREATAS=
    CREATAS_cluster_means$x, GROSAGR=GROSAGR_cluster_means$x)

# univariate
d <- ggplot(PISA_short_cluster_means, aes(x=CREATAS, fill=country)) +
  geom_histogram(show.legend = FALSE, position = "identity", alpha=0.5) +
  scale_x_continuous(name="Creative Activities at School (CREATAS)", expand
    =c(0, 0),
                     limits=c(-6, 6)) +
  scale_y_continuous(name="Frequency", limits=c(0, 150), expand=c(0, 0),) +
  scale_fill_manual(name="Country", values=c("#002654", "#ffce00")) +
  theme_minimal() +
  theme(text = element_text(family="serif"), panel.grid.minor = element_
    blank(),
        panel.border = element_rect(color = "grey",  fill = NA, linewidth =
    0.5)) +
  guides(colour = guide_legend(override.aes = list(alpha = 1))) +
  labs(title="Cluster Means", subtitle="D")


e <- ggplot(PISA_short_cluster_means, aes(x=GROSAGR, fill=country)) +
  geom_histogram(show.legend = FALSE, position = "identity", alpha=0.5) +
  scale_x_continuous(name="Growth Mindset (GROSAGR)", expand=c(0,0),
                     limits=c(-6, 6)) +
  scale_y_continuous(name="Frequency", limits=c(0, 150), expand=c(0, 0),) +
  scale_fill_manual(name="Country", values=c("#002654", "#ffce00")) +
  theme_minimal() +
  theme(text = element_text(family="serif"), panel.grid.minor = element_
    blank(),
        panel.border = element_rect(color = "grey",  fill = NA, linewidth =
    0.5)) +
  guides(colour = guide_legend(override.aes = list(alpha = 1))) +
```

```
  labs(subtitle="E")

# bivariate
f <- ggplot(PISA_short_cluster_means, aes(x=CREATAS, y=GROSAGR, col=country)) +
  geom_point(alpha=0.3) +
  scale_x_continuous(name="Creative Activities at School (CREATAS)", expand=c(0, 0),
                     limits=c(-6, 6)
  ) +
  scale_y_continuous(name="Growth Mindset (GROSAGR)", expand=c(0,0),
                     limits=c(-6, 6)
  ) +
  scale_color_manual(name="Country", values=c("#002654", "#ffce00")) +
  theme_minimal() +
  theme(text = element_text(family="serif"), panel.grid.minor = element_blank(),
        panel.border = element_rect(color = "grey",  fill = NA, linewidth = 0.5)) +
  guides(colour = guide_legend(override.aes = list(alpha = 1)))  +
  guides(colour = guide_legend(override.aes = list(alpha = 1))) +
  labs(subtitle="F")

a + b + c + d + e + f + plot_layout(nrow=2, guides='collect') & theme(text = element_text("serif"), legend.position = "bottom") # Fig.4



##### (5) Inspecting the Data IV: Missing Data

## What is the proportion of missingness?
vis_miss(PISA_short)
# 28% of CREATAS and 20% of GROSAGR missing
# for each country:
table(is.na(PISA_short$CREATAS), PISA_short$CNT)#/nrow(PISA_short)
table(is.na(PISA_short$GROSAGR), PISA_short$CNT)#/nrow(PISA_short)
# numbers from footnote Fig.3

## Is the missingness systematical?
# MCAR: missings are completely independent of other variables and the missing value itself
# MAR: missings are dependent on other variables but not on the missing itself
# MNAR: missings are independent of the other variables but they are not random

## Let's check the missing patterns (= co-occurence of missings in multiple variables).
```

```
239
240 ## (a) descriptive
241 # by figure with percentages
242 md.pattern(PISA_short, rotate.names = TRUE) # note this function is from
    package mice but mcar_test is from package naniar
243
244 # rows: missing patterns
245 # numbers to left: cases for each missing pattern
246 # number to right: number of missings in missing pattern
247 # numbers at bottom: number of missing cases for each variable (column) -->
     absolute numbers we got in figure before
248
249 # 4 patterns
250 # most often all variables existent (1. row),
251 8137 / (8137 + 1182 + 312 + 2067) # approx. 70% cases without any missings,
     thus, 30% of cases with at least one missing!
252 # then one missing in CREATAS (2. row),
253 (1182) / (8137 + 1182 + 312 + 2067) # approx 10% of only missing CREATAS
254 # then missings in CREATAS and GROSAGR (4. row),
255 (2067) / (8137 + 1182 + 312 + 2067) # approx 18% of missing CREATAS and
    GROSAGR
256 # Note 10% + 18% add up to the 28% missing cases reported for CREATAS
    before
257 # then one missing in GROSAGR (3. row)
258 (312) / (8137 + 1182 + 312 + 2067) # approx 3% of only missing GROSAGR
259
260
261 ## (b) inferential
262 # by using Little's (1988) test that compares patterns of missingness
263 # H0: MCAR
264 # H1: not MCAR
265 # Note CNT and CNTSCHID are perfectly correlated and can thus not be used
    in the same test bc of multicollinearity (i.e., singularity)
266 # we drop CNT
267 mcar_test(PISA_short[, c("CNTSCHID", "CREATAS", "GROSAGR")])
268 # test is significant, thus evidence that MCAR does not hold
269
270 ## explore MAR assumption
271
272 # create missing data indicators (missing=1, existent=0)
273 PISA_short$missing_CREATAS <- ifelse(is.na(PISA_short$CREATAS), yes=1, no
    =0)
274 PISA_short$missing_GROSAGR <- ifelse(is.na(PISA_short$GROSAGR), yes=1, no
    =0)
275
276 ## (a) descriptive
277 # by correlation table
278 cor_data <- PISA_short
```

```r
cor_data$CNT <- ifelse(cor_data$CNT == "Albania", yes=1, no=0) # recode to numeric bc character does not work
cor <- cor(cor_data, use = "pairwise.complete.obs")
cor[upper.tri(cor)] <- NA
print(round(cor, 2), na.print="")

# missingness has large correlation with country (0.393 and 0.431)
# missingness has large correlation with cluster (-0.393 and -0.431)
# contingency of missingness (or presence) of both variables is quite large (0.667), we see this in the missing patterns
# together, this suggest a design effect (i.e., questionnaires not administered in certain clusters in countries)
# missingness has small correlation with other variable (-0.065 and 0.120)
# most importantly, country has moderate to large correlation with the other variable (0.425 and -0.235)

## (b) inferential
# by fitting logistic mixed-effects models to predict missingness
# Note that a variable and their missingness indicator cannot be used in the same model because of multicollinearity (e.g. GROSAGR and missing_GROSAGR).
# Thus, we consider one model for each.

# CREATAS
model_CREATAS <- glmer(missing_CREATAS ~ CNT * GROSAGR + (1 | CNTSCHID), family = binomial, data = PISA_short)
summary(model_CREATAS)
# CNT and GROSAGR predict NA in CREATAS
model_CREATAS_mi <- glmer(missing_CREATAS ~ CNT * missing_GROSAGR + (1 | CNTSCHID), family = binomial, data = PISA_short)
summary(model_CREATAS_mi)
# CNT, NA in GROSAGR, and their interaction predict NA in CREATAS

# GROSAGR
model_GROSAGR <- glmer(missing_GROSAGR ~ CNT * CREATAS + (1 | CNTSCHID), family = binomial, data = PISA_short)
summary(model_GROSAGR)
# CNT predicts NA in CREATAS
model_GROSAGR_mi <- glmer(missing_GROSAGR ~ CNT * missing_CREATAS + (1 | CNTSCHID), family = binomial, data = PISA_short)
summary(model_GROSAGR_mi)
# CNT, NA in CREATAS, and their interaction predict NA in GROSAGR

# evidence for MAR: missingness can be predicted by other variables (or missingness of other variables) in data and country
# thus imputation is warranted, but first we inspect another source of missingness and estimation problems

```

```
315
316
317 ##### (6) Reformating I: Balanced Cluster Sizes in LF
318 # necessary for imputing unbalanced data, and to reformat to WF later
319
320 ## create new data frame with balanced number of students
321 PISA_short_balanced <- data.frame(
322   j = rep(1:g, each=n_max),
323   i = rep(1:n_max, times=g),
324   CNTSCHID = rep(NA, n_max*g) , # incomplete
325   CNTSTUID = rep(NA, n_max*g), # incomplete
326   CNT = rep(NA, n_max*g),
327   CREATAS = rep(NA, n_max*g),
328   GROSAGR = rep(NA, n_max*g),
329   missing_CREATAS = rep(1, n_max*g),
330   missing_GROSAGR = rep(1, n_max*g)
331 )
332
333 # sort data by school
334 PISA_short <- PISA_short[with(PISA_short, order(CNTSCHID)), ]
335
336 # fill in existing data
337 for (j in 1:g) {
338   school <- unique(PISA_short$CNTSCHID)[j]
339   students <- filter(PISA_short, CNTSCHID == school)$CNTSTUID
340   nSchool <- length(students)
341   PISA_short_balanced$CNTSCHID[((j - 1) * n_max + 1):((j - 1) * n_max +
      nSchool)] <- school
342   PISA_short_balanced$CNTSTUID[((j - 1) * n_max + 1):((j - 1) * n_max +
      nSchool)] <- students
343   PISA_short_balanced$CNT[((j - 1) * n_max + 1):((j - 1) * n_max + n_max)]
      <- unique(PISA_short$CNT[which(PISA_short$CNTSCHID == school)])
344   PISA_short_balanced$CREATAS[((j - 1) * n_max + 1):((j - 1) * n_max +
      nSchool)] <- PISA_short$CREATAS[which(PISA_short$CNTSCHID == school)]
345   PISA_short_balanced$GROSAGR[((j - 1) * n_max + 1):((j - 1) * n_max +
      nSchool)] <- PISA_short$GROSAGR[which(PISA_short$CNTSCHID == school)]
346   PISA_short_balanced$missing_CREATAS[((j - 1) * n_max + 1):((j - 1) * n_
      max + nSchool)] <- PISA_short$missing_CREATAS[which(PISA_short$CNTSCHID
      == school)]
347   PISA_short_balanced$missing_GROSAGR[((j - 1) * n_max + 1):((j - 1) * n_
      max + nSchool)] <- PISA_short$missing_GROSAGR[which(PISA_short$CNTSCHID
      == school)]
348 }
349
350 # Now N=n_max*g = 19980 level-1 units.
351 # Final subsample per country: g*n_max = N
352 table(nData$country)*n_max
353
```

```r
# "genuine" missings and unbalanced data
table(PISA_short_balanced$missing_CREATAS, PISA_short_balanced$CNT)
table(PISA_short_balanced$missing_GROSAGR, PISA_short_balanced$CNT)
# numbers from footnote Fig.4



##### (7) Multiple Imputation
# in LF and country-wise

# set imputation method for CREATAS and GROSAGR
meth <- mice(PISA_short_balanced, maxit = 0)$method
meth["CNTSCHID"] <- ""
meth[c("CREATAS", "GROSAGR")] <- "2l.pan"  # homogeneous variances in each group (i.e., country) assumed

# create imputation models for CREATAS and GROSAGR
pred <- make.predictorMatrix(PISA_short_balanced)
pred[ , "j"] <- -2  # Set cluster variable
pred[c("j", "i", "CNTSCHID", "CNTSTUID", "CNT", "missing_CREATAS", "missing_GROSAGR"), ] <- 0  # no models for these variables
pred[ , c("i", "CNTSCHID", "CNTSTUID", "CNT", "missing_CREATAS", "missing_GROSAGR") ] <- 0  # not used as predictors ###### no CNT

# impute
imp_Albania <- mice(filter(PISA_short_balanced, CNT == "Albania"), predictorMatrix = pred, method = meth, seed = 123)
imp_Ireland <- mice(filter(PISA_short_balanced, CNT == "Ireland"), predictorMatrix = pred, method = meth, seed = 123)

# inspect single imputed data sets
stripplot(imp_Albania, CREATAS, pch = 19, xlab = "Imputation number")
stripplot(imp_Ireland, CREATAS, pch = 19, xlab = "Imputation number")
stripplot(imp_Albania, GROSAGR, pch = 19, xlab = "Imputation number")
stripplot(imp_Ireland, GROSAGR, pch = 19, xlab = "Imputation number")
# Because the imputed data sets appear quite similar, we will combine them instead of estimating models for each
# data set and pooled the results.

# compare descriptive stats of existent and imputed data (Tab.1)
ex_Alb <- describe(select(PISA_short_balanced[PISA_short_balanced$CNT == "Albania",], CREATAS, GROSAGR))
imp_Alb <- describe(select(complete(imp_Albania), CREATAS, GROSAGR))
ex_Ire <- describe(select(PISA_short_balanced[PISA_short_balanced$CNT == "Ireland",], CREATAS, GROSAGR))
imp_Ire <- describe(select(complete(imp_Ireland), CREATAS, GROSAGR))
# for both countries, mean and sd are quite similar in the existent and imputed data
```

```
393
394 # combine imputed data sets of both groups (i.e., countries)
395 PISA_short_balanced_imp <- rbind(complete(imp_Albania), complete(imp_
    Ireland))
396
397 # plot imputed data (Fig.5)
398 ggplot(PISA_short_balanced_imp, aes(x=CREATAS, y=GROSAGR, col=CNT)) +
399   geom_point(data = transform(PISA_short_balanced_imp, CNT = NULL), col="
     grey85", alpha=0.5) +
400   geom_point(show.legend = FALSE, alpha=0.3) +
401   facet_grid(CNT ~ missing_CREATAS, margins=TRUE, # adds an additional
     facet for all levels combined
402              labeller=as_labeller(c('0'="Existent", '1'="Imputed", '(all)'=
     "All", 'Albania'="Albania", 'Ireland'="Ireland"))) +
403   scale_x_continuous(name="Creative Activities at School (CREATAS)", expand
     =c(0, 0), limits=c(-6.5, 6.5)) +
404   scale_y_continuous(name="Growth Mindset (GROSAGR)", expand=c(0.05,0.05),
     limits=c(-6, 6)) +
405   scale_color_manual(values=c("#002654", "#ffce00", "black")) +
406   theme_minimal() +
407   theme(text = element_text(family="serif"), panel.grid.minor = element_
     blank(),
408         panel.border = element_rect(color = "grey",  fill = NA, linewidth =
      0.5))
409
410
411
412 ##### (8) Reformating II: Format LF to WF
413 # where each row corresponds to a school
414 PISA_short_balanced_imp_WF <- select(PISA_short_balanced_imp, -c("CNTSCHID"
    , "CNTSTUID", "missing_CREATAS", "missing_GROSAGR")) # drop variables,
    otherwise formating faulty
415 PISA_short_balanced_imp_WF <- pivot_wider(PISA_short_balanced_imp_WF, names
    _from = i, values_from = c("CREATAS", "GROSAGR"), names_sep = ".")
416
417
418
419 ##### (9) Model Estimation
420
421 ## Homogeneity/Heterogeneity is set differently for both levels:
422 # Level-1: in model syntax (by using same or different parameter labels)
423 # Level-2: with function parameter   group  .equal = c("lv.variances", "lv.
    covariances")    (by setting it or leaving it out)
424 # Thus, the model syntax below is the same for all models.
425 # (Note that the variances at each level in the homogeneous models equal
    the pooled variances in the heterogeneous models.)
426
427 varNames <- c("CREATAS", "GROSAGR") # variable names in vector required for
```

```
    loop
p <- length(varNames)

## within (p*n)

# means set to 0
means_w <- c()
tmp <- c()
count <- 0
for (j in 1:p){
  for (i in 1:n_max){
    count <- count + 1
    tmp[count] <- paste0(varNames[j], ".", i, "~0*1")
  }
}
means_w <- paste(tmp, collapse = "; ")


## between (p)

# factor loadings
fac_load_b <- c()
tmp <- c()
for (j in 1:p){
  for (i in 1:n_max){
    tmp[i] <- paste0("1*", varNames[j], ".", i)
  }
  fac_load_b[j] <- paste0("f", varNames[j], " =~", paste(tmp, collapse="+")
   )
}
fac_load_b <- paste(fac_load_b, collapse="; ")

# variances and means
fac_var_b <- c()
fac_int_b <- c()
for (j in 1:p){
  fac_var_b[j] <- paste0("f", varNames[j], "~~f", varNames[j])
  fac_int_b[j] <- paste0("f", varNames[j], "~1")
}
fac_var_b <- paste(fac_var_b, collapse="; ")
fac_int_b <- paste(fac_int_b, collapse="; ")

# covariances
fac_cov_b <- c()
count <- 0
for(j in 1:p){
  for(m in 1:p){
    if(j != m & m > j){
```

```
      count <- count + 1
      fac_cov_b[count] <- paste0("f", varNames[j], "~~", "f", varNames[m])
    }
  }
}
fac_cov_b <- paste(fac_cov_b, collapse = "; ")

model_WF_B <- paste(fac_load_b, fac_var_b, fac_cov_b, fac_int_b, sep="; ")


### Model with Homogeneous Within- and Between-Cluster (Co)variances

## within (p*n)

# variances
tmp2 <- c()
resid_var_w_homo <- c()
tmp3 <- c()
for (j in 1:p){
  for (i in 1:n_max){
    tmp2[i] <- paste0(varNames[j], ".", i)
    tmp3[i] <- paste0(tmp2[i], "~~c(", varNames[j], "_both, ", varNames[j],
    "_both)*", tmp2[i]) # same label for parameter ACROSS groups
  }
  resid_var_w_homo[j] <- paste(tmp3, collapse="; ")
}
resid_var_w_homo <- paste(resid_var_w_homo, collapse="; ")

# covariances
resid_cov_w_homo <- c()
count <- 0
for (i in 1:n_max){
  for(j in 1:p){
    for(m in 1:p){
      if(j != m & m > j){
        count <- count + 1
        resid_cov_w_homo[count] <- paste0(varNames[j], ".", i, "~~c(",
   varNames[j], "_", varNames[m], "_both, ", varNames[j], "_", varNames[m],
    "_both)*", varNames[m], ".", i) # same label for parameter ACROSS
   groups
      }
    }
  }
}
resid_cov_w_homo <- paste(resid_cov_w_homo, collapse="; ")

model_WF_W_homo <- paste(resid_var_w_homo, resid_cov_w_homo, means_w, sep =
    "; ")
```

```r

model_WFmultigroup_homo <- paste(model_WF_W_homo, model_WF_B, sep="; ")

fit_WFmultigroup_homo <- sem(model = model_WFmultigroup_homo,
                             data = PISA_short_balanced_imp_WF,
                             group="CNT",
                             group.equal = c("lv.variances", "lv.covariances"))
summary(fit_WFmultigroup_homo)


## the "genuine" lavaan ML MG SEM returns very similar estimates in the most recent version on Github (0.6-19.2186).
# Note that here only the function parameter "group.equal" controls whether a fully homogeneous/heterogeneous model is estimated.

model_MLMGSEM <- c(
  "
  Group: 1
  Level: 1
  CREATAS ~~ CREATAS
  GROSAGR ~~ GROSAGR
  CREATAS ~~ GROSAGR
  Level: 2
  CREATAS ~~ CREATAS
  GROSAGR ~~ GROSAGR
  CREATAS ~~ GROSAGR

  Group: 2
  Level: 1
  CREATAS ~~ CREATAS
  GROSAGR ~~ GROSAGR
  CREATAS ~~ GROSAGR
  Level: 2
  CREATAS ~~ CREATAS
  GROSAGR ~~ GROSAGR
  CREATAS ~~ GROSAGR
  "
) # Note that the same model syntax is used for the fully heterogeneous model later.
# Alternatively, one could use parameter labels to denote homogeneous (i.e., same label in both groups) or heterogeneous (i.e., differen labels in both groups) parameters
# (just as in the WFmultigroup approach; see also the models that are heterogeneous at one level in the genuine ML MG SEM approach).

fit_MLMGSEM_homo <- sem(model = model_MLMGSEM,
                        data = PISA_short_balanced_imp, # data in LF!
```

```
                        cluster="j",
                        group="CNT",
                        group.equal = c("residuals", "residual.covariances")
    # homogeneous
)
summary(fit_MLMGSEM_homo)



### Model with Heterogeneous Within-Cluster (Co)variances

## within (p*n)

# variances (with n-wise equality constraints)
tmp2 <- c()
tmp3 <- c()
resid_var_w_hetero <- c()
for (j in 1:p){
  for (i in 1:n_max){
    tmp2[i] <- paste0(varNames[j], ".", i)
    tmp3[i] <- paste0(tmp2[i], "~~c(", varNames[j], "_albania, ", varNames[
    j], "_ireland)*", tmp2[i]) # same label for parameter WITHIN groups
  }
  resid_var_w_hetero[j] <- paste(tmp3, collapse="; ")
}
resid_var_w_hetero <- paste(resid_var_w_hetero, collapse="; ")

# covariances (with n-wise equality constraints)
resid_cov_w_hetero <- c()
count <- 0
for (i in 1:n_max){
  for(j in 1:p){
    for(m in 1:p){
      if(j != m & m > j){
        count <- count + 1
        resid_cov_w_hetero[count] <- paste0(varNames[j], ".", i, "~~c(",
    varNames[j], "_", varNames[m], "_albania, ", varNames[j], "_", varNames[
    m], "_ireland)*", varNames[m], ".", i) # same label for parameter WITHIN
     groups
      }
    }
  }
}
resid_cov_w_hetero <- paste(resid_cov_w_hetero, collapse="; ")

model_WF_W_hetero <- paste(resid_var_w_hetero, resid_cov_w_hetero, means_w,
     sep = "; ")

```

```r

model_WFmultigroup_hetero_B <- paste(model_WF_W_homo, model_WF_B, sep="; ")

fit_WFmultigroup_hetero_B <- sem(model = model_WFmultigroup_hetero_B,
                                 data = PISA_short_balanced_imp_WF,
                                 group="CNT"#,
                                 #group.equal = c("lv.variances", "lv.covariances")
                                 )
summary(fit_WFmultigroup_hetero_B)


## Here you have to use the most recent version on Github (0.6-19.2186) again with its "genuine" ML MG SEM which yields very similar estimates.

model_MLMGSEM_hetero_B <- c(
  "
  Group: 1
  Level: 1
  CREATAS ~~ CREATAS_both*CREATAS
  GROSAGR ~~ GROSAGR_both*GROSAGR
  CREATAS ~~ CREATAS_GROSAGR_both*GROSAGR
  Level: 2
  CREATAS ~~ CREATAS_albania*CREATAS
  GROSAGR ~~ GROSAGR_albania*GROSAGR
  CREATAS ~~ CREATAS_GROSAGR_albania*GROSAGR

  Group: 2
  Level: 1
  CREATAS ~~ CREATAS_both*CREATAS
  GROSAGR ~~ GROSAGR_both*GROSAGR
  CREATAS ~~ CREATAS_GROSAGR_both*GROSAGR
  Level: 2
  CREATAS ~~ CREATAS_ireland*CREATAS
  GROSAGR ~~ GROSAGR_ireland*GROSAGR
  CREATAS ~~ CREATAS_GROSAGR_ireland*GROSAGR
  "
)

fit_MLMGSEM_hetero_B <- sem(model = model_MLMGSEM_hetero_B,
                            data = PISA_short_balanced_imp,
                            cluster="j",
                            group="CNT"
)
summary(fit_MLMGSEM_hetero_B)


```

```
### Model with Heterogeneous Within and Between-Cluster (Co)variances

model_WFmultigroup_hetero_WB <- paste(model_WF_W_hetero, model_WF_B, sep="; ")

fit_WFmultigroup_hetero_WB <- sem(model = model_WFmultigroup_hetero_WB,
                                  data = PISA_short_balanced_imp_WF,
                                  group="CNT"#,
                                  #group.equal = c("lv.variances", "lv.covariances")
)
summary(fit_WFmultigroup_hetero_WB)

## the "genuine" lavaan ML MG SEM returns very similar estimates

fit_MLMGSEM_hetero_WB <- sem(model = model_MLMGSEM,
                         data = PISA_short_balanced_imp,
                         cluster="j",
                         group="CNT"#,
                         #group.equal = c("residuals", "residual.covariances")
)
summary(fit_MLMGSEM_hetero_WB)



### Model Comparisons

anova(fit_WFmultigroup_homo, fit_WFmultigroup_hetero_W)
anova(fit_WFmultigroup_homo, fit_WFmultigroup_hetero_B)
anova(fit_WFmultigroup_homo, fit_WFmultigroup_hetero_WB)
anova(fit_WFmultigroup_hetero_B, fit_WFmultigroup_hetero_WB)
# the most complex model, that has heterogeneous within- and between-cluster (co)variances, fits the data best
```

Bibliography

Adams, G., Gulliford, M. C., Ukoumunne, O. C., Eldridge, S., Chinn, S., & Campbell, M. J. (2004). Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *Journal of Clinical Epidemiology*, *57*(8), 785–794. https://doi.org/10.1016/j.jclinepi.2003.12.013

Afshartous, D. (1995). Determination of Sample Size for Multilevel Model Design. *Annual Meeting of the American Educational, San Francisco, CA*.

Ahn, S. C., & Horenstein, A. R. (2013). Eigenvalue Ratio Test for the Number of Factors. *Econometrica*, *81*(3), 1203–1227. https://doi.org/10/f4x9dd

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society: Series A (General)*, *149*(1), 1–26. https://doi.org/10.2307/2981882

Allais, D. C. (1964). *The Selection of Measurements for Prediction*. Stanford, California.

Anand, S., & Segal, P. (2015). The Global Distribution of Income. In *Handbook of Income Distribution* (pp. 937–979, Vol. 2). Elsevier. https://doi.org/10.1016/B978-0-444-59428-0.00012-6

Anderson, T. W. (2003). *Introduction to multivariate statistical analysis* (3rd ed). Wiley.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Arrindell, W. A., & Van Der Ende, J. (1985). An Empirical Test of the Utility of the Observations-To-Variables Ratio in Factor and Components Analysis. *Applied Psychological Measurement*, *9*(2), 165–178. https://doi.org/10.1177/014662168500900205

Arruda, E. H., & Bentler, P. M. (2017). A Regularized GLS for Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(5), 657–665. https://doi.org/10/gcmfh5

Arruda, E. H. (2017). *Applications of Regularization to SEM: Shrinking Eigenvalues to Improve Stability of Covariance Matrices* [Doctoral dissertation, UCLA].

Banerjee, S., & Monni, S. (2021). An Orthogonally Equivariant Estimator of the Covariance Matrix in High Dimensions and for Small Sample Sizes. *Journal of Statistical Planning and Inference*, *213*, 16–32. https://doi.org/10/gpgn3t

Barendse, M. T., & Rosseel, Y. (2020). Multilevel Modeling in the 'Wide Format' Approach with Discrete Data: A Solution for Small Cluster Sizes. *Structural Equation Modeling: A*

Multidisciplinary Journal, *27*(5), 696–721. https://doi.org/10.1080/10705511.2019.1689366

Barendse, M. T., & Rosseel, Y. (2023). Multilevel SEM with random slopes in discrete data using the pairwise maximum likelihood. *British Journal of Mathematical and Statistical Psychology*, *76*(2), 327–352. https://doi.org/10.1111/bmsp.12294

Barrett, P. T., & Kline, P. (1981). The observation to variable ratio in factor analysis. *Personality Study & Group Behaviour*, *1*(1), 23–33.

Bauer, D. J. (2003). Estimating Multilevel Linear Models as Structural Equation Models. *Journal of Educational and Behavioral Statistics*, *28*(2), 135–167. https://doi.org/10.3102/10769986028002135

Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical Considerations for Using Exploratory Factor Analysis in Educational Research. *Practical Assessment, Research, and Evaluation*, *18*(1), 1–13. https://doi.org/10.7275/qv2q-rk76

Bentler, P. M., & Yuan, K.-H. (2011). Positive Definiteness via Off-Diagonal Scaling of a Symmetric Indefinite Matrix. *Psychometrika*, *76*(1), 119–123. https://doi.org/10.1007/s11336-010-9191-3

Bhargava, A. K., & Disch, D. (1982). Exact probabilities of obtaining estimated non-positive definite between-group covariance matrices. *Journal of Statistical Computation and Simulation*, *15*(1), 27–32. https://doi.org/10.1080/00949658208810561

Bickel, P. J., Li, B., Tsybakov, A. B., Van De Geer, S. A., Yu, B., Valdés, T., Rivero, C., Fan, J., & Van Der Vaart, A. (2006). Regularization in statistics. *Test*, *15*(2), 271–344. https://doi.org/10.1007/BF02607055

Blanca, M. J., Alarcón, R., & Bono, R. (2018). Current Practices in Data Analysis Procedures in Psychology: What Has Changed? *Frontiers in Psychology*, *9*, 2558. https://doi.org/10.3389/fpsyg.2018.02558

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in lisrel maximum likelihood estimation. *Psychometrika*, *50*(2), 229–242. https://doi.org/10/fdx7bj

Boyd, L. H., & Iversen, G. R. (1979). *Contextual analysis: Concepts and statistical techniques*. Wadsworth Publishing Company.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, *24*(6), 2350–2383. https://doi.org/10.1214/aos/1032181158

Bryk, A. S., & Raudenbush, S. W. (1987). Application of Hierarchical Linear Models to Assessing Change. *Psychological Bulletin*, *101*(1), 147–158. https://doi.org/10.1037/0033-2909.101.1.147

Bryk, A. S., & Weisberg, H. I. (1977). Use of the nonequivalent control group design when subjects are growing. *Psychological Bulletin*, *84*(5), 950. https://doi.org/10.1037/0033-2909.84.5.950

Burghgraeve, E., De Neve, J., & Rosseel, Y. (2021). Estimating Structural Equation Models Using James–Stein Type Shrinkage Estimators. *Psychometrika*, *86*(1), 96–130. https://doi.org/10/gpgpdq

Candel, M. J., & van Breukelen, G. J. (2015). Sample size calculation for treatment effects in randomized trials with fixed cluster sizes and heterogeneous intraclass correlations and variances. *Statistical Methods in Medical Research*, *24*(5), 557–573. https://doi.org/10.1177/0962280214563100

Cattell, R. B. (1978). *The Scientific Use of Factor Analysis in Behavioral and Life Sciences*. Springer US.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper Solutions in Structural Equation Models: Causes, Consequences, and Strategies. *Sociological Methods & Research*, *29*(4), 468–508. https://doi.org/10/cs43xw

Chou, C.-P., Bentler, P. M., & Pentz, M. A. (1998). Comparisons of two statistical approaches to study growth curves: The multilevel model and the latent curve analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(3), 247–266. https://doi.org/10.1080/10705519809540104

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A Nondegenerate Penalized Likelihood Estimator for Variance Parameters in Multilevel Models. *Psychometrika*, *78*(4), 685–709. https://doi.org/10.1007/s11336-013-9328-2

Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *Journal of Epidemiology & Community Health*, *62*(8), 752–758. https://doi.org/10.1136/jech.2007.060798

Cole, S. R., Chu, H., & Greenland, S. (2014). Maximum likelihood, profile likelihood, and penalized likelihood: A primer. *American Journal of Epidemiology*, *179*(2), 252–260. https://doi.org/10.1093/aje/kwt245

Curran, P. J. (2003). Have Multilevel Models Been Structural Equation Models All Along? *Multivariate Behavioral Research*, *38*(4), 529–569. https://doi.org/10.1207/s15327906mbr3804_5

De Jonckere, J., & Rosseel, Y. (2023). A Model-Based Shrinkage Target to Avoid Non-convergence in Small Sample SEM. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(6), 941–955. https://doi.org/10.1080/10705511.2023.2171420

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D., & Lee, R. S. (2009). Multilevel Modeling: A Review of Methodological Issues and Applications. *Review of Educational Research*, *79*(1), 69–102. https://doi.org/10.3102/0034654308325581

Dempster, A. P. (1972). Covariance selection. *Biometrics*, *28*(1), 157–175. https://doi.org/10.2307/2528966

Deng, L., Yang, M., & Marcoulides, K. M. (2018). Structural Equation Modeling With Many Variables: A Systematic Review of Issues and Developments. *Frontiers in Psychology*, *9*, 1–14. https://doi.org/10.3389/fpsyg.2018.00580

Depaoli, S., & Clifton, J. P. (2015). A Bayesian Approach to Multilevel Structural Equation Modeling With Continuous and Dichotomous Outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(3), 327–351. https://doi.org/10.1080/10705511.2014.937849

Dey, D. K., & Srinivasan, C. (1985). Estimation of a covariance matrix under Stein's loss. *The Annals of Statistics*, *13*(4), 1581–1591.

Dufour, J.-M., & King, M. L. (1991). Optimal invariant tests for the autocorrelation coefficient in linear regressions with stationary or nonstationary AR (1) errors. *Journal of Econometrics*, *47*(1), 115–143.

Duncan, T. E., Duncan, S. C., Alpert, A., Hops, H., Stoolmiller, M., & Muthen, B. (1997). Latent Variable Modeling of Longitudinal and Multilevel Substance Use Data. *Multivariate Behavioral Research*, *32*(3), 275–318. https://doi.org/10.1207/s15327906mbr3203_3

Engel, J., Buydens, L., & Blanchet, L. (2017). An overview of large-dimensional covariance and precision matrix estimators with applications in chemometrics. *Journal of Chemometrics*, *31*(4), e2880. https://doi.org/10/f95v95

Everitt, B. S. (1975). Multivariate Analysis: The Need for Data, and other Problems. *British Journal of Psychiatry*, *126*(3), 237–240. https://doi.org/10.1192/bjp.126.3.237

Fan, J., Liao, Y., & Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, *19*(1), C1–C32. https://doi.org/10/f8k6fz

Finch, W. H., & French, B. F. (2011). Estimation of MIMIC Model Parameters with Multilevel Data. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(2), 229–252. https://doi.org/10.1080/10705511.2011.557338

Fisher, T. J., & Sun, X. (2011). Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. *Computational Statistics & Data Analysis*, *55*(5), 1909–1918. https://doi.org/10/c94f7n

George, E. I., & Oman, S. D. (1996). Multiple-Shrinkage Principal Component Regression. *The Statistician*, *45*(1), 111. https://doi.org/10/dgv4hr

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, *78*(1), 45–51. https://doi.org/10.1093/biomet/78.1.45

Goldstein, H. (2005). Heteroscedasticity and complex variation. *Encyclopedia of statistics in behavioral science*, *2*, 790–795.

Goldstein, H. (2011). *Multilevel statistical models*. John Wiley & Sons.

Goldstein, H., Kounali, D., & Robinson, A. (2008). Modelling measurement errors and category misclassifications in multilevel models. *Statistical Modelling*, *8*(3), 243–261. https://doi.org/10.1177/1471082X0800800302

Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika*, *53*(4), 455–467. https://doi.org/10.1007/BF02294400

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D., & Thomas, S. (1993). A Multilevel Analysis of School Examination Results. *Oxford Review of Education*, *19*(4), 425–433. https://doi.org/10.1080/0305498930190401

Gorsuch, R. L. (1983a). *Factor Analysis* (2nd ed.). Psychology Press. https://doi.org/10.4324/9780203781098

Gorsuch, R. L. (1983b). Factor Analysis. Hillsdale, NJ: Lawrence Earlbaum Associates.

Gray, H., Leday, G. G. R., Vallejos, C. A., & Richardson, S. (2018). *Shrinkage estimation of large covariance matrices using multiple shrinkage targets* [Unpublished Manuscript].

Greenland, S., Schwartzbaum, J. A., & Finkle, W. D. (2000). Problems due to Small Samples and Sparse Data in Conditional Logistic Regression Analysis. *American Journal of Epidemiology*, *151*(5), 531–539. https://doi.org/10.1093/oxfordjournals.aje.a010240

Grilli, L., & Rampichini, C. (2011). The Role of Sample Cluster Means in Multilevel Models. *Methodology*, *7*(4), 121–133. https://doi.org/10/fhzx29

Gulliford, M. C., Ukoumunne, O. C., & Chinn, S. (1999). Components of Variance and Intraclass Correlations for the Design of Community-based Surveys and Intervention Studies: Data from the Health Survey for England 1994. *American Journal of Epidemiology*, *149*(9), 876–883. https://doi.org/10/gn2gxn

Hamaker, E. L., Dolan, C. V., & Molenaar, P. C. M. (2003). ARMA-Based SEM When the Number of Time Points T Exceeds the Number of Cases N: Raw Data Maximum Likelihood. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(3), 352–379. https://doi.org/10/dkmnbp

Hart, R. A., & Clark, D. H. (1999). Does size matter? Exploring the small sample properties of maximum likelihood estimation. *Annual Meeting of the Midwest Political Science Association, Chicago, IL*.

Hayashi, K., Yuan, K.-H., & Liang, L. (2018). On the Bias in Eigenvalues of Sample Covariance Matrix. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative Psychology* (pp. 221–233, Vol. 233). Springer International Publishing. https://doi.org/10.1007/978-3-319-77249-3_19

Hecht, M., Gische, C., Vogel, D., & Zitzmann, S. (2020). Integrating Out Nuisance Parameters for Computationally More Efficient Bayesian Estimation – An Illustration and Tutorial. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(3), 483–493. https://doi.org/10.1080/10705511.2019.1647432

Hecht, M., Walther, J.-K., Arnold, M., & Zitzmann, S. (2023). Finding the Optimal Number of Persons (N) and Time Points (T) for Maximal Power in Dynamic Longitudinal Models Given a Fixed Budget. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*(3), 535–551. https://doi.org/10.1080/10705511.2023.2230520

Hedeker, D., & Mermelstein, R. J. (2007). Mixed-effects regression models with heterogeneous variance: Analyzing ecological momentary assessment (EMA) data of smoking. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling Ecological and Contextual Effects in Longitudinal Studies of Human Development* (pp. 183–206). Mahwah, NJ: Erlbaum.

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2008). An Application of a Mixed-Effects Location Scale Model for Analysis of Ecological Momentary Assessment (EMA) Data. *Biometrics*, *64*(2), 627–634. https://doi.org/10.1111/j.1541-0420.2007.00924.x

Hedges, L. V., & Hedberg, E. C. (2013). Intraclass Correlations and Covariate Outcome Correlations for Planning Two- and Three-Level Cluster-Randomized Experiments in Education. *Evaluation Review*, *37*(6), 445–489. https://doi.org/10.1177/0193841X14529126

Herzog, W., Boomsma, A., & Reinecke, S. (2007). The Model-Size Effect on Traditional and Modified Tests of Covariance Structures. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 361–390. https://doi.org/10.1080/10705510701301602

Hill, P. W., & Rowe, K. J. (1996). Multilevel Modelling in School Effectiveness Research. *School Effectiveness and School Improvement*, *7*(1), 1–34. https://doi.org/10.1080/0924345960070101

Hill, W. G., & Thompson, R. (1978). Probabilities of Non-Positive Definite between-Group or Genetic Covariance Matrices. *Biometrics*, *34*(3), 429–439. https://doi.org/10.2307/2530605

Hodges, J. S. (1998). Some algebra and geometry for hierarchical models, applied to diagnostics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*(3), 497–536. https://doi.org/10.1111/1467-9868.00137

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10/cznwqw

Hoffman, L. (2007). Multilevel Models for Examining Individual Differences in Within-Person Variation and Covariation Over Time. *Multivariate Behavioral Research*, *42*(4), 609–629. https://doi.org/10.1080/00273170701710072

Hox, D., & McNeish, J. (2020). Small Samples in Multilevel Modeling. In R. van de Schoot & M. Milocevic (Eds.), *Small Sample Size Solutions* (pp. 215–225). Routledge.

Hox, J. J., & Maas, C. J. M. (2001). The Accuracy of Multilevel Structural Equation Modeling With Pseudobalanced Groups and Small Samples. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(2), 157–174. https://doi.org/10.1207/S15328007SEM0802_1

Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010a). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, *64*(2), 157–170. https://doi.org/10.1111/j.1467-9574.2009.00445.x

Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010b). The effect of Estimation Method and Sample Size in Multilevel Structural Equation Modeling. *Statistica Neerlandica*, *64*(2), 157–170. https://doi.org/10.1111/j.1467-9574.2009.00445.x

Hox, J. J., & Maas, C. J. (2002). Sample sizes for multilevel modeling. *Proceedings of the Fifth International Conference on Logic and Methodology, Opladen, RG*.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Hox, J. J., van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods*, *6*(2), 87–93. https://doi.org/10.18148/srm/2012.v6i2.5033

Huang, F. L., Wiedermann, W., & Zhang, B. (2022). Accounting for Heteroskedasticity Resulting from Between-Group Differences in Multilevel Models. *Multivariate Behavioral Research*, *58*(3), 637–657. https://doi.org/10.1080/00273171.2022.2077290

Huang, P.-H., Chen, H., & Weng, L.-J. (2017). A Penalized Likelihood Method for Structural Equation Modeling. *Psychometrika*, *82*(2), 329–354. https://doi.org/10/gbm65j

Huang, Y., & Bentler, P. M. (2015). Behavior of Asymptotically Distribution Free Test Statistics in Covariance Versus Correlation Structure Analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(4), 489–503. https://doi.org/10/gcz6zp

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *1*(1), 221–233.

Jackson, D. L. (2001). Sample Size and Number of Parameter Estimates in Maximum Likelihood Confirmatory Factor Analysis: A Monte Carlo Investigation. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(2), 205–223. https://doi.org/10.1207/S15328007SEM0802_3

Jackson, D. L. (2003). Revisiting Sample Size and Number of Parameter Estimates: Some Support for the N:q Hypothesis. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(1), 128–141. https://doi.org/10.1207/S15328007SEM1001_6

Jackson, D. L., Voth, J., & Frey, M. P. (2013). A Note on Sample Size and Solution Propriety for Confirmatory Factor Analytic Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(1), 86–97. https://doi.org/10.1080/10705511.2013.742388

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 555–566. https://doi.org/10/gcmfh8

James, W., & Stein, C. (1961). Estimation with Quadratic Loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Scientific Software International.

Julian, M. (2001). The Consequences of Ignoring Multilevel Data Structures in Nonhierarchical Covariance Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(3), 325–352. https://doi.org/10.1207/S15328007SEM0803_1

Jung, S., & Takane, Y. (2007). Regularized Common Factor Analysis. *New Trends in Psychometrics*, 141–149.

Kamada, A., & Kano, Y. (2012). Statistical inference in structural equation modeling with a near singular covariance matrix. *2nd Institute of Mathematical Statistics Asia Pacific Rim Meeting, Tsukuba, Japan*.

Kamada, A., Yanagihara, H., Wakaki, H., & Fukui, K. (2014). Selecting a shrinkage parameter in structural equation modeling with a near singular covariance matrix by the GIC minimization method. *Hiroshima Mathematical Journal*, *44*(3), 315–326. https://doi.org/10/gpgn8b

Kelley, C. T. (1995). *Iterative Methods for Linear and Nonlinear Equations*. Society for Industrial and Applied Mathematics.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical Practices of Educational Researchers: An Analysis of their ANOVA, MANOVA, and ANCOVA Analyses. *Review of Educational Research*, *68*(3), 350–386. https://doi.org/10.3102/00346543068003350

Korendijk, E. J., Maas, C. J., Moerbeek, M., & Van der Heijden, P. G. (2008). The Influence of Misspecification of the Heteroscedasticity on Multilevel Regression Parameter and Standard Error Estimates. *Methodology*, *4*(2), 67–72. https://doi.org/10.1027/1614-2241.4.2.67

Kreft, I. G., & De Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

Kreft, I. G., & Yoon, B. (1994). Are multilevel techniques necessary? An attempt at demystification. *Annual Meeting of the American Educational Research Association, New Orleans (LA)*.

Laird, N., Lange, N., & Stram, D. (1987). Maximum Likelihood Computations with Repeated Measures: Application of the EM Algorithm. *Journal of the American Statistical Association*, *82*(397), 97–105. https://doi.org/10.1080/01621459.1987.10478395

Lange, K., Chambers, J., & Eddy, W. (1999). *Numerical analysis for statisticians* (Vol. 2). Springer.

Leckie, G., French, R., Charlton, C., & Browne, W. (2014). Modeling Heterogeneous Variance–Covariance Components in Two-Level Models. *Journal of Educational and Behavioral Statistics*, *39*(5), 307–332. https://doi.org/10.3102/1076998614546494

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4

Ledoit, O., & Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, *40*(2). https://doi.org/10.1214/12-AOS989

Ledoit, O., & Wolf, M. (2018). Optimal estimation of a large-dimensional covariance matrix under Stein's loss. *Bernoulli*, *24*, 3791–3832. https://doi.org/10.3150/17-BEJ979

Ledoit, O., & Wolf, M. (2022). The Power of (Non-)Linear Shrinking: A Review and Guide to Covariance Matrix Estimation. *Journal of Financial Econometrics*, *20*(1), 187–218. https://doi.org/10.1093/jjfinec/nbaa007

Liang, X., & Jacobucci, R. (2020). Regularized Structural Equation Modeling to Detect Measurement Bias: Evaluation of Lasso, Adaptive Lasso, and Elastic Net. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(5), 722–734. https://doi.org/10.1080/10705511.2019.1693273

Lorenzo-Seva, U., & Ferrando, P. J. (2021). Not Positive Definite Correlation Matrices in Exploratory Item Factor Analysis: Causes, Consequences and a Proposed Solution. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(1), 138–147. https://doi.org/10.1080/10705511.2020.1735393

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological Methods*, *16*(4), 444–467. https://doi.org/10.1037/a0024376

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*(3), 203–229. https://doi.org/10/c8qdh2

Luo, W., Li, H., Baek, E., Chen, S., Lam, K. H., & Semma, B. (2021). Reporting Practice in Multilevel Modeling: A Revisit After 10 Years. *Review of Educational Research*, *91*(3), 311–355. https://doi.org/10.3102/0034654321991229

Maas, C. J. M., & Hox, J. J. (2004a). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*(2), 127–137. https://doi.org/10.1046/j.0039-0402.2003.00252.x

Maas, C. J. M., & Hox, J. J. (2004b). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*, *46*(3), 427–440. https://doi.org/10.1016/j.csda.2003.08.006

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

MacCallum, R. C., Kim, C., Malarkey, W. B., & Kiecolt-Glaser, J. K. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research*, *32*(3), 215–253. https://doi.org/10.1207/s15327906mbr3203_1

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

Marcoulides, K. M., Yuan, K.-H., & Deng, L. (2023). Structural Equation Modeling with Small Samples and Many Variables. In R. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (Second Edition, pp. 525–542). Guilford Press.

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child development*, *58*(1), 110–133. https://doi.org/10.2307/1130295

McCoach, D. B., Rifenbark, G. G., Newton, S. D., Li, X., Kooken, J., Yomtov, D., Gambino, A. J., & Bellara, A. (2018). Does the Package Matter? A Comparison of Five Common Multilevel Modeling Software Packages. *Journal of Educational and Behavioral Statistics*, *43*(5), 594–627. https://doi.org/10.3102/1076998618776348

McDonald, R. P. (1993). A general model for two-level data with responses missing at random. *Psychometrika*, *58*(4), 575–585. https://doi.org/10.1007/BF02294828

McDonald, R. P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of mathematical and statistical psychology*, *42*(2), 215–232. https://doi.org/10.1111/j.2044-8317.1989.tb00911.x

McNabb, C. B., & Murayama, K. (2021). Unnecessary reliance on multilevel modelling to analyse nested data in neuroscience: When a traditional summary-statistics approach suffices. *Current Research in Neurobiology*, *2*, 100024. https://doi.org/10.1016/j.crneur.2021.100024

McNeish, D. (2021). Specifying Location-Scale Models for Heterogeneous Variances as Multilevel SEMs. *Organizational Research Methods*, *24*(3), 630–653. https://doi.org/10.1177/1094428120913083

McNeish, D., & Stapleton, L. M. (2016). Modeling Clustered Data with Very Few Clusters. *Multivariate Behavioral Research*, *51*(4), 495–518. https://doi.org/10.1080/00273171.2016.1167008

McNeish, D. M. (2015). Using Lasso for Predictor Selection and to Assuage Overfitting: A Method Long Overlooked in Behavioral Sciences. *Multivariate Behavioral Research*, *50*(5), 471–484. https://doi.org/10.1080/00273171.2015.1036965

McNeish, D. M., & Stapleton, L. M. (2016). The Effect of Small Sample Size on Two-Level Model Estimates: A Review and Illustration. *Educational Psychology Review*, *28*(2), 295–314. https://doi.org/10.1007/s10648-014-9287-x

McQuitty, S. (1997). Effects of employing ridge regression in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *4*(3), 244–252. https://doi.org/10.1080/10705519709540074

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*(3), 259–284. https://doi.org/10.1037/1082-989X.10.3.259

Meredith, W., & Tisak, J. (1990). Latent Curve Analysis. *55*(1), 107–122.

Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *3*(1), 45–58.

Mok, M. (1995). Sample Size Requirements for 2-level Designs in Educational Research. *Multilevel modelling newsletter*, *7*(2), 11–15.

Monk, D. H. (1992). Education productivity research: An update and assessment of its role in education finance reform. *Educational Evaluation and Policy Analysis*, *14*(4), 307–332. https://doi.org/10.3102/01623737014004

Muthen, B. O., & Satorra, A. (1995). Complex Sample Data in Structural Equation Modeling. *Sociological Methodology*, *25*, 267–316. https://doi.org/10.2307/271070

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585. https://doi.org/10.1007/BF02296397

Muthén, B. O. (1990). *Mean and covariance structure analysis of hierarchical data* [Unpublished manuscript].

Muthén, B. O. (1994). Multilevel Covariance Structure Analysis. *Sociological Methods & Research*, *22*(3), 376–398. https://doi.org/10.1177/0049124194022003006

Muthén, B. O., Khoo, S.-T., & Gustafsson, J.-E. (1997). *Multilevel latent variable modeling in multiple populations* (Technical Report). Graduate School of Education & Information Studies. University of California, Los Angeles.

Muthén, B. O., & Satorra, A. (1989). Multilevel Aspects of Varying Parameters in Structural Models. In Bock (Ed.), *Multilevel Analysis of Educational Data* (pp. 87–99). San Diego: Academic Press.

Muthén, L. K., & Muthén, B. O. (2002). How to Use a Monte Carlo Study to Decide on Sample Size and Determine Power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8

Orth, U., Clark, D. A., Donnellan, M. B., & Robins, R. W. (2021). Testing prospective effects in longitudinal research: Comparing seven competing cross-lagged models. *Journal of Personality and Social Psychology*, *120*(4), 1013–1034. https://doi.org/10.1037/pspp0000358

Orzek, J. H., & Voelkle, M. C. (2023). Regularized continuous time structural equation models: A network perspective. *Psychological Methods*, 1–35. https://doi.org/10.1037/met0000550

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*(1), 85–112. https://doi.org/10.1016/j.jsp.2009.09.002

Pourahmadi, M. (2013). *High-dimensional covariance estimation* (Vol. 882). John Wiley & Sons.

Preacher, K. J. (2011). Multilevel SEM Strategies for Evaluating Mediation in Three-Level Data. *Multivariate Behavioral Research*, *46*(4), 691–731. https://doi.org/10.1080/00273171.2011.589280

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2016). Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychological Methods*, *21*(2), 189–205. https://doi.org/10.1037/met0000052

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata*. College Station, TX: Stata Press.

Raudenbush, S. W. (2001). Toward a coherent framework for comparing trajectories of individual change. In *New methods for the analysis of change* (pp. 35–64). American Psychological Association. https://doi.org/10.1037/10409-002

Raudenbush, S. W., & Bryk, A. S. (1987). Examining Correlates of Diversity. *Journal of Educational Statistics*, *12*(3), 241–269. https://doi.org/10.3102/10769986012003241

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (Vol. 1). SAGE.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*(2), 185–205. https://doi.org/10.1037/1082-989X.8.2.185

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A Growth Curve Approach to the Measurement of Change. *Psychological Bulletin*, *92*(3), 726–748. https://doi.org/10.1037/0033-2909.92.3.726

Rogosa, D. R., & Willett, J. B. (1985). Understanding Correlates of Change by Modeling Individual Differences in Growth. *Psychometrika*, *50*(2), 203–228. https://doi.org/10.1007/BF02294247

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

Rovine, M. J., & Molenaar, P. C. (2000). A Structural Modeling Approach to a Multilevel Random Coefficients Model. *Multivariate Behavioral Research*, *35*(1), 51–88. https://doi.org/10.1207/S15327906MBR3501_3

Ryu, E. (2014). Factorial invariance in multilevel confirmatory factor analysis. *British Journal of Mathematical and Statistical Psychology*, *67*(1), 172–194. https://doi.org/10.1111/bmsp.12014

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, *50*(1), 83–90. https://doi.org/10.1007/BF02294150

Schäfer, J., & Strimmer, K. (2005). A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*(1), 1–30. https://doi.org/10.2202/1544-6115.1175

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance Components*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316856

Shi, D., Lee, T., & Terry, R. A. (2018). Revisiting the Model Size Effect in Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(1), 21–40. https://doi.org/10.1080/10705511.2017.1369088

Shin, Y., & Raudenbush, S. W. (2010). A Latent Cluster-Mean Approach to the Contextual Effects Model With Missing Data. *Journal of Educational and Behavioral Statistics*, *35*(1), 26–53. https://doi.org/10/c63vdm

Singer, H. (2010). SEM Modeling with Singular Moment Matrices Part I: ML-Estimation of Time Series. *The Journal of Mathematical Sociology*, *34*(4), 301–320. https://doi.org/10.1080/0022250X.2010.509524

Singer, J. D. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2. ed). SAGE.

Steele, F. (2008). Multilevel Models for Longitudinal Data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *171*(1), 5–19. https://doi.org/10.1111/j.1467-985X.2007.00509.x

Stegmueller, D. (2013). How Many Countries for Multilevel Modeling? A Comparison of Frequentist and Bayesian Approaches. *American Journal of Political Science*, *57*(3), 748–761. https://doi.org/10.1111/ajps.12001

Stein, C. (1956). Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. In J. Neyman (Ed.). University of California Press.

Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Annals of the Institute of Statistical Mathematics*, *16*(1), 155–160. https://doi.org/10.1007/BF02868569

Stein, C. (1975). Estimation of a covariance matrix Rietz Lecture. *39th Annual Meeting IMS, Atlanta, Georgia*.

Sun, Y. (2015). *Regularization in High-dimensional Statistics* [Doctoral dissertation, Stanford University].

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R., & Hastie, T. (2007). Margin Trees for High-dimensional Classification. *Journal of Machine Learning Research*, *8*(3), 637–652.

Tikhonov, A. N. (1943). On the stability of inverse problems. *Doklady Akademii Nauk SSSR*, *39*, 195–198.

Touloumis, A. (2015). Nonparametric Stein-type shrinkage covariance matrix estimators in high-dimensional settings. *Computational Statistics & Data Analysis*, *83*, 251–261. https://doi.org/10.1016/j.csda.2014.10.018

Townsend, Z., Buckley, J., Harada, M., & Scott, M. (2013). The Choice between Fixed and Random Effects. In *The SAGE Handbook of Multilevel Modeling* (pp. 73–88). SAGE Publications Ltd. https://doi.org/10.4135/9781446247600.n5

Trendafilov, N. T., & Unkel, S. (2011). Exploratory Factor Analysis of Data Matrices With More Variables Than Observations. *Journal of Computational and Graphical Statistics*, *20*(4), 874–891. https://doi.org/10.1198/jcgs.2011.09211

Tsukuma, H. (2016). Estimation of a high-dimensional covariance matrix with the Stein loss. *Journal of Multivariate Analysis*, *148*, 1–17. https://doi.org/10.1016/j.jmva.2016.02.012

Ulitzsch, E., Lüdtke, O., & Robitzsch, A. (2023). Alleviating estimation problems in small sample structural equation modeling—A comparison of constrained maximum likelihood, Bayesian estimation, and fixed reliability approaches. *Psychological Methods*, *28*(3), 527–557. https://doi.org/10.1037/met0000435.supp

Unkel, S., & Trendafilov, N. T. (2010). Simultaneous Parameter Estimation in Exploratory Factor Analysis: An Expository Review. *International Statistical Review*, *78*(3), 363–382. https://doi.org/10.1111/j.1751-5823.2010.00120.x

Van Hoa, T. (1985). The inadmissibility of the Stein estimator in normal multiple regression equations. *Economics Letters*, *19*(1), 39–42. https://doi.org/10.1016/0165-1765(85)90099-0

Van Montfort, K., Oud, J. H., & Voelkle, M. C. (2018). *Continuous time modeling in the behavioral and related sciences*. Springer.

Vandenberg, R. J., & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Voelkle, M. C., Oud, J. H. L., von Oertzen, T., & Lindenberger, U. (2012). Maximum Likelihood Dynamic Factor Modeling for Arbitrary N and T Using SEM. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(3), 329–350. https://doi.org/10.1080/10705511.2012.687656

Walther, J.-K., Hecht, M., Nagengast, B., & Zitzmann, S. (2024). To Be Long or To Be Wide: How Data Format Influences Convergence and Estimation Accuracy in Multilevel Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*(5), 759–774. https://doi.org/10.1080/10705511.2024.2320050

Walther, J.-K., Hecht, M., & Zitzmann, S. (2024). Shrinking Small Sample Problems in Multilevel Structural Equation Modeling via Regularization of the Sample Covariance Matrix. *Structural Equation Modeling: A Multidisciplinary Journal*, *32*(1), 45–65. https://doi.org/10.1080/10705511.2024.2380919

West, B. T., Welch, K. B., & Galecki, A. T. (2022). *Linear mixed models: A practical guide using statistical software*. CRC Press.

West, K. D. (1997). Another heteroskedasticity- and autocorrelation-consistent covariance matrix estimator. *Journal of Econometrics*, *76*(1–2), 171–191. https://doi.org/10.1016/0304-4076(95)01788-7

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, *48*(4), 817–838. https://doi.org/10.2307/1912934

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, *50*(1), 1–25. https://doi.org/10.2307/1912526

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin*, *116*(2), 363–381. https://doi.org/10.1037/0033-2909.116.2.363

Williams, D. R., & Rodriguez, J. E. (2022). Why overfitting is not (usually) a problem in partial correlation networks. *Psychological Methods*, *27*(5), 822–840. https://doi.org/http://dx.doi.org/10.1037/met0000437

Wothke, W. (1993). Nonpositive Definite Matrices in Structural Modeling. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 256–293). Sage Publications.

Wu, W. B. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, *90*(4), 831–844. https://doi.org/10.1093/biomet/90.4.831

Yang, M., Jiang, G., & Yuan, K.-H. (2018). The Performance of Ten Modified Rescaled Statistics as the Number of Variables Increases. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 414–438. https://doi.org/10.1080/10705511.2017.1389612

Yang, R., & Berger, J. O. (1994). Estimation of a Covariance Matrix Using the Reference Prior. *The Annals of Statistics*, *22*(3), 1195–1211. https://doi.org/10.1214/aos/1176325625

Yuan, K.-H., & Bentler, P. M. (2017). Improving the convergence rate and speed of Fisher-scoring algorithm: Ridge and anti-ridge methods in structural equation modeling. *Annals of the Institute of Statistical Mathematics*, *69*(3), 571–597. https://doi.org/10/gpgn83

Yuan, K.-H., & Chan, W. (2008). Structural equation modeling with near singular covariance matrices. *Computational Statistics & Data Analysis*, *52*(10), 4842–4858. https://doi.org/10.1016/j.csda.2008.03.030

Yuan, K.-H., Wu, R., & Bentler, P. M. (2011). Ridge structural equation modelling with correlation matrices for ordinal and continuous data: Ridge SEM with correlation matrices. *British Journal of Mathematical and Statistical Psychology*, *64*(1), 107–133. https://doi.org/10/cwd74t

Zhang, Q. (2022). High-Dimensional Mediation Analysis with Applications to Causal Gene Identification. *Statistics in Biosciences*, *14*(3), 432–451. https://doi.org/10.1007/s12561-021-09328-0

Zhao, Y., Lindquist, M. A., & Caffo, B. S. (2020). Sparse principal component based high-dimensional mediation analysis. *Computational Statistics & Data Analysis*, *142*, 106835. https://doi.org/10.1016/j.csda.2019.106835

Zitzmann, S. (2018). A Computationally More Efficient and More Accurate Stepwise Approach for Correcting for Sampling Error and Measurement Error. *Multivariate Behavioral Research*, *53*(5), 612–632. https://doi.org/10/gpgn86

Zitzmann, S., Lüdtke, O., & Robitzsch, A. (2015). A Bayesian Approach to More Stable Estimates of Group-Level Effects in Contextual Studies. *Multivariate Behavioral Research*, *50*(6), 688–705. https://doi.org/10/gg5fg2

Zitzmann, S., Lüdtke, O., Robitzsch, A., & Marsh, H. W. (2016). A Bayesian Approach for Estimating Multilevel Latent Contextual Models. *Structural Equation Modeling: A*

Multidisciplinary Journal, 23(5), 661–679. https://doi.org/10.1080/10705511.2016.1207179

Zitzmann, S., Nagengast, B., Hübner, N., & Hecht, M. (2024). A simple solution to heteroscedasticity in multilevel nonlinear structural equation modeling [Manuscript submitted for publication].

Zitzmann, S., Wagner, W., Hecht, M., Helm, C., Fischer, C., Bardach, L., & Göllner, R. (2022). How Many Classes and Students Should Ideally be Sampled When Assessing the Role of Classroom Climate via Student Ratings on a Limited Budget? An Optimal Design Perspective. *Educational Psychology Review, 34*, 511–536. https://doi.org/10.1007/s10648-021-09635-4

Zitzmann, S., Walther, J.-K., Hecht, M., & Nagengast, B. (2022). What Is the Maximum Likelihood Estimate When the Initial Solution to the Optimization Problem Is Inadmissible? The Case of Negatively Estimated Variances. *Psych, 4*(3), 343–356. https://doi.org/10.3390/psych4030029

Zitzmann, S., Weirich, S., & Hecht, M. (2023). Accurate Standard Errors in Multilevel Modeling with Heteroscedasticity: A Computationally More Efficient Jackknife Technique. *Psych, 5*(3), 757–769. https://doi.org/10.3390/psych5030049

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x