STRUCTURAL EQUATION MODELING
2025, VOL. 32, NO. 5
ISSN: 1070-5511 | E-ISSN: 1532-8007

OPEN ACCESS

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,

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).

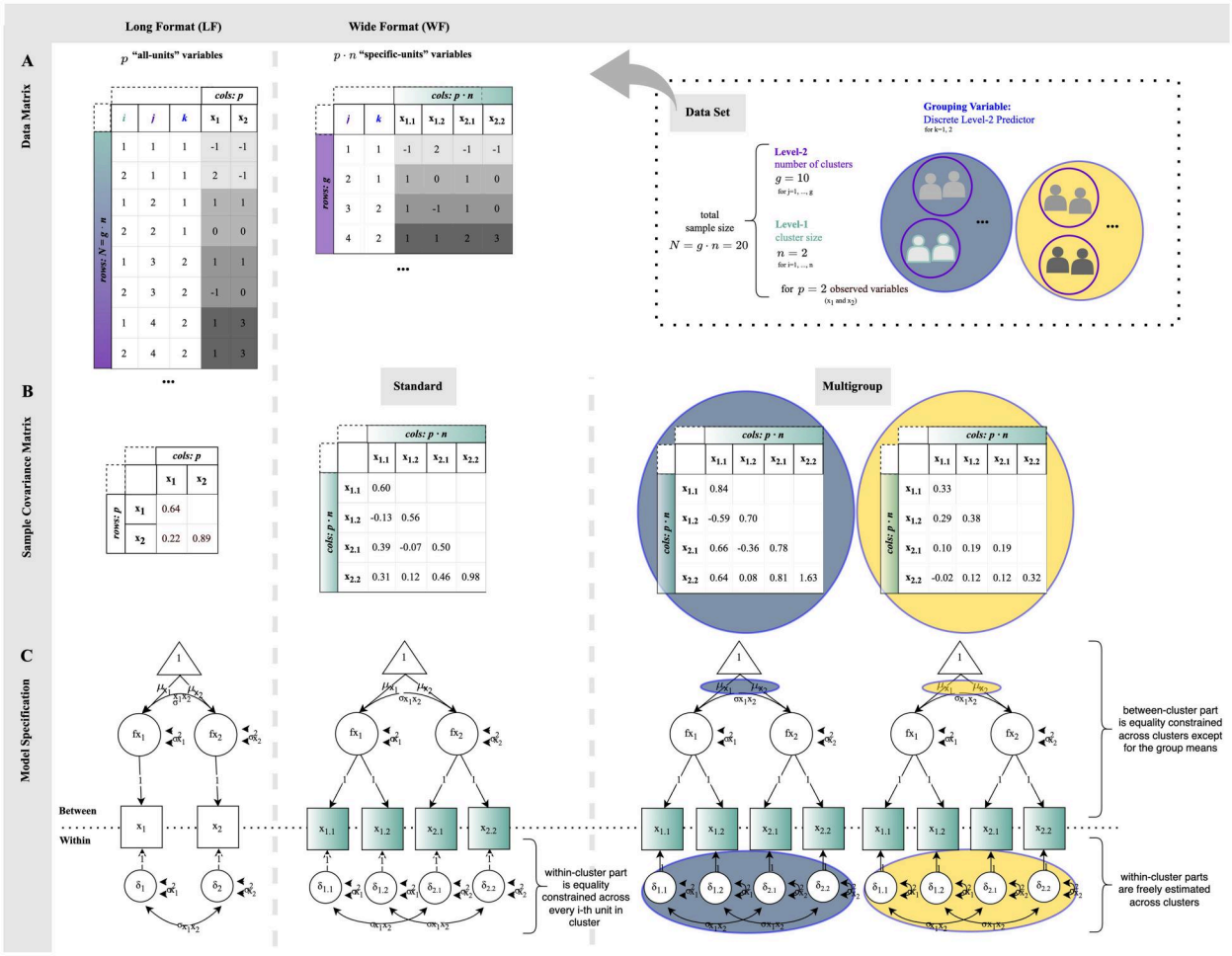

Figure 1. The LF, WF, and WF multigroup approaches. *Data set*: the data collected in a given setting. *Data Matrix*: the data set in matrix form, where columns refer to observed variables and rows to observed units. *Data Format*: one of two possible formats of the data matrix, long format (LF) or wide 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;

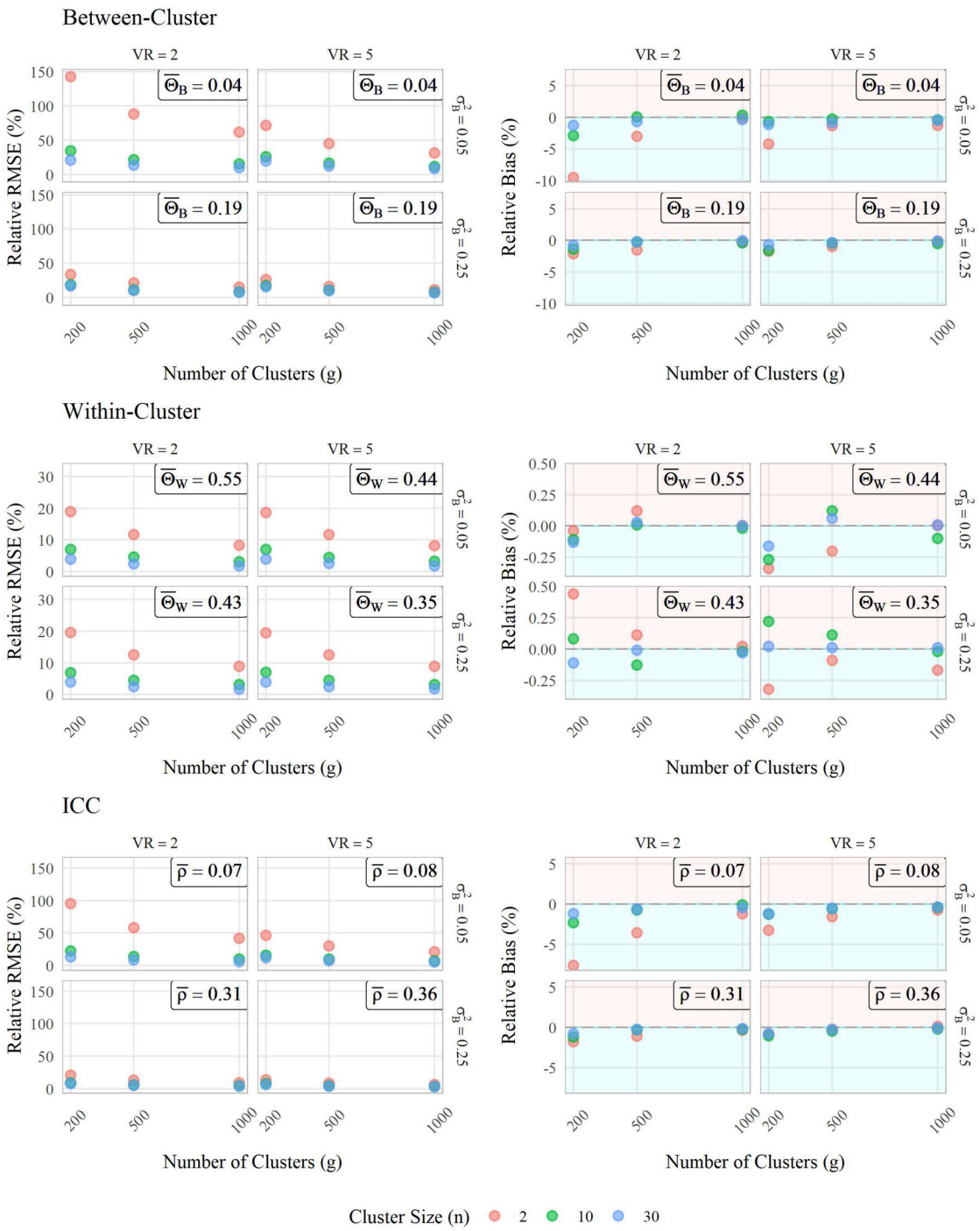

Figure 2. Estimation accuracy of between-cluster, within-cluster, and ICC parameter estimates. VR = variance ratio. Only (co)variance parameter estimates are considered. In a bivariate two-level intercept-only model with heterogeneous within-cluster (co)variances for two groups, this comprises three 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

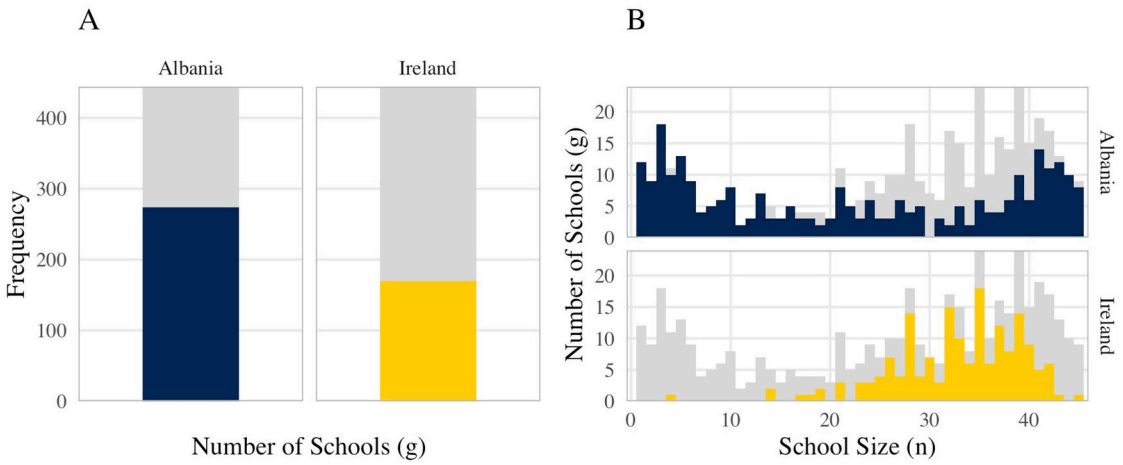

Figure 3. Number of schools and school sizes by country. Number of Schools = Clusters (i.e. Level-2 units); school size = cluster size (i.e. level-1 units students).

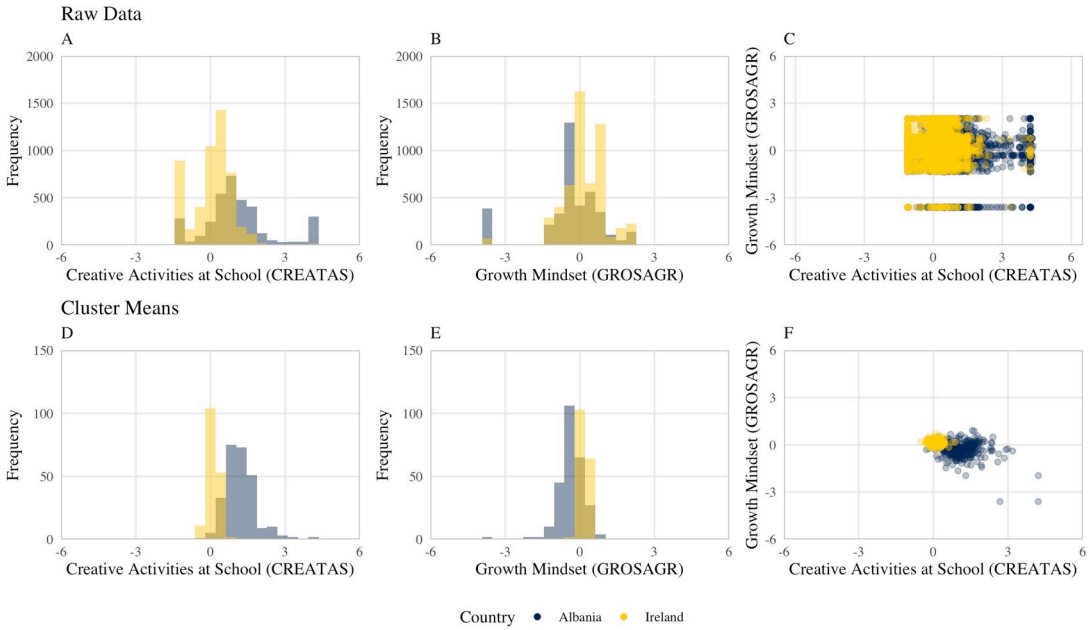

Figure 4. The distributions of raw data and cluster means. $N_{CREATAS(All)} = 8,449$ (28% missings) with $N_{CREATAS(Albania)} = 3,398$ (23.5% of all missings and 44.5% of missings in Albania) and $N_{CREATAS(Ireland)} = 5,051$ (4.5% of overall missings and 10% of missings in Ireland); $N_{GROSAGR(All)} = 9,319$ (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

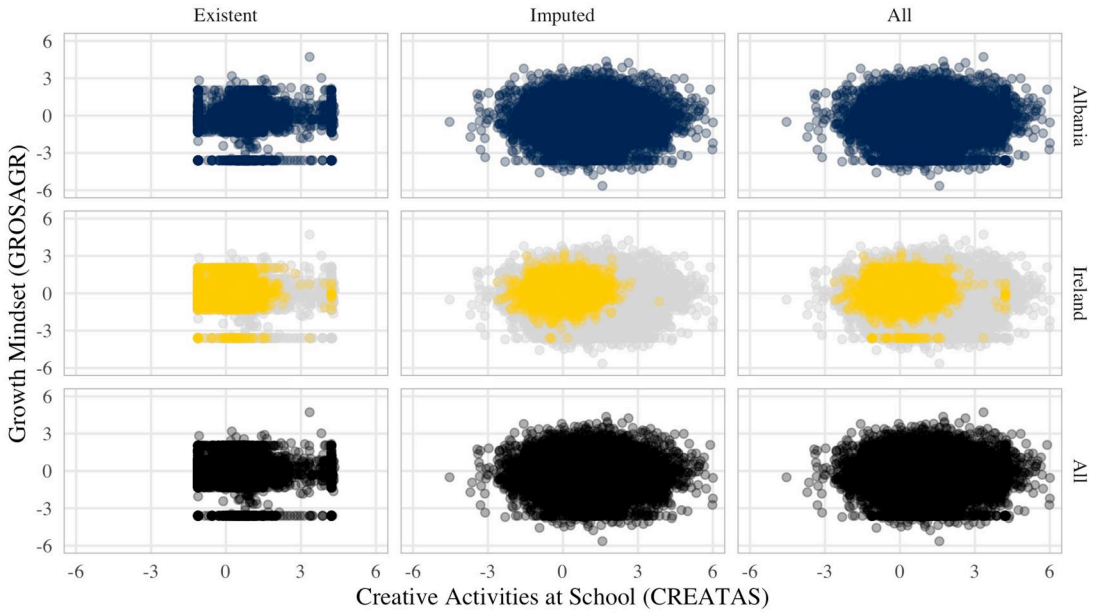

Figure 5. Existent and imputed data. "All (data)" refers to $N = 19,980$ for each variable where $N_{Albania} = 12,330$ and $N_{Ireland} = 7,650$; "Existent (data)": $N_{CREATAS(Albania)} = 3,398$ and $N_{CREATAS(Ireland)} = 5,051$, $N_{GROSAGR(Albania)} = 3,870$ and $N_{GROSAGR(Ireland)} = 5,449$ (but 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

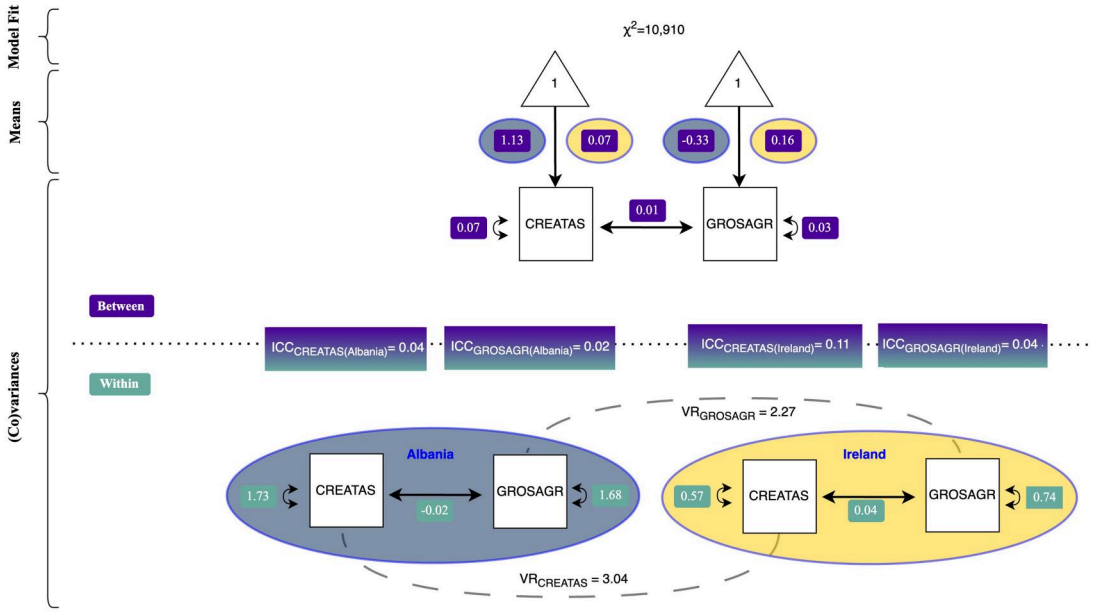

Figure 6. Models with heterogeneous and homogeneous variances. The figure was created manually with the free software draw.io (https://www.drawio.com/).

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.

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

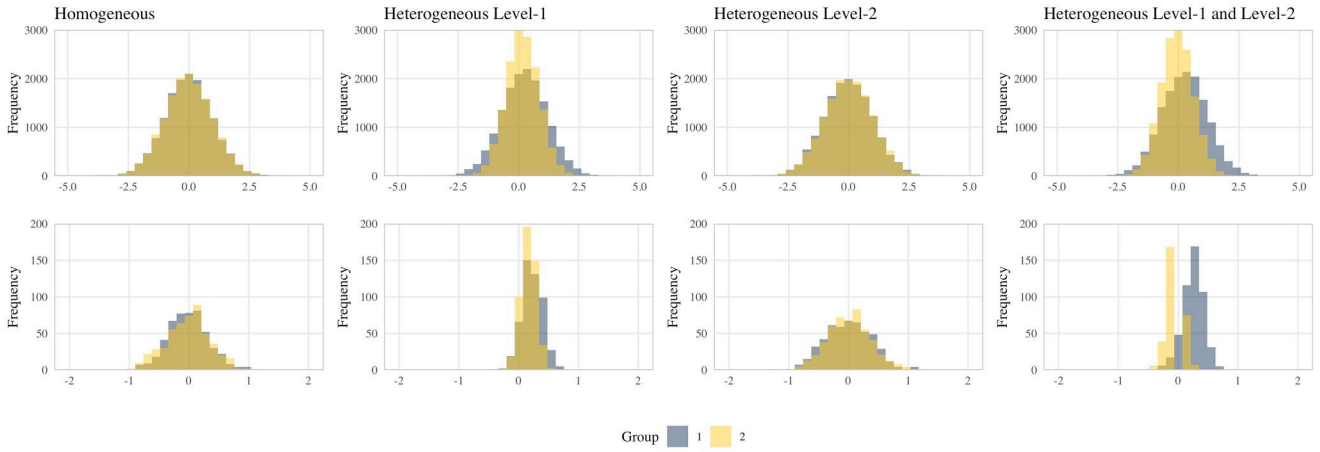

Figure A1. The distributions of raw data and cluster means under homogeneous and heterogeneous conditions. The upper row shows raw data and the lower row cluster means of one observed variable. The simulated heterogeneous conditions have been adapted from the PISA data from the empirical illustration 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)

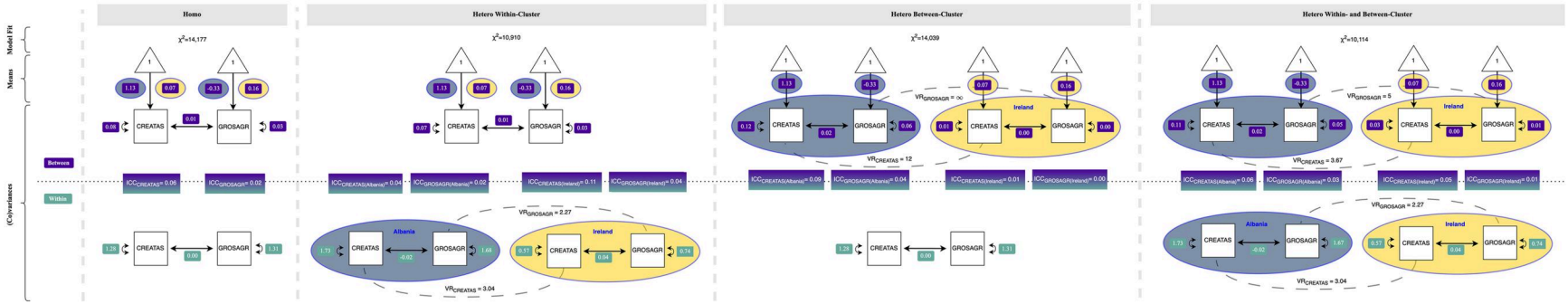

Figure A2. Four types of models with homogeneity and heterogeneity at the between- and within-cluster levels. The code to estimate all parameters for all models can be found in Appendix B. These are the estimates from the WFmultigroup approach, but those of the "genuine" ML MG SEM are very similar.

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),]
```

```
158
159 j <- c()
160 country <- c()
161 for (i in 1:nrow(PISA_short)){
162   tmp_j <- PISA_short$CNTSCHID[i]
163   if (i==1){
164     country <- append(country, PISA_short$CNT[i])
165     j <- append(j, tmp_j)
166   } else {
167     if (tmp_j > tail(j, n=1)){
168       country <- append(country, PISA_short$CNT[i])
169       j <- append(j, tmp_j )
170     }
171   }
172 }
173
174 PISA_short_cluster_means <- data.frame(j=1:444, country=country, CREATAS=
    CREATAS_cluster_means$x, GROSAGR=GROSAGR_cluster_means$x)
175
176 # univariate
177 d <- ggplot(PISA_short_cluster_means, aes(x=CREATAS, fill=country)) +
178   geom_histogram(show.legend = FALSE, position = "identity", alpha=0.5) +
179   scale_x_continuous(name="Creative Activities at School (CREATAS)", expand
     =c(0, 0),
180                      limits=c(-6, 6)) +
181   scale_y_continuous(name="Frequency", limits=c(0, 150), expand=c(0, 0),) +
182   scale_fill_manual(name="Country", values=c("#002654", "#ffce00")) +
183   theme_minimal() +
184   theme(text = element_text(family="serif"), panel.grid.minor = element_
     blank(),
185         panel.border = element_rect(color = "grey",  fill = NA, linewidth =
      0.5)) +
186   guides(colour = guide_legend(override.aes = list(alpha = 1))) +
187   labs(title="Cluster Means", subtitle="D")
188
189
190 e <- ggplot(PISA_short_cluster_means, aes(x=GROSAGR, fill=country)) +
191   geom_histogram(show.legend = FALSE, position = "identity", alpha=0.5) +
192   scale_x_continuous(name="Growth Mindset (GROSAGR)", expand=c(0,0),
193                      limits=c(-6, 6)) +
194   scale_y_continuous(name="Frequency", limits=c(0, 150), expand=c(0, 0),) +
195   scale_fill_manual(name="Country", values=c("#002654", "#ffce00")) +
196   theme_minimal() +
197   theme(text = element_text(family="serif"), panel.grid.minor = element_
     blank(),
198         panel.border = element_rect(color = "grey",  fill = NA, linewidth =
      0.5)) +
199   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).
```

```r

## (a) descriptive
# by figure with percentages
md.pattern(PISA_short, rotate.names = TRUE) # note this function is from package mice but mcar_test is from package naniar

# rows: missing patterns
# numbers to left: cases for each missing pattern
# number to right: number of missings in missing pattern
# numbers at bottom: number of missing cases for each variable (column) --> absolute numbers we got in figure before

# 4 patterns
# most often all variables existent (1. row),
8137 / (8137 + 1182 + 312 + 2067) # approx. 70% cases without any missings, thus, 30% of cases with at least one missing!
# then one missing in CREATAS (2. row),
(1182) / (8137 + 1182 + 312 + 2067) # approx 10% of only missing CREATAS
# then missings in CREATAS and GROSAGR (4. row),
(2067) / (8137 + 1182 + 312 + 2067) # approx 18% of missing CREATAS and GROSAGR
# Note 10% + 18% add up to the 28% missing cases reported for CREATAS before
# then one missing in GROSAGR (3. row)
(312) / (8137 + 1182 + 312 + 2067) # approx 3% of only missing GROSAGR


## (b) inferential
# by using Little's (1988) test that compares patterns of missingness
# H0: MCAR
# H1: not MCAR
# Note CNT and CNTSCHID are perfectly correlated and can thus not be used in the same test bc of multicollinearity (i.e., singularity)
# we drop CNT
mcar_test(PISA_short[, c("CNTSCHID", "CREATAS", "GROSAGR")])
# test is significant, thus evidence that MCAR does not hold

## explore MAR assumption

# create missing data indicators (missing=1, existent=0)
PISA_short$missing_CREATAS <- ifelse(is.na(PISA_short$CREATAS), yes=1, no=0)
PISA_short$missing_GROSAGR <- ifelse(is.na(PISA_short$GROSAGR), yes=1, no=0)

## (a) descriptive
# by correlation table
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

```

```r

##### (6) Reformating I: Balanced Cluster Sizes in LF
# necessary for imputing unbalanced data, and to reformat to WF later

## create new data frame with balanced number of students
PISA_short_balanced <- data.frame(
  j = rep(1:g, each=n_max),
  i = rep(1:n_max, times=g),
  CNTSCHID = rep(NA, n_max*g) , # incomplete
  CNTSTUID = rep(NA, n_max*g), # incomplete
  CNT = rep(NA, n_max*g),
  CREATAS = rep(NA, n_max*g),
  GROSAGR = rep(NA, n_max*g),
  missing_CREATAS = rep(1, n_max*g),
  missing_GROSAGR = rep(1, n_max*g)
)

# sort data by school
PISA_short <- PISA_short[with(PISA_short, order(CNTSCHID)), ]

# fill in existing data
for (j in 1:g) {
  school <- unique(PISA_short$CNTSCHID)[j]
  students <- filter(PISA_short, CNTSCHID == school)$CNTSTUID
  nSchool <- length(students)
  PISA_short_balanced$CNTSCHID[((j - 1) * n_max + 1):((j - 1) * n_max +
    nSchool)] <- school
  PISA_short_balanced$CNTSTUID[((j - 1) * n_max + 1):((j - 1) * n_max +
    nSchool)] <- students
  PISA_short_balanced$CNT[((j - 1) * n_max + 1):((j - 1) * n_max + n_max)]
    <- unique(PISA_short$CNT[which(PISA_short$CNTSCHID == school)])
  PISA_short_balanced$CREATAS[((j - 1) * n_max + 1):((j - 1) * n_max +
    nSchool)] <- PISA_short$CREATAS[which(PISA_short$CNTSCHID == school)]
  PISA_short_balanced$GROSAGR[((j - 1) * n_max + 1):((j - 1) * n_max +
    nSchool)] <- PISA_short$GROSAGR[which(PISA_short$CNTSCHID == school)]
  PISA_short_balanced$missing_CREATAS[((j - 1) * n_max + 1):((j - 1) * n_
    max + nSchool)] <- PISA_short$missing_CREATAS[which(PISA_short$CNTSCHID
    == school)]
  PISA_short_balanced$missing_GROSAGR[((j - 1) * n_max + 1):((j - 1) * n_
    max + nSchool)] <- PISA_short$missing_GROSAGR[which(PISA_short$CNTSCHID
    == school)]
}

# Now N=n_max*g = 19980 level-1 units.
# Final subsample per country: g*n_max = N
table(nData$country)*n_max

```

```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
```

```r

# combine imputed data sets of both groups (i.e., countries)
PISA_short_balanced_imp <- rbind(complete(imp_Albania), complete(imp_Ireland))

# plot imputed data (Fig.5)
ggplot(PISA_short_balanced_imp, aes(x=CREATAS, y=GROSAGR, col=CNT)) +
  geom_point(data = transform(PISA_short_balanced_imp, CNT = NULL), col="grey85", alpha=0.5) +
  geom_point(show.legend = FALSE, alpha=0.3) +
  facet_grid(CNT ~ missing_CREATAS, margins=TRUE, # adds an additional facet for all levels combined
             labeller=as_labeller(c('0'="Existent", '1'="Imputed", '(all)'="All", 'Albania'="Albania", 'Ireland'="Ireland"))) +
  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(0.05,0.05), limits=c(-6, 6)) +
  scale_color_manual(values=c("#002654", "#ffce00", "black")) +
  theme_minimal() +
  theme(text = element_text(family="serif"), panel.grid.minor = element_blank(),
        panel.border = element_rect(color = "grey",  fill = NA, linewidth = 0.5))



##### (8) Reformating II: Format LF to WF
# where each row corresponds to a school
PISA_short_balanced_imp_WF <- select(PISA_short_balanced_imp, -c("CNTSCHID", "CNTSTUID", "missing_CREATAS", "missing_GROSAGR")) # drop variables, otherwise formating faulty
PISA_short_balanced_imp_WF <- pivot_wider(PISA_short_balanced_imp_WF, names_from = i, values_from = c("CREATAS", "GROSAGR"), names_sep = ".")



##### (9) Model Estimation

## Homogeneity/Heterogeneity is set differently for both levels:
# Level-1: in model syntax (by using same or different parameter labels)
# Level-2: with function parameter   group  .equal = c("lv.variances", "lv.covariances")    (by setting it or leaving it out)
# Thus, the model syntax below is the same for all models.
# (Note that the variances at each level in the homogeneous models equal the pooled variances in the heterogeneous models.)

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)


```

```r
### 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
```