Published in final edited form as:

Multivariate Behav Res. 2020 ; 55(1): 120–136. doi:10.1080/00273171.2019.1627513.

Testing variance components in linear mixed modeling using permutation

Dr Han Du, Ph.D,

University of California, Los Angeles, Department of Psychology, Franz Hall, 502 Portola Plaza, Los Angeles, 90095 United States

Dr Lijuan Wang

University of Notre Dame, Notre Dame, United States

Abstract

Inference of variance components in linear mixed modeling (LMM) provides evidence of heterogeneity between individuals or clusters. When only nonnegative variances are allowed, there is a boundary (i.e., 0) in the variances' parameter space, and regular inference statistical procedures for such a parameter could be problematic. The goal of this paper is to introduce a practically feasible permutation method to make inferences about variance components while considering the boundary issue in LMM. The permutation tests with different settings (i.e., constrained vs. unconstrained estimation, specific vs. generalized test, different ways of calculating p -values, and different ways of permutation) were examined with both normal data and non-normal data. In addition, the permutation tests were compared to likelihood ratio (LR) tests with a mixture of chi-squared distributions as the reference distribution. We found that the unconstrained permutation test with the one-sided p -value approach performed better than the other permutation tests and is a useful alternative when the LR tests are not applicable. An R function is provided to facilitate the implementation of the permutation tests, and a real data example is used to illustrate the application. We hope our results will help researchers choose appropriate tests when testing variance components in LMM.

1. Introduction

Psychological, educational, and social science research often involve repeated measures, clustered subjects, or both. To deal with such data, linear mixed modeling is frequently used, which contains fixed effects and random effects (e.g., Demidenko, 2013; Fitzmaurice et al.,

Corresponding Author hdu@psych.ucla.edu.

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

2012; Hedeker & Gibbons, 2006; Laird & Ware, 1982; Singer & Willett, 2003). Fixed effects are the structural components and random effects capture the variance/covariance parts. For example, in a linear growth curve model, fixed effects are usually the population-average intercept and slope across individuals, and random effects are the individual deviations from the overall intercept and slope (Hoffman, 2015). When focusing on the random effects, the corresponding variance components could summarize the variation in the individual intercepts and slopes around the population averages (Singer & Willett, 2003).

Although researchers are usually interested in the fixed effects, variance components can be of interest in order to examine whether there is heterogeneity between individuals or clusters and how large the heterogeneity is. For example, in educational research, such tests could investigate the variation in student achievements across schools/regions; in medical research, it may be of interest to test the between-person variation in treatment effects. Additionally, in longitudinal studies, the tests could answer questions about whether individual differences in change exist. If there is no heterogeneity between individuals or clusters, there is no need to specify a multilevel model. If heterogeneity exists, then variance components may have influences on the estimation of the whole model and results of the fixed effects even when they are not of interest. For example, when the statistical inference of variance components is challenging, Heywood cases and nonpositive definite matrices may occur which lead to nonconvergence using iterative methods, such as maximum likelihood (ML) methods (e.g., Bentler & Chou, 1987; Kolenikov & Bollen, 2012). In this case, whether the estimation of fixed effects reaches a local maximum or the global maximum is unknown.

Nonconvergence is only one of the problems in variance components inferences, and there are other problems in inferring variance components such as hypothesis testing, confidence interval calculation, and statistical power calculation. One of the essential reasons leading to the difficulty in the inferences is that there is a boundary (i.e., 0) in the variances' parameter space when only admissible (nonnegative) variances are allowed (i.e., constrained estimation). This is the so-called boundary issue. More specifically, in testing whether a variance component is 0, the null hypothesis $H_0: \sigma^2 = 0$ assumes that the true parameter is exactly on the boundary, thus the statistical distributions from the regular procedures no longer hold (Saville et al., 2011). For example, the asymptotic central chi-squared distributions of the likelihood ratio (LR) test statistics and Wald test statistics are not valid when the test value of the null hypothesis lies on the boundary (e.g., Self & Liang, 1987; Shapiro, 1985, 1988; Stram & Lee, 1994). More details will be discussed later. A consequence is that the p -value from the classical procedures tends to be inflated compared to the p -value from the true distribution, leading to too conservative results (e.g., Stoel et al., 2006).

A possible solution to avoid the boundary issue is to use unconstrained estimation where negative variance estimates are allowed. In this case, the boundary of the parameter space disappears. There are controversial opinions towards unconstrained estimation. On the one hand, it may result in Heywood cases in factor analysis which would then lead to nonconvergence (e.g., Dijkstra, 1992). Furthermore, negative variances do not have any practical meaning: it cannot necessarily imply that the true variance is very small or 0. On the other hand, researchers have argued that it is more than a "convenient choice", because

as long as other necessary statistical conditions are met (e.g., the model is identified), the regular test statistics and the corresponding distributions are unaffected, which is a great advantage (e.g., Savalei & Kolenikov, 2008). Savalei and Kolenikov (2008) also pointed out that unconstrained estimation provides a way to distinguish model misspecification from the boundary issue.

Besides unconstrained estimation, other methods have been proposed to address the boundary issue. However, many of the methods (e.g., LR tests with a mixture of chi-squared distributions) are limited by their strict assumptions and applicable models, which are not easy to implement. The permutation test, as a nonparametric method, provides a convenient and general way to test variance components, and avoids the restriction to a specific model and the corresponding assumptions to a large degree. Permutation methods empirically generate the reference distributions conditional on the observed data. The reference distributions are conditionally distribution-free and do not rely on the true underlying distributions that are hard to analytically calculate (Ernst et al., 2004). We propose to combine unconstrained estimation with permutation to take advantage of both methods. That is, we propose to (1) handle the boundary issue by unconstrained estimation, and (2) approximate the reference distribution by permutation to avoid distributional assumptions and large data requirements. More specifically, the permutation test does not require the regular normality assumption on both data and Level-1/Level-2 residuals in linear mixed models, and large sample sizes at different levels that are necessary for using the asymptotic reference distributions in parametric variance tests. In addition, we use least squares estimation instead of ML methods for estimating the parameters in LMM to avoid nonconvergence and computational issues. Thus, the main objective of this study is to introduce a practically feasible permutation method to make inferences about variance components while considering the boundary issue in linear mixed modeling.

The outline of this paper is as follows: in Section 2, a brief overview of linear mixed modeling is given. In Section 3, several existing methods that address the boundary issue in variance components are introduced, some of which motivated the current study. In Section 4, we present the proposed permutation tests. In Section 5, the permutation tests with different settings (i.e., constrained vs. unconstrained estimation, specific vs. generalized test, different ways of calculating p -values, and different ways of permutation) are thoroughly examined via Monte Carlo simulations. In Section 6, an R function is provided for implementing the recommended permutation methods. In Section 7, a real data example is provided to illustrate the recommended permutation methods. We end the paper with some concluding remarks in Section 7.

2. Review of Linear mixed modeling

A typical linear mixed model (LMM) is

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i. \quad (1)$$

Using the longitudinal context as an example for illustration, $\mathbf{Y}_i$ is an $n_i \times 1$ vector of repeated measurements for individual i ($i = 1, \dots, N$; n_i is the Level-1 sample size and N is

the Level-2 sample size), $\boldsymbol{\beta}$ is an $m \times 1$ vector of population parameters (fixed effects; m is the number of fixed effects), $\mathbf{X}_i$ is an $n_i \times m$ design matrix for the fixed effects, $\mathbf{b}_i$ is a $k \times 1$ vector for random effects with $E(\mathbf{b}_i) = \mathbf{0}$ and $\text{var}(\mathbf{b}_i) = \mathbf{D}$ ($\mathbf{D}$ is the Level-2 variance-covariance matrix; k is the number of random effects), $\mathbf{Z}_i$ is an $n_i \times k$ design matrix for the random effects, and $\boldsymbol{\varepsilon}_i$ is an $n_i \times 1$ independently distributed error term with $E(\boldsymbol{\varepsilon}_i) = \mathbf{0}$ and $\text{var}(\boldsymbol{\varepsilon}_i) = \mathbf{I}\sigma_e^2$. When $\mathbf{X}_i = \mathbf{Z}_i = \mathbf{X}$, the model is called a balanced random-coefficient model (Demidenko, 2013; Fitzmaurice et al., 2012). It is not rare to have multiple Level-1 predictors such as time-varying covariates in real data analysis (e.g., Cillessen & Borch, 2011; Mayeux & Cillessen, 2008), which could lead to multiple Level-2 variances of interest. Researchers may be interested in testing whether a single variance component, a subset, or even all the variance components in $\mathbf{D}$ are 0. For example, when a model contains intercept, linear change, and quadratic change, one may be interested in whether there is between-individual heterogeneity in any possible changes (variances of linear change and quadratic change); in this case, the researcher can conduct a test of a subset of the variance components (2 out of 3 variances). In contrast, when one is interested in testing whether there is any between-individual heterogeneity, the researcher can test all of the variance components (variances of intercept, linear change, and quadratic change) simultaneously. As another example, when there are time-varying covariates included, one may be interested in whether there is heterogeneity in any of the effects of the time-varying covariates (TVC), which would lead to a test of a subset of the variance components (variances of the TVC effects) as well. However, it's challenging or even not applicable to use existing parametric tests such as likelihood ratio tests to test multiple variance components simultaneously. Therefore, one of the goals of this study is to propose and evaluate a method that can simultaneously test a set of variance components.

Based on whether considering the information in the covariances, there are specific tests and generalized tests, where a specific test focuses on only testing the variance components, whereas a generalized test examines whether variances and corresponding covariances are 0 (Hertzog et al., 2008; Ke & Wang, 2015; Du, 2018).¹ In this paper, the definition of specific versus generalized tests is based on whether the information of covariances is used in test statistics. Therefore, another goal of this study is to compare the performance of specific permutation tests and generalized permutation tests in testing variance components.

3. Review of existing methods for variance components inferences

3.1 Likelihood ratio (LR)

As mentioned earlier, the LR test with central chi-squared distributions would fail when the parameters under the null hypothesis lie on the boundary with constrained estimation.² Mathematically, Self and Liang (1987), Shapiro (1985), and Shapiro (1988) derived a general form for the distribution of the LR test statistic under the null hypothesis with the boundary issue as a mixture of chi-squared distributions. When the Fisher information matrix of the estimated parameters is diagonal, it is straightforward to calculate the weights

¹Some researchers distinguish specific versus generalized tests by whether setting the covariances at 0 in alternative hypotheses (e.g., Drikvandi et al., 2013). That is, if all covariances are freely estimated, it is a generalized test; otherwise, it is a specific test.

²The Wald test is asymptotically equivalent to the LR test, thus it is not elaborated here.

of the mixture distributions. Otherwise, the weights depend on the covariances of the parameters. Shapiro (1985) provided an analytical form for the weights when the number of boundary parameters is four or smaller; when the number exceeds four, the analytical form of the weights cannot be easily expressed.

With regard to variance testing, Stram and Lee (1994), Stram and Lee (1995), and Stoel et al. (2006) generalized the previous findings, and found that the mixture ratio for testing a single variance component is 0.5:0.5. However, in testing multiple variance components, the situation becomes complicated: the diagonal information matrix of the variances is an unrealistic assumption, and positive semi-definite constraints exist. Therefore, there is no simple analytical form for the weights when testing multiple variances. And which chi-squared distributions comprise the mixture distribution depends on the number of unconstrained parameters of interest, the number of constrained parameters of interest, and the constraints on the estimation, which is very complex (Wu & Neale, 2013). In another line of research, instead of using the approximation with a mixture of chi-squared distributions, Crainiceanu & Ruppert (2004) and Greven et al. (2008) constructed the reference distribution of the LR test for testing a single variance component based on spectral decomposition. This approximation performs better than the mixture of chi-squared distributions when the model has nonparametric smoothing functions.

One may wonder whether we can test variance components one by one to avoid the dilemma of finding the analytical form. There are two drawbacks of this approach. First, post hoc corrections, such as Bonferroni correction, are needed to adjust for multiple testing. It is well-known that the Bonferroni correction usually leads to too conservative results. Second and more importantly, Self and Liang (1987), Greven et al. (2008), and Wu and Neale (2013) noticed that when testing one of the variance components, if any other untested variance component (nuisance parameter) is on the boundary, non-chi-squared distribution components exist in the mixture distribution, therefore the reference distribution of the LR test is not a pure mixture of chi-squared distributions. Based on our pilot simulation results (the variance of random slopes is the parameter of interest, and there are 2 or 5 other variance components in the multilevel model; see the supplemental material), the 0.5:0.5 mixture of chi-squared distributions failed to control the Type I error rate when one or more untested variance components were 0. More specifically, the specific LR tests could yield too liberal Type I error rates and the generalized LR tests could yield too conservative Type I error rates. The more nuisance parameters that were on the boundary, the more severely the Type I error rate deviated from the nominal level. In addition, with untested variances on the boundary, nonconvergence occurred when sample sizes were small in the multilevel modeling (MLM) framework.

Berkhof and Snijders (2001) found that the asymptotic mixture of chi-squared distributions provided acceptable Type I error rates (i.e., between 0.025 and 0.075 for $\alpha = 0.05$; Bradley, 1978) using the specific LR test to test the slope variance in a two-level model with a random intercept and a random slope. Ke and Wang (2015) used the mixtures of chi-squared distributions to test the slope variance component in a latent linear growth curve model. Specifically, Ke and Wang (2015) adopted the Satorra–Bentler scaling correction to conduct robust generalized and specific LR tests with 0.5:0.5 mixture distributions when data are

normal and non-normal, and conducted normal-theory-based generalized and specific LR tests with mixture distributions when data are normal. They found that the (robust) LR tests with mixture distributions generally could control the Type I error rates well. When the true covariance is not 0, the (robust) generalized LR tests had higher power than the other methods discussed in their paper (e.g., unconstrained estimation and (robust) specific LR tests); when the true covariance is 0, the (robust) specific LR tests were more powerful.

3.2 Permutation methods

As discussed earlier, the LR test with a central chi-squared distribution would fail when the parameters under the null hypothesis lie on the boundary with constrained estimation. In addition, the LR test with a mixture of chi-squared distributions would fail when the nuisance parameters are on the boundary and is relatively inapplicable when the Fisher information matrix of the estimated parameters is not diagonal. Thus, to address the limitations of LR tests for testing variance components, nonparametric methods were proposed. Drikvandi et al. (2013) proposed a permutation test for testing all variance components in LMM with longitudinal data. They proposed a test statistic for LMM in Equation (1),³

$$G = \frac{1}{N} \text{tr}(\mathbf{Z}_* (\mathbf{I} \otimes \hat{\mathbf{D}}) \mathbf{Z}_*') \quad (2)$$

where $\mathbf{Z}_* = \text{diag}(\mathbf{Z}_1, \dots, \mathbf{Z}_N)$, $\otimes$ is the Kronecker product, and $\hat{\mathbf{D}}$ is a distribution-free unbiased estimate of $\mathbf{D}$. Suppose a two-level random intercept and random slope model is fit to longitudinal data,

$$\begin{aligned} Y_{ij} &= \beta_{0i} + \beta_{1i}t_j + \varepsilon_{ij} \\ \beta_{0i} &= \beta_0 + b_{0i} \\ \beta_{1i} &= \beta_1 + b_{1i} \end{aligned} \quad (3)$$

where Y_{ij} is the score for individual i ($i = 1, \dots, N$) at time j ($j = 1, \dots, T$), t_j is the time for every individual at occasion j (i.e., all individuals are measured at the same time points), ε_{ij} is the Level-1 error term with variance σ_ε^2 , β_0 is the average intercept, β_1 is the average slope, and b_{0i} and b_{1i} are the random effects respectively with variances σ_0^2 and σ_1^2 . In this case, the design matrix for the random effects equals that for the fixed effects,

$$\mathbf{Z}_i = \mathbf{X}_i = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_T \end{pmatrix}. \text{ The test statistic for testing the intercept variance and slope variance is}$$

³Drikvandi et al. (2013) denoted the test statistics as T .

$$\begin{aligned}
G &= \frac{1}{N} \text{tr} \begin{pmatrix} X_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & X_N \end{pmatrix} \begin{pmatrix} \hat{D} & \cdots & \mathbf{0} \\ \vdots & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \hat{D} \end{pmatrix} \begin{pmatrix} X_1' & \cdots & \mathbf{0} \\ \vdots & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & X_N' \end{pmatrix} \\
&= \text{tr}(\mathbf{X} \hat{D} \mathbf{X}') \\
&= \text{tr} \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_T \end{pmatrix} \begin{pmatrix} \hat{\sigma}_0^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{12} & \hat{\sigma}_1^2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ t_1 & \cdots & t_T \end{pmatrix} \\
&= T\hat{\sigma}_0^2 + 2 \sum_j t_j \hat{\sigma}_{12} + \sum_j t_j^2 \hat{\sigma}_1^2.
\end{aligned}$$

The unbiased variance least squares estimator of $\mathbf{D}$ is presented in the supplemental material.

Under the null hypothesis $H_0: \mathbf{D} = \mathbf{0}$, $E(G) = 0$. Drikvandi et al. (2013) considered only constrained estimation. When $\hat{D}$ is not positive definite (e.g., negative variance estimates), $\hat{D}$ needs to be transformed to $\hat{D}^+ = \mathbf{P} \mathbf{\Lambda}_+ \mathbf{P}'$, where $\mathbf{P}$ and $\mathbf{\Lambda}$ are the matrix of eigenvectors and diagonal matrix of eigenvalues of $\hat{D}$ respectively, and $\mathbf{\Lambda}_+ = \max(\mathbf{0}, \mathbf{\Lambda})$. $\hat{D}^+$ is the closest matrix to $\hat{D}$ in the nonnegative definite parameter space (Demidenko, 2013). For example, when $\hat{D} = \begin{pmatrix} 1 & 0.2 \\ 0.2 & -0.2 \end{pmatrix}$, $\hat{D}^+$ is $\begin{pmatrix} 1 & 0.163 \\ 0.163 & 0.026 \end{pmatrix}$. For the two-random-effect longitudinal model in Equation (3), the procedure of Drikvandi's permutation test for testing D is as follows: (1) remove the fixed effects, $\mathbf{Y}_i^* = \mathbf{Y}_i - \mathbf{X}_i \hat{\beta}$; (2) calculate the G test statistic G^O for the original sample; (3) assuming $H_0: \mathbf{D} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ is true, permute the individual indices for data at the first time point, then for the second time point, ..., and for the last time point; (4) calculate the permutation G test statistics G^P ; (5) repeat steps 3 and 4 B times; (6) the empirical p -value is calculated as the proportion of G^P s larger than G^O . The permutation test is theoretically valid, because under the null hypothesis, there are no intercept and slope variances across individuals (i.e., all individuals have the same intercept and slope), and the data at each time point are exchangeable across individuals.

Drikvandi et al. (2013) further provided a way to test a subset of variance components. To guarantee that the permuted data are from the null hypothesis that the variances of interest are 0, we need to remove the influence from the variances that we are not interested in during testing. For example, for the two-random-effect longitudinal model in Equation (3), when only the variance of the random slope (σ_1^2) is the parameter of interest, we need to compute the predicted random effects related to the intercept and remove them from the data. A good candidate for such prediction is the empirical Bayes estimator (also called empirical best linear unbiased predictor; Fitzmaurice et al., 2012), $\hat{\mathbf{b}}_i = \hat{D} \mathbf{Z}_i' \mathbf{\Sigma}_i^{-1} (\mathbf{Y}_i - \mathbf{X}_i \hat{\beta})$, where $\mathbf{\Sigma}_i = \text{cov}(\mathbf{Y}_i) = (\mathbf{I} \hat{\sigma}_\epsilon^2 + \mathbf{Z}_i \hat{D} \mathbf{Z}_i')$. Accordingly, the first step of the aforementioned process is changed to: removing both the estimated untested random effects (i.e., $\hat{\mathbf{b}}_i^{(1)}; \hat{\mathbf{b}}_i^{(1)} = \begin{pmatrix} \hat{b}_{0i} \\ 0 \end{pmatrix}$ in this example) and estimated fixed effects from the data,

$Y_i^* = Y_i - X_i\hat{\beta} - Z_i\hat{b}_i^{(1)}$. Drikvandi et al. (2013) defined the test statistic in testing a subset of the variance components as

$$G = \frac{1}{N} \text{tr} \left(Z_*^{(2)} (I \otimes \hat{D}_{22}) Z_*^{(2)'} \right), \quad (4)$$

where $Z_*^{(2)} = \text{diag}(Z_1^{(2)}, \dots, Z_N^{(2)})$ corresponding to the design matrix of the tested random effects and $\hat{D}_{22}$ is the sub-matrix of D that we are interested in testing. In the two-random-effect longitudinal model example, $G = \sum_j^T t_j^2 \hat{\sigma}_1^2$. One advantage of permutation tests in testing a subset of variance components (including one variance) is that it works regardless of whether the untested variances are on the boundary, which appears as a concern in LR tests.

Drikvandi's method in testing a subset of variance components is a "semi-specific" test based on our definition, because the covariances between $b_i^{(1)}$ (untested random effects) and $b_i^{(2)}$ (tested random effects) are not involved in Equation (4) but only the covariances within $b_i^{(2)}$ are considered. When $b_i^{(2)}$ is a scalar, the test is a specific test. We denote the tests using G (Equation (4)) as specific tests in this paper to distinguish them from the conventional generalized tests which consider all the covariances. Their simulation results showed that Drikvandi's method could control the Type I error rate well in testing all the variance components within their studied scenarios, no matter whether the distribution of the random effects is normal. In this case, because there is no untested random effect, Drikvandi's method is not a "semi-specific" test and there is no distinction between specific and generalized tests. However, Drikvandi's method could lead to liberal Type I error rates in testing a subset of variance components (e.g., 0.087 in Drikvandi et al., 2013).

4. New permutation tests for variance components in linear mixed modeling

The limitation of the aforementioned LR test statistics is that they follow mixtures of chi-squared distributions only under specific conditions, and one can calculate the weights of the mixture distributions analytically only in some limited scenarios. Even if the analytical form of the weights exists, the weights depend on each specific model and each specific sample if the information matrix is not diagonal, thus it is hard to generalize the conclusion from one scenario to other scenarios. Permutation-based methods, however, do not require the derivation of asymptotic distributions, and instead they construct the reference distribution empirically, which becomes a big advantage. In this paper, we will focus on evaluating and extending Drikvandi's permutation test for two reasons: (1) it is easy to implement, and (2) the test statistic in Drikvandi et al. (2013) adopts noniterative estimation (the variance least squares estimation and ordinary least squares estimation are used), which can avoid nonconvergence problems.

There are several questions that have not been answered in the previous permutation studies. (1) Drikvandi et al. (2013) used constrained estimation, and considering the good statistical properties of unconstrained estimation (e.g., Savalei & Kolenikov, 2008; Saville et al., 2011), unconstrained estimation coupled with the permutation test needs to be explored. In addition, Drikvandi's method failed to control Type I error rates in testing a subset of the variance components. Whether unconstrained estimation can fix such a problem needs to be explored. (2) it is found that generalized tests could be more powerful than specific tests when the covariances are non-zero (e.g., Hertzog et al., 2008; Ke & Wang, 2015), but Drikvandi et al. (2013) only considered specific tests. The performance of the generalized permutation tests in dealing with the boundary issue of variance components is unknown. (3) Drikvandi et al. (2013) proposed a way to permute individual indices at each time point based on the fact that every individual has the same intercept and slope under the null hypothesis. But after removing all the estimated fixed effects from the data, permutation across individuals at each time point is not the only approach to conduct permutation, and another way of implementing permutation becomes available, which will be discussed further in a following subsection.

4.1 Generalized variance test for testing a subset of the variance components

We propose generalized variance tests for testing a subset of variance components (including testing a single variance). When testing a subset of variance components (variances in D_{22}), the basic idea of the generalized tests is to include the covariances between $b_i^{(1)}$ (the random effects corresponding to D_{11}) and $b_i^{(2)}$ (the random effects corresponding to D_{22}) in test statistics (Du, 2018). When the variances in D_{22} are 0, not only the covariance within D_{22} but also the covariance between D_{11} and D_{22} are 0. Thus testing the variances within D_{22} and all relevant covariances simultaneously can increase statistical power. Extending the specific test statistic G by Drikvandi et al. (2013) in Equation (4), we propose the following generalized variance test statistic,

$$G_{\Delta} = \frac{1}{N} \text{tr}(Z_* (I \otimes \hat{D}^*) Z_*') - \frac{1}{N} \text{tr}(Z_*^{(1)} (I \otimes \hat{D}_{11}^*) Z_*^{(1)'}), \quad (5)$$

where $Z_*^{(1)} = \text{diag}(Z_1^{(1)}, \dots, Z_N^{(1)})$ corresponding to the random effects with variance-covariance matrix D_{11} , and $(\hat{D}^*) = \text{diag}(\hat{D})$ except that the off-diagonal elements of $\hat{D}^*$ are the square of those in $\hat{D}$ to ensure non-negativeness of the generalized variance test statistic. The first part of Equation (5) is the same as Equation (2) which is the summary of information in all the variances and covariances; the second part of Equation (5) is to exclude the information in D_{11} for testing the subset of variance components in D_{22} . When Z_i is the same across individuals, G can be simplified as $G_{\Delta} = \text{tr}(Z_i \hat{D}^* Z_i') - \text{tr}(Z_i^{(1)} \hat{D}_{11}^* Z_i^{(1)'})$. For example, $G_{\Delta} = 2 \sum_j^T t_j \hat{\sigma}_{12}^* + \sum_j^T t_j^2 \hat{\sigma}_2^{*2}$ in the two-random-effect longitudinal model in Equation (3).⁴

4.2 Different ways of calculating p-value

p -value is defined as the probability of observing such extreme data or more extreme data under the null hypothesis. Based on the permutation reference distribution and the observed test statistic G^O , the empirical p -value in permutation tests for a one-sided test $H_a: \theta > 0$ (Good, 2013; Efron & Tibshirani, 1994; Ernst et al., 2004) is calculated as

$$p = P(G^P > G^O \mid H_0). \quad (6)$$

With unconstrained estimation, G and G^P can be positive or negative. Therefore, besides the one-sided p -value, we also consider the two-sided p -value. However, the p -value in the permutation test for a two-sided test has an ambiguous definition. One of the problems is that the permutation reference distribution is not symmetric, thus it is not meaningful to double the probability at one tail. There are two ways to calculate a two-sided p -value. Efron & Tibshirani (1994) suggested taking the absolute value of G^P and G^O ,

$$p = P(|G^P| > |G^O| \mid H_0). \quad (7)$$

Ernst et al. (2004) noticed the mean of the reference distribution may not be 0 and suggested centering the reference distribution first,

$$p = P(|G^P - \bar{G}^P| > |G^O - \bar{G}^P| \mid H_0), \quad (8)$$

where $\bar{G}^P$ is the mean of the test statistics across permutations. However, even using this method the proportions on the two tails are still unequal. Another problem of the two-sided p -value is that an extremely negative test statistic is not necessarily caused by extremely negative variance estimates, but is also possibly due to the extremely negative covariance estimates. And even if extremely negative variance estimates are the main reason leading to the extremely negative test statistic, the negative variance estimates are not evidence of nonzero population variances. When calculating the statistical power, we would lose power by counting both extremely positive and negative variance estimates (i.e., a larger p -value) instead of counting only extremely positive variance estimates (i.e., a smaller p -value). Therefore, we expect that the two-sided tests have smaller statistical power than the one-sided tests.

4.3 Different ways of permutation

Taking the two-random-effect longitudinal model in Equation (3) as an example, when all the estimated fixed effects are removed from the data, and there are no individual differences in intercepts and slopes under the null hypothesis, each individual's data indeed are $Y_i = \epsilon_i$,

⁴We examined three more generalized variance test statistics. In the first one, $\hat{D}$ and $\hat{D}_{11}$ instead of $\hat{D}^*$ and $\hat{D}_{11}^*$ were used in Equation (5). In the second one, $\hat{D}_{11}$ was replaced by $\hat{D}_{12}\hat{D}_{22}^{-1}\hat{D}_{12}$, and the updated whole matrix $\hat{D}$ was used in Equation (2). In the third one, we took the absolute values of the off-diagonal elements of $\hat{D}$, and applied Equation (5). But none of them controlled the Type I error rates well in the simulation.

where ε_i is an $n_i \times 1$ independently distributed error term with $E(\varepsilon_i) = 0$ and $var(\varepsilon_i) = \sigma_\varepsilon^2$ (Figure 1a). Drikvandi et al. (2013) proposed to permute individual indices at each time point (Figure 1b). But since each individual's data are randomly distributed around $Y_i = 0$, it is possible to conduct another permutation by shuffling all the observations, ignoring individual and time indices, and then assigning the shuffled observations to each individual with the same number of measurement occasions (Figure 1c). Although both ways of permutation work conceptually, the performance of different ways of permutation with finite Monte Carlo permutation samples needs to be empirically compared. Note that regardless of permuting individual indices at each time point or permuting all observations, an individual's observation at a specific time point is bonded with the corresponding design matrix (i.e., the corresponding row in $\mathbf{Z}_i$). That is, the measurement occasion and the time-vary covariates will travel with the observation during permutation.

5. Simulation studies

5.1 Simulation Design

In terms of whether to allow negative variance estimates, both constrained and unconstrained estimation will be examined. With unconstrained estimation, negative variance estimates are allowed, and the reference distributions of G and G^- cover negative parameter space. Therefore, one-sided tests can be conducted based on Equation (6), and two-sided tests can be conducted based on Equation (7) or (8). With constrained estimation, G and G^- are always positive and only the one-sided test can be used. Equation (3) was used for generating data in the simulations, which is a longitudinal data analysis model with a random intercept and a random slope. Thus there are 2 Level-2 variance components. In the supplemental material, we considered more scenarios, where researchers may want to test one out of three variances, two out of three variances, or all of the three variances (please see the supplemental material for more details and simulation results).

In total, there were 6 factors considered in the simulation setting: (1) testing one out of two variance components or testing both variance components, (2) permutation across individuals at each time point or permutation across all observations, (3) N was set to be 10, 30, 50, or 100, (4) T was set to be 3, 5, or 50 (large N and T are specified to explore the asymptotic property of different methods), (5) the values of diagonal elements and off-diagonal elements of the variance-covariance matrix $\mathbf{D}$ were varied to examine Type I error rates and statistical power, and (6) the random effects followed a multivariate normal distribution or a multivariate t distribution ($\mathbf{t}_{df} = \mathbf{3} \left(\mathbf{0}, \frac{df-2}{df} \mathbf{D} \right)$), following Drikvandi et al., 2013). In terms of the fixed effects, $\beta_0 = 1$, $\beta_1 = 2$ (following Drikvandi et al., 2013), and the effect of time-varying covariate $\beta_2 = 0.5$. t_j was from 1 to T , and the time-varying covariate X_{ij} was generated from a normal distribution $N(0, 1)$ with no Level-2 structure. The permutation tests were conducted with 1000 Monte Carlo permutation samples; there were 1000 replications under each condition. We consider a Type I error rate between 2.5% and 7.5% as satisfactory (Bradley, 1978). Since the non-normal distribution cases have consistent conclusions to the normal distribution cases, we refer the readers to the supplemental materials for detailed results of the non-normal distribution cases.

There were 8 types of permutation tests considered in both permutation across individuals at each time point and permutation across all observations. The 8 tests are listed in Table 1. Among them, the constrained specific test statistic with the one-sided p -value, P_C , is the Drikvandi's test statistic. When testing all the variance components, since all the information in the variance-covariance matrix is used, there is no distinction in our specific and generalized permutation tests.⁵

The permutation tests were compared with the LR tests with mixture distributions. When testing one out of two variance components, the specific and generalized LR tests based on normality assumption, LR_{NS} and LR_{NG} , are compared to the critical values 2.71 ($.5\chi_0^2 + .5\chi_1^2$ at $\alpha = 0.05$) and 5.14 ($.5\chi_1^2 + .5\chi_2^2$ at $\alpha = 0.05$) respectively (Ke & Wang, 2015). When testing multiple variance components, the LR tests are not easy to implement because the number, degrees of freedom, and weights of chi-squared components in the mixture distribution are not easily accessible and even do not have an analytical form. Therefore the LR tests will not be discussed in the multiple variance components testing. Moreover, robust specific and generalized LR tests in accounting for non-normal data are adopted with the Satorra–Bentler scaling correction and denoted as LR_{RS} and LR_{RG} (Satorra, 2000; Satorra & Bentler, 2001; more details refer to Ke & Wang, 2015).⁶

5.2 Type I error rates

5.2.1 Scenario 1: Testing only the slope variance (two variance components in total).—The intercept variance σ_0^2 was specified at 1 when generating data, and the null

hypothesis is $H_0: D = \begin{bmatrix} \sigma_0^2 & 0 \\ 0 & 0 \end{bmatrix}$. When the number of time points and/or the number of

individuals are large, the calculation using Equations (4) and (5) is time consuming. For example, 1000 permutations ($B = 1000$) using Equations (4) and (5) would take about 25 hours when $N = 1000$ and $T = 3$. In the LR tests, typical latent growth curve modeling with ML in the structural equation modeling (SEM) framework cannot handle cases where the number of time points is not smaller than the number of individuals, because the variance-covariance matrix is singular and not invertible. In this case, nonconvergence will occur. With constrained estimation, the LR tests in SEM also had convergence problems even when the number of time points was smaller than the number of individuals. Under constrained estimation, the convergence rates were 0.286, 0.638, 0.292, and 0.651 for LR_{NS} , LR_{NG} , LR_{RS} , and LR_{RG} respectively when $N = 100$ and $T = 50$; the convergence rates were 0.217,

⁵We also evaluated a constrained estimation of D which is calculated in two steps, as a layman's method. We first allowed negative variances, and if the estimated variances are negative, we constrained the negative variances to be 0 and re-estimated D . In the simulation, we found that the two-step constrained estimation method generally led to either too liberal or too conservative test results.

⁶We also evaluated unconstrained specific and generalized LR tests with two steps, which may be adopted by researchers for convenience in real data analysis. That is, when the slope variance is estimated to be negative, the decision is to fail to reject the null

hypothesis $H_0: D = \begin{bmatrix} \sigma_0^2 & 0 \\ 0 & 0 \end{bmatrix}$; otherwise the specific and generalized LR test statistics are compared to the critical values of χ_1^2 and

χ_2^2 at $\alpha = 0.05$ respectively. Thus, there are two-step unconstrained normal-theory-based specific and generalized LR tests for normal data, and two-step unconstrained robust specific and generalized LR tests for non-normal data. Our simulation results showed that the two-step unconstrained LR tests provided too conservative Type I error rates. The main reason is that only a large positive variance leads to the rejection of the null hypothesis at the second step, thus we lose power to allow for negative variances and fail to reject the null hypothesis at the first step.

0.368, 0.226, and 0.386 for LR_{NS} , LR_{NG} , LR_{RS} , and LR_{RG} respectively when $N=300$ and $T=50$; the convergence rates were 0.197, 0.206, 0.216, and 0.233 for LR_{NS} , LR_{NG} , LR_{RS} , and LR_{RG} respectively when $N=1000$ and $T=50$; and the convergence rates were 100% in the other conditions. Fitting in the multilevel modeling (MLM) framework, however, did not have the non-convergence problem. Specifically, when SEM was fitted, the small number of participants and big number of time points (i.e., small n big p) issue existed in some conditions because of the use of the wide/multivariate data format, which could result in non-convergence. In contrast, the long/univariate data format is used in the MLM framework and thus the small n big p issue did not exist in MLM in the simulation. Therefore, the results of LR tests are presented from MLM in Figure 2 for the conditions where SEM had the convergence issue. The robust LR tests proposed by Satorra (2000) and Satorra & Bentler (2001) are available only in the SEM framework. When SEM did not converge, the robust LR test results were not available.

When the random effects follow a multivariate normal distribution, the results of the permutation tests across individuals are presented in Figure 2. The constrained estimation $\hat{D}^+$ could yield too conservative results in P_C (i.e., the constrained specific test which is the Drikvandi's test statistic) and P_C (i.e., the constrained generalized test). The Type I error rates of the unconstrained permutation tests, P_{1U} (i.e., the unconstrained specific test with the one-sided p -value), P_{2CU} (i.e., the unconstrained specific test with the two-sided centered p -value), and P_{1U} (i.e., the unconstrained generalized test with the one-sided p -value), were closer to 0.05 and more stable than the other permutation statistics. And among these permutation test statistics, P_{1U} outperformed the others. The pattern of the test statistics in the permutation tests across all observations was similar to that of the permutation tests across individuals, but overall the permutation tests across all observations performed worse than the permutation tests across individuals in controlling Type I error rates (summarized in Figure S1; Figures S1-S11 are in the supplemental materials). Therefore, the permutation tests across all observations will not be discussed in the following sections. LR_{NG} provided acceptable Type I error rates, and LR_{NS} had similar but relatively worse performance than LR_{NG} (Figure 2). LR_{RG} and LR_{RS} could fail to control Type I error rates. Both P_{1U} in the permutation tests across individuals and LR_{NG} generally provided acceptable rates and had similar performances. When the random effects followed a multivariate t distribution, the performance of the permutation tests was not obviously influenced (see the supplemental material for more details; Figures S2 and S3).

5.2.2 Scenario 2: Testing intercept variance and slope variance together (two variance components in total).—

When testing intercept variance and slope variance together, the null hypothesis is $H_0: D = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$. And there is no distinction between our specific and generalized tests based on our definition. Only the permutation tests were considered because the LR tests are not practically applicable for this scenario. In both multivariate normal distribution and multivariate t distribution conditions, the constrained and unconstrained permutation tests well controlled the Type I error rates (the normal case is presented in Figure 3 and the non-normal case is presented in Figure S4).

5.2.3 Summary of the performance of the tests on Type I error rates.—The performance of the examined methods are summarized based on the Type I error rates in Table 2. In sum, the permutation tests across all observations were inferior to the permutation tests across individuals in controlling Type I error rates. In testing both the variance components, the unconstrained permutation tests with the one-sided p -value (P_{1U}), the two-sided p -value (P_{2U}), and the two-sided centered p -value (P_{2CU}), and the constrained tests with the one-sided p -value (P_C ; Drikvandi's test statistic) well controlled the Type I error rates across the simulation conditions. When testing one out of two variances, there is a distinction between specific tests and generalized tests. The unconstrained specific test with the one-sided p -value (P_{1U}) outperformed the others (including the Drikvandi's specific test statistic, P_C) in the current simulation conditions, and LR_{NG} performed well regardless of whether the random effects followed a multivariate normal distribution, followed by the performance of LR_{NS} . For the extra simulation scenarios in the supplemental material, we found that in testing all of the three variance components, P_{1U} , P_{2U} , P_{2CU} , and P_C well controlled the Type I error rates across the simulation conditions. In testing two out of three variances, P_{1U} outperformed the others. However, in testing one out of three variance components, none of the discussed permutation tests well controlled the Type I error rates. For more detailed results, please see the supplemental material.

5.3 Statistical power

The same scenarios for examining Type I error rates are considered in this section to explore power. Based on the Type I error rates, the permutation tests across all observations are not discussed further due to their failure of controlling Type I error rates.

5.3.1 Scenario 1: Testing only the slope variance (two variance components in total).—Only the unconstrained permutation tests that relatively well controlled the Type I error rates (i.e., P_{1U} , P_{2CU} , and P_{1U}) were considered in power calculation. Intercept variance was set at 1, and different values of slope variance (σ_1^2) and covariance (σ_{01}) were considered in D , which are similar to the covariance matrix setting in Drikvandi et al. (2013): (0.02, 0.05), (0.02, 0.1), (0.05, 0.05), and (0.05, 0.1) when generating data. No matter whether the random effects followed a multivariate normal distribution (Figure 4) or a multivariate t distribution (Figure S8), the LR tests, as parametric methods, had higher power than the permutation tests in testing only the slope variance across all the conditions when the LR tests had converged results. Specifically, the specific normal-theory-based LR tests (i.e., LR_{NS}) demonstrated the highest power. Due to the reasons that were highlighted in Section 4.2, the one-sided test, P_{1U} , had higher power compared to the two-sided permutation test P_{2CU} , as expected.

5.3.2 Scenario 2: Testing intercept variance and slope variance together (two variance components in total).—Only permutation tests that provided acceptable Type I error rates (i.e., P_{1U} , P_{2U} , P_{2CU} , and P_C) were considered in exploring power. Different D matrices were considered: $\begin{bmatrix} 0.05 & 0.02 \\ 0.02 & 0.05 \end{bmatrix}$ and $\begin{bmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{bmatrix}$. Similar to testing only the slope variance, the unconstrained permutation test with the one-sided p -value, P_{1U} , had the highest power for the normal conditions and was more powerful than Drikvandi's test statistic, P_C .

(Figure 3). And P_{1U} , P_{2U} , P_{2CU} , and P_C yielded similar power values for the non-normal conditions (Figure S9).

5.3.5 Summary of the performance of the tests on statistical power.—In summary, when testing only the slope variance, the unconstrained specific permutation test with the one-sided p -value (P_{1U}) was the most powerful test among the permutation tests, and the specific LR tests (LR_{NS}) were more powerful than the generalized LR tests (LR_{NG}) (Figure 4). In testing all the variance components, the unconstrained permutation test with the one-sided p -value (P_{1U}) usually was the most powerful test among the permutation tests (Figure 4 and Figures S9-S11).

6. Real data example

A sample of 204 children were repeatedly measured by the Wechsler Intelligence Scale for Children (WISC) with 4 waves: the spring before first grade, the end of first grade, the end of third grade, and the end of fifth grade (McArdle & Epstein, 1987). The intelligence quotient (IQ) scores from the verbal subscale and performance subscale were collected. We are interested in whether the 204 children had a significant difference in linear changes across the 6 years and whether the children were homogenous in terms of both initial levels and changes. No covariates were considered. Therefore the parameters of interest are the slope variance (σ_1^2) and intercept variance (σ_0^2) in a linear growth curve model. The unconstrained specific permutation test with the one-sided p -value (P_{1U}) and the two LR tests (LR_{NS} and LR_{NG}) that well controlled the Type I error rates in the previous simulation of testing only the slope variance were employed in this real data illustration. The LR tests were implemented based on latent growth curve modeling. When testing both the intercept and slope variances, only P_{1U} was used because the LR tests are not applicable.

In terms of the changes in the verbal scores, the variance least squares estimate of σ_1^2 was 1.54, and the one-sided p -value of P_{1U} was 0.090, which indicates that there was no significant slope variance at $\alpha=0.05$. The ML estimate of σ_1^2 in the latent growth curve model was 1.53. LR_{NS} and LR_{NG} provided p -values < 0.001 , which indicated that there was a significant slope variance at $\alpha=0.05$. This result was consistent with the simulation results that the LR tests were more powerful than the permutation tests. In terms of the changes in the performance scores, the variance least squares estimate of σ_1^2 was 1.22, and the one-sided p -value of P_{1U} was 0.002, which indicated that there was a significant slope variance at $\alpha=0.05$. The ML estimate of σ_1^2 in the latent growth curve model was 1.21. And LR_{NS} and LR_{NG} provided p -values < 0.001 .

In terms of testing both initial levels and changes in the verbal scores, p -value of P_{1U} was smaller than 0.001, which indicated that there existed heterogeneity across individuals in one or both variances at $\alpha=0.05$. Testing both variances is practically challenging using the LR tests because both the weights and components of the mixture of chi-squared distributions are complex. Thus the simultaneous test results from the LR tests are not available.

7. Discussion

In real data analysis, researchers may be interested in examining variance(s) in linear mixed modeling, which reveals the evidence about heterogeneity between individuals or clusters. However, the relatively widely used parametric methods, such as the LR tests, are useful only within limited scenarios. Thus, this paper aims to provide a practically feasible permutation test for variance components that can be applied to various models. Specifically, we considered constrained vs. unconstrained estimation, specific vs. generalized test, different ways of calculating p -values, and different ways of permutation. Based on the simulation results, we provide an R function, `permutation.p`, to facilitate the practical application of the recommended permutation methods (please see the Appendix for more details).

Three questions were posed at the beginning as unsolved issues. Firstly, what is the performance of permutation tests with unconstrained estimation? On the one hand, both the unconstrained and the constrained permutation tests performed equally well in controlling Type I error rates when testing all the variance components under the studied scenarios. On the other hand, the unconstrained permutation tests generally performed better than the constrained permutation tests (including Drikvandi's test statistic) in controlling Type I error rates when testing one out of two variances (summarized in Table 2). And in most of the examined cases, the unconstrained permutation test with the one-sided p -value, P_{1U} , was more powerful than the constrained permutation test P_C (Drikvandi's test statistic) when testing all the variance components.

Secondly, do the generalized tests provide better statistical inference? The unconstrained generalized permutation test with the one-sided p -value, P_{1U} , could relatively well control the Type I error rate in testing a single variance out of two variances, but was inferior to the unconstrained specific permutation test with the one-sided p -value, P_{1U} , in controlling Type I error rates and providing statistical power. The LR tests yielded similar results in the sense that the specific LR tests (LR_{NS}) provided higher power than the generalized LR test (LR_{NG}) in testing only the slope variance. Note that the covariance between the random intercepts and random slopes was set at 0.02 or 0.05 in Scenario 1, which was small. Ke and Wang (2015) and Hertzog et al. (2008) found that when the true covariance is 0, the specific LR tests were more powerful than the generalized LR tests. Therefore, the small covariance probably could explain why the specific permutation tests were more powerful than the generalized permutation tests in the examined conditions.

Thirdly, what is the performance of different ways of permutation? The permutation tests across all observations were inferior to the permutation tests across individuals in controlling the Type I error rates. The cause of the underperformance of the permutation tests across all observations is probably due to their being less efficient. That is, when there are a relatively large number of individuals and/or a relatively large number of time points, it is more difficult for the permutation tests across all observations to shuffle the data well and test whether there is a between-individual variance.

Different ways of calculating p -values were explored in the paper. In testing a subset of variances, the specific unconstrained permutation tests with the one-sided p -value (P_{1U}) performed better than the specific unconstrained permutation tests with the two-sided p -value (P_{2U}) or with the two-sided centered p -value (P_{2CU}) in terms of controlling Type I error rates. As mentioned earlier, how to define the counterpart of the observed statistics in two-sided tests is controversial in skewed permutation distributions, and how to calculate the probability of observing as extreme as the observed statistic or more extreme statistics in two-sided tests is ambiguous. Moreover, the unconstrained permutation tests with the one-sided p -value were more powerful than the ones with the two-sided p -value. And only positive variance estimates and thus positive permutation statistics provide valid evidence of the existence of variance, thus it is more reasonable to treat the test as a one-sided test. Therefore, the unconstrained permutation tests with the one-sided p -value are suggested.

In terms of the comparison between LR tests and permutation tests, the permutation tests have several advantages. First, the discussed permutation tests in the paper are nonparametric tests, which are based on noniterative estimation (see supplemental material for details). Therefore, the permutation tests avoid the nonconvergence issue in the LR tests coupled with ML estimation, as illustrated earlier in the paper. Second, the permutation tests can be applied to relatively complex models; in contrast, the LR tests considering the boundary issue could not be generalized to many models, because the analytical form for the mixtures of chi-squared distributions in LR tests could be complex or even does not exist, except for a few exceptions. In addition, LR tests assume that the untested variances are not on the boundary; otherwise, LR tests could fail to control the Type I error rates. However, as a nonparametric method, the discussed permutation tests usually have lower statistical power compared to the LR tests when the LR tests are available.

The challenge of testing a subset of the variance components in permutation tests is that it is difficult to completely clean the random effects that are associated with the untested variances. For example, in testing only the slope variance, the predicted random effects related to the intercept are removed, however these random effects cannot be completely cleaned up. The not fully cleaned intercept variance and the corresponding covariance imply that the "cleaned" data may contain other systematic variance/covariance besides the slope variance, which makes the test of only the slope variance not very accurate. In other words, the permutation p -value is no longer calculated under the null hypothesis.

Based on the simulation results, in testing one out of two variances, the specific unconstrained permutation test with the one-sided p -value, P_{1U} , is recommended, because it was generally more powerful than the other examined permutation tests and had better Type I error rates than the other examined tests; LR_{NG} is also recommended, because it also generally provided acceptable Type I error rates and was even more powerful than P_{1U} . In testing all the variance components, the LR tests are relatively inaccessible due to the difficulty of calculating the mixture weights and mixture components. Thus, the unconstrained permutation test with the one-sided p -value, P_{1U} , is recommended, because it well controlled the Type I error rates and was more powerful than the other permutation tests. Based on the extra simulations in the supplemental material, we found in testing a subset of multiple variance components, if the number of total variances and/or the number

of untested variance are not large, P_{1U} is also recommended given its relatively large statistical power compared to the other permutation tests; otherwise, none of the permutation tests succeeded in controlling the Type I error rates. In testing one out of multiple variance components, LR tests could be considered if it is clear that all of the nuisance parameters (estimated but untested parameters) are not on the boundary, which can be a strong assumption. Improving the performance of the permutation tests in testing a subset of the variance components should be done in the future.

In the current study, we focus on the use of the least squares estimation coupled with the G test (Equation (2)) for permutation tests. One future direction is to explore the performance of maximum likelihood estimation for permutation tests using other test statistics, such as LR tests and Wald tests. The limitation of the current study is that the proposed permutation tests cannot deal with unbalanced data where measurement timing differs across individuals, and cannot directly deal with missing data. However, if we impute the missing data using missing data techniques, such as multiple imputation (e.g., Enders et al., in press), permutation tests can be applied to the imputed data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments:

ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions or the NIH is not intended and should not be inferred.

Funding: Lijuan Wang was supported by NIH grants 1R01HD087319 and 1R01HD091235 during the study period.

Appendix

R function for testing variances using permutation

We provide an R function for testing a single variance, a subset of variance components, or all variance components using permutation in linear mixed-effects modeling. The R function can be downloaded from https://github.com/hduquant/quantlab/blob/master/Permutation_Test/permutation.R. Based on the simulation results, when testing all variance components, the R function uses the unconstrained permutation test with the one-sided p -value; when testing a subset of variance components (including a single variance), the R function uses the unconstrained specific permutation test with the one-sided p -value. A specification of the function permutation.p is

```
permutation.p(B = 1000, N, t, p = 2, r = 2, target = c(1:r), x,z,y).
```

The argument B reads in the number of permutation samples. t specifies the Level-1 sample size, such as the number of occasions. N reads in the Level-2 sample size, such as the number of individuals. p reads in the number of fixed effects. r reads in the number of random effects, which can be different from the number of fixed effects. x reads in the

design matrix for the fixed effects, which is a $t \times N$ by p matrix. In x , the first individual's t by p matrix is presented first, then the second individual's t by p matrix is presented, and so on. The Level-2 covariates should be after the Level-1 covariates for each individual in x . And z reads in the design matrix for the random effects, which is a $t \times N$ by r matrix. In z , the first individual's t by r matrix is presented first, then the second individual's t by r matrix is presented, and so on. `target` specifies the tested variance components, such as testing only the first variance, `target = 1`, or testing all the variances, `target = c(1 : r)`.

References

- Bentler PM, & Chou C-P (1987). Practical issues in structural modeling. *Sociological Methods & Research*, 16(1), 78–117.
- Berkhof J, & Snijders TA (2001). Variance component testing in multilevel models. *Journal of Educational and Behavioral Statistics*, 26(2), 133–152. doi: 10.3102/10769986026002133
- Bradley JV (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2), 144–152. doi: 10.1111/j.2044-8317.1978.tb00581.x
- Cillessen AH, & Borch C (2011). Analyzing social networks in adolescence In Card NA, Selig JP, Little T, et al. (Eds.), *Modeling dyadic and interdependent data in the developmental and behavioral sciences* (pp. 61–85). Routledge, NY: New York.
- Crainiceanu CM, & Ruppert D (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1), 165–185.
- Demidenko E (2013). *Mixed models: theory and applications*. John Wiley & Sons.
- Dijkstra TK (1992). On statistical inference with parameter estimates on the boundary of the parameter space. *British Journal of Mathematical and Statistical Psychology*, 45(2), 289–309. doi: 10.1111/j.2044-8317.1992.tb00994.x
- Drikvandi R, Verbeke G, Khodadadi A, & Nia VP (2013). Testing multiple variance components in linear mixed-effects models. *Biostatistics*, 14(1), 144–159. doi: 10.1093/biostatistics/kxs028 [PubMed: 22930674]
- Du H (2018). *Testing variance components in linear mixed modeling using permutation* (Unpublished doctoral dissertation). University of Notre Dame.
- Efron B, & Tibshirani RJ (1994). *An introduction to the bootstrap*. CRC press.
- Enders CK, Du H, & Keller BT (in press). Substantive model-compatible imputation for multilevel regression models with random coefficients, interaction effects, and other non-linear terms. *Psychological Methods*.
- Ernst MD, et al. (2004). Permutation methods: a basis for exact inference. *Statistical Science*, 19(4), 676–685. doi: 10.1214/088342304000000396
- Fitzmaurice GM, Laird NM, & Ware JH (2012). *Applied longitudinal analysis* (Vol. 998). John Wiley & Sons.
- Good P (2013). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media.
- Greven S, Crainiceanu CM, Küchenhoff H, & Peters A (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics*, 17(4), 870–891.
- Hedeker D, & Gibbons RD (2006). *Longitudinal data analysis* (Vol. 451). John Wiley & Sons.
- Hertzog C, von Oertzen T, Ghisletta P, & Lindenberger U (2008). Evaluating the power of latent growth curve models to detect individual differences in change. *Structural Equation Modeling*, 15(4), 541–563. doi: 10.1080/10705510802338983
- Hoffman L (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. Routledge.
- Ke Z, & Wang L (2015). Detecting individual differences in change: Methods and comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(3), 382–400. doi: 10.1080/10705511.2014.936096

- Kolenikov S, & Bollen KA (2012). Testing negative error variances: Is a heywood case a symptom of misspecification? *Sociological Methods & Research*, 41(1), 124–167.
- Laird NM, & Ware JH (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4), 963–974. doi: 10.2307/2529876 [PubMed: 7168798]
- Mayeux L, & Cillessen AH (2008). It's not just being popular, it's knowing it, too: The role of self-perceptions of status in the associations between peer status and aggression. *Social Development*, 17(4), 871–888. doi: 10.1111/j.1467-9507.2008.00474.x
- McArdle JJ, & Epstein D (1987). Latent growth curves within developmental structural equation models. *Child Development*, 58, 110–133. doi: 10.1111/j.1467-8624.1987.tb03494.x [PubMed: 3816341]
- Satorra A (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures In Heijmans RD, Pollock D, & Satorra A (Eds.), *Innovations in multivariate statistical analysis: A festschrift for heinz neuendecker* (pp. 233–247). Dordrecht, The Netherlands: Kluwer Academic.
- Satorra A, & Bentler PM (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66(4), 507–514. doi: 10.1007/bf02296192
- Savalei V, & Kolenikov S (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods*, 13(2), 150. doi: 10.1037/1082-989x.13.2.150 [PubMed: 18557683]
- Saville BR, Herring AH, & Kaufman JS (2011). Assessing variance components in multilevel linear models using approximate bayes factors: a case-study of ethnic disparities in birth weight. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(3), 785–804. doi: 10.1111/j.1467-985x.2011.00685.x
- Self SG, & Liang K-Y (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398), 605–610. doi: 10.1080/01621459.1987.10478472
- Shapiro A (1985). Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika*, 72(1), 133–144. doi: 10.1093/biomet/72.1.133
- Shapiro A (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. *International Statistical Review/Revue Internationale de Statistique*, 56(1), 49–62. doi: 10.2307/1403361
- Singer JD, & Willett JB (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press.
- Stoel RD, Garre FG, Dolan C, & Van Den Wittenboer G (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods*, 11(4), 439. doi: 10.1037/1082-989x.11.4.439 [PubMed: 17154756]
- Stram D, & Lee J (1995). Corrections: Variance component testing in the longitudinal mixed effects model. *Biometrics*, 51(3), 1196–1196. doi: 10.2307/2533038
- Stram D, & Lee JW (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50, 1171–1177. doi: 10.2307/2533455 [PubMed: 7786999]
- Wu H, & Neale MC (2013). On the likelihood ratio tests in bivariate acde models. *Psychometrika*, 78(3), 441–463. doi: 10.1007/s11336-012-9304-2 [PubMed: 25106394]

Figure 1.

Different ways of permutation

(a). Figure 1: Original data before permutation

(b). Figure 1: Permutation across individuals at each time point

(c). Figure 1: Permutation across all observations ignoring individual indices

Figure 2:

Type I error rates of the permutation tests across individuals at each time point and the likelihood ratio tests when the random effects follow a multivariate normal distribution in Scenario 1 (testing the slope variance out of two variances)

Note: P_{1U} , P_{2U} , and P_{2CU} represent unconstrained specific permutation tests with the one-sided p -value, the two-sided p -value, and the two-sided centered p -value, respectively. P_{1U} , P_{2U} , and P_{2CU} represent unconstrained generalized permutation tests with the one-sided p -value, the two-sided p -value, and the two-sided centered p -value, respectively. P_C and P_{CA} represent constrained specific and generalized with the one-sided p -value, respectively. LR_{NS} and LR_{NG} represent specific and generalized LR tests based on normality assumption, respectively. LR_{RS} and LR_{RG} represent robust specific and generalized LR tests, respectively.

Figure 3:

Type I error rates and power of the permutation tests across individuals at each time point when the random effects follow a multivariate normal distribution in Scenario 2 (testing both variances)

Note: P_{1U} , P_{2U} , and P_{2CU} represent unconstrained permutation tests with the one-sided p -value, the two-sided p -value, and the two-sided centered p -value, respectively. P_C represents constrained permutation tests the one-sided p -value.

Figure 4:

Statistical power of the permutation tests across individuals at each time point and the likelihood ratio tests when the random effects follow a multivariate normal distribution in Scenario 1 (testing the slope variance out of two variances)

Note: P_{1U} and P_{2CU} represent unconstrained specific permutation tests with the one-sided p -value and the two-sided centered p -value, respectively. P_{1U} represents unconstrained generalized permutation tests with the one-sided p -value. LR_{NS} and LR_{NG} represent specific and generalized LR tests based on normality assumption, respectively.

Table 1:

Summary of the evaluated permutation tests

Type of test	Ways of calculating p -value	Unconstrained	Constrained
Specific (G)	One-sided p -value	P_{1U}	P_C
	Two-sided p -value	P_{2U}	
	Two-sided centered p -value	P_{2CU}	
Generalized (G)	One-sided p -value	P_{1U}	P_C
	Two-sided p -value	P_{2U}	
	Two-sided centered p -value	P_{2CU}	

Note: G denotes the test statistic from Equation (4), and G denotes the test statistic from Equation (5). The one-sided p -value is based on Equation (6), the two-sided p -value is based on Equation (7) from Efron, and the two-sided centered p -value is based on Equation (8) from Ernst.

Table 2:

Summary of the performance of the methods in controlling Type I error rates

			Permutation test		LR test with constrained estimation	
			Unconstrained	Constrained	Normal-based	Robust
Scenario 1: Test one out of two variances						
Specific	G	One-sided p -value			LR_S	
		Two-sided p -value				
		Two-sided centered p -value				
Generalized	G	One-sided p -value			LR_G	
		Two-sided p -value				
		Two-sided centered p -value				
Scenario 2: Test both two variances						
No distinction between specific and generalized tests	G	One-sided p -value			LR	
		Two-sided p -value				
		Two-sided centered p -value				

Note: A check mark means that the method generally provided acceptable Type I error rates in the simulations. The LR tests were not evaluated in Scenario 2 because the LR tests are not practically applicable for this scenario.