





Understanding, Testing, and Relaxing Sphericity of Repeated Measures ANOVA with Manifest and Latent Variables Using SEM

Benedikt Langenberg¹ , Jonathan L. Helm² , Thomas Günther^{3,4} , Axel Mayer¹ 

[1] *Psychological Methods, Faculty of Psychology and Sports Science, Bielefeld University, Bielefeld, Germany.*
[2] *Department of Psychology, San Diego State University, San Diego, CA, USA.* [3] *Child Neuropsychology Section, Department of Child and Adolescent Psychiatry, Psychotherapy and Psychosomatics, RWTH Aachen University, Aachen, Germany.* [4] *Psychological Diagnostics and Intervention, Institute of Psychology, RWTH Aachen University, Aachen, Germany.*

Methodology, 2023, Vol. 19(1), 60–95, <https://doi.org/10.5964/meth.8415>

Received: 2022-02-27 • **Accepted:** 2023-01-16 • **Published (VoR):** 2023-03-31

Handling Editor: Isabel Benitez, University of Granada, Granada, Spain

Corresponding Author: Benedikt Langenberg, Psychological Methods, Faculty of Psychology and Sport Science, Bielefeld University, Universitätsstraße 25, 33615 Bielefeld, Germany. E-mail: benedikt.langenberg@uni-bielefeld.de

Supplementary Materials: Materials [see [Index of Supplementary Materials](#)]



Abstract

This article demonstrates how to perform univariate repeated measures ANOVA (U-RM-ANOVA) as a special case of structural equation models (SEMs). In the literature, sphericity is usually defined in terms of variances of pairwise differences of within-subject conditions. This article illustrates the original definition by Huynh and Feldt (1970) in terms of (co)variances of contrasts using SEM and demonstrates how to impose, test, and relax sphericity, and how to test main/interaction effects with and without the assumption of sphericity in SEM. We perform two simulation studies. The first study compares Mauchly's sphericity test with an SEM based test and shows that the two approaches have a very similar Type 1 error and power. The second study compares U-RM-ANOVA with SEM for different degrees of departure from sphericity and shows that U-RM-ANOVA and SEM have similar statistical properties in terms of Type 1 error, power, as well as similar bias and efficiency of effect size estimates of main and interaction effects. We furthermore show how to implement sphericity in latent variable models and provide software to perform the proposed tests and analyses.



Keywords

sphericity, structural equation modeling, analysis of variance, growth curve modeling

Over the past 30 years, structural equation models (SEMs) have become common for analyzing longitudinal data. For instance, growth curve models or latent change models are prominent examples and are special cases of SEMs (for more examples, see e.g., [Newsom, 2015](#)). However, less is known regarding the direct connection between SEMs and univariate repeated measures ANOVA (U-RM-ANOVA). More specifically, even though it is well known that both regression and ANOVA are special cases of SEM, to our knowledge, there is no documented approach on how to test main and interaction effects commonly implemented in U-RM-ANOVA using SEM. Although the connection between U-RM-ANOVA and SEM may seem trivial, there are at least three challenges that quickly become apparent while attempting to identify that connection. First, it is not obvious how to; (1) impose or test sphericity in SEM, (2) test the main/interaction effects of U-RM-ANOVA (e.g., test the main/interaction effects of, say, a 2×3 fully within subjects repeated measures design), and (3) impose sphericity on latent variables (i.e., perform the U-RM-ANOVA using latent rather than manifest variables). This article tackles these challenges by identifying and demonstrating how to perform each in SEM, and thereby builds on the expansive range of literature on methods for analyzing repeated measures using SEM (e.g., growth curve models, [McArdle, 1988](#); [McArdle & Epstein, 1987](#); [Meredith, 1993](#); latent change models, [McArdle, 2009](#); [McArdle & Hamagami, 2001](#); [Raykov, 1999](#); [Steyer et al., 1997](#); for an overview, see [Newsom, 2015](#)).

Furthermore, there are at least four benefits to formally identifying the connection between U-RM-ANOVA and SEM. First, sphericity is often a difficult concept for researchers to grasp, and has a colloquial definition based on the variances of differences between all possible pairs of within-subject conditions (e.g., [Field, 1998](#); [Field et al., 2012](#); [Lane, 2016](#); [Nimon, 2012](#)) that only holds in designs with one factor. In contrast, this article demonstrates how sphericity may be specified in SEM, which may help researchers understand its meaning. Second, researchers may also test and/or relax the assumptions of sphericity in SEM (without having to use post-hoc adjustments that are common to U-RM-ANOVA; e.g., Greenhouse-Geisser ([Greenhouse & Geisser, 1959](#)) and Huynh-Feldt ([Huynh & Feldt, 1970](#)) adjustments. Third, once in the SEM framework, researchers may capitalize on the other benefits of SEM, such as built-in approaches to handle both missing data and adjustments for non-normality (see the [Conclusions and Future Directions](#) section for further detail). Fourth, and which will be demonstrated in the article, the SEM framework allows for measurement models, which researchers may want to use in cases wherein manifest variables likely contain measurement error.

Although the analysis of repeated measures via SEM is not new, virtually all of the literature describes analyses in terms of models for repeated measures across time (e.g., growth curves, [McArdle, 1988](#); [McArdle & Epstein, 1987](#); [Meredith, 1993](#); latent

change models, McArdle, 2009; McArdle & Hamagami, 2001; Raykov, 1999; Steyer et al., 1997). These analyses are vitally important, but they do not apply to all empirical investigations. For example, many experimental psychologists aim to examine repeated measures across treatment or other experimental conditions, and the repeated measures do not follow a function of time (as assumed in growth curves and latent change models). Hence, this article builds the SEM literature for repeated measures by reconsidering how U-RM-ANOVA (which is not constrained to a specific ordering of time) can be implemented in SEM.

We limit the scope of this article in two ways. First, the article only considers within-subjects designs in order to focus on the sphericity considerations that underlie U-RM-ANOVA, and because mixed designs have been described elsewhere (see Langenberg et al., 2020). Yet, for interested readers, the discussion section briefly characterizes how to include between-subjects factors, and points to other articles that describe this topic in greater detail. Second, because there is extensive literature on these topics, this article assumes that readers have knowledge of the identification of latent variables in the SEM framework, and considerations of measurement invariance to compare means of latent variables across repeated measures. Relevant citations are provided (e.g., Newsom, 2015; Pitts et al., 1996; Widaman et al., 2010).

The remainder of this article includes: (1) a guiding empirical example, (2) a review of orthogonal contrasts (which will be central to testing both sphericity and main/interaction effects in SEM), (3) a review of sphericity, (4) how to impose, test, and relax sphericity in SEM, (5) a simulation study that compares Mauchly's sphericity test to SEM, (6) how to test main/interaction effects of U-RM-ANOVA in SEM (including how to reproduce F-values from U-RM-ANOVA in SEM), (7) a simulation study that compares RM-ANOVA to SEM with and without assuming sphericity, (8) an extension of SEMs that include measurement models, and (9) some conclusions and future directions. By the article's close, readers will have a clear understanding of sphericity, how it may be imposed in SEM, and how to test the hypotheses of U-RM-ANOVA in SEM.

Guiding Empirical Example

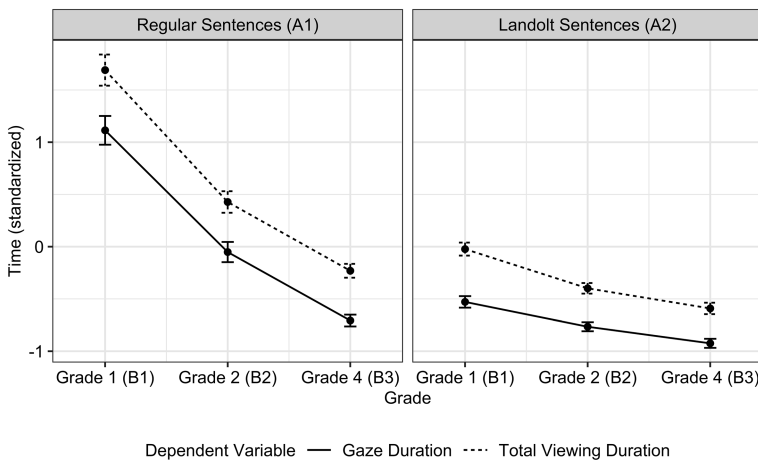
Here we describe an empirical example to guide readers through the remainder of the article. The example stems from a study aimed to investigate the development of different processes involved in reading; including both the necessary motor skills for fixating on a sentence, and the cognitive skills to process the sentence. A total of $N = 268$ children were asked to read a set of sentences during Grades One, Two, and Four (each child completed three repeated measures, $N = 169$ children had complete data). Among other variables, the researchers measured mean gaze and mean total viewing duration. The variables were log-transformed and standardized.

The comparison (i.e., control) condition used “Landolt sentences” (e.g., Heim et al., 2018; Hillen et al., 2013), which simply replace each character of a regular sentence with a circle. The difference in mean gaze and mean total viewing duration across regular and Landolt sentences measures processing time (i.e., processing time for a regular sentence includes time for both sentence fixation and processing; whereas processing time for Landolt sentences only includes time for sentence fixation).

The experimental design conforms to a 2×3 repeated measures design, with sentence type (Factor A: A1 = real sentences; A2 = Landolt sentences) and grades (Factor B: B1 = Grade One; B2 = Grade Two; B3 = Grade Four) as within-subject factors. Figure 1 displays mean gaze duration (solid line) and the mean total viewing duration (dashed line) for the two sentence types across the three measurement occasions. In the next section, we only focus on the dependent variable mean gaze duration. The section [Testing Main and Interaction Effects of U-RM-ANOVA Using L-RM-ANOVA](#) uses both dependent variables as indicators of a common latent variable in a measurement model.

Figure 1

Standardized Mean Gaze Duration and Mean Total Viewing Duration for the Two Sentence Types



Note. The solid line indicates standardized mean gaze duration; the dashed line indicates mean total viewing duration for the two sentence types (left panel: regular sentences; right panel: Landolt sentences) across the three measurement occasions. Error bars indicate standard errors.

Review of Orthogonal Contrast Matrices

A matrix of orthogonal contrasts will need to be included in SEM to test both the assumption of sphericity, and the main/interaction effects. This section reviews orthogonal contrast matrices in preparation for their inclusion in SEM later in this article.

For R repeated measures (in the empirical example, $R = 6$) a set of $R - 1$ orthogonal contrasts may be constructed to test the main and interaction effects of U-RM-ANOVA, and are usually organized into an $(R - 1) \times R$ matrix. For example, if the data from the empirical example are organized into an R dimensional column vector, denoted \mathbf{y} , whose elements conform to the order {A1B1, A1B2, A1B3, A2B1, A2B2, A2B3}, then the following orthogonal contrast matrix, \mathbf{C} , can be defined as

$$\mathbf{C} = \begin{bmatrix} -0.4082 & -0.4082 & -0.4082 & 0.4082 & 0.4082 & 0.4082 \\ -0.5000 & 0.0000 & 0.5000 & -0.5000 & 0.0000 & 0.5000 \\ 0.2887 & -0.5774 & 0.2887 & 0.2887 & -0.5774 & 0.2887 \\ 0.5000 & 0.0000 & -0.5000 & -0.5000 & 0.0000 & 0.5000 \\ -0.2887 & 0.5774 & -0.2887 & 0.2887 & 0.5774 & 0.2887 \end{bmatrix} \quad (1)$$

wherein the first, second/third, and fourth/fifth rows respectively enable tests for the main effect of A, main effect of B, and the interaction. In particular, multiplying \mathbf{y} by \mathbf{C} , such that

$$\boldsymbol{\pi} = \mathbf{C}\mathbf{y} \quad (2)$$

creates an $R - 1$ dimensional column vector (i.e., $\boldsymbol{\pi}$) of contrast variables; and if the null hypothesis for the main effect of A, the main effect of B, or the interaction effect is true, then we would respectively expect the means of the first, second/third, or fourth/fifth elements of $\boldsymbol{\pi}$ to equal zero. Stated differently, if five means of $\boldsymbol{\pi}$ are respectively labeled μ_{π_1} through μ_{π_5} , then the null hypothesis for the main effect of A prescribes $\mu_{\pi_1} = 0$; the null hypothesis for the main effect of B prescribes that both $\mu_{\pi_2} = 0$ and $\mu_{\pi_3} = 0$; and the null hypotheses for the interaction prescribes that both $\mu_{\pi_4} = 0$ and $\mu_{\pi_5} = 0$. All that remains is implementing a method (e.g., calculating p -values) to test whether the sample contrast means significantly differ from zero.

There are two necessary criteria for \mathbf{C} to be termed an orthogonal contrast matrix. First, the rows must sum to zero. Second, the rows (but not the columns) must be independent from one another (i.e., $\mathbf{C}\mathbf{C}^T$ equals a $(R - 1) \times (R - 1)$ diagonal matrix, with zeros in the off-diagonal). The construction of orthogonal contrast matrices for repeated measures designs can be complex (especially for designs with two or more factors). Hence, we encourage readers to use statistical software, such as R (R Core Team, 2021) to construct an orthogonal contrast matrix based on a specific design (see Appendix A for R code on how to construct orthogonal contrast matrices; or see UCLA Statistical Consulting Group, 2011 for an overview of different types of orthogonal contrast matrices, along with code to produce those matrices),

A specific type of orthogonal contrast matrix, termed an *orthonormal* contrast matrix, is of special interest in the context of U-RM-ANOVA. Orthonormal contrast matrices are orthogonal contrast matrices (i.e., orthonormal matrices satisfy the two criteria of

orthogonal matrices), whose sum of squared row elements equals 1 (i.e., after squaring each value in the matrix, the sum across each row equals 1). Consequently, the \mathbf{C} matrix defined in Equation 1 is an orthonormal contrast matrix. The small benefit of the orthonormal matrix (which will be demonstrated later in this article) is that the sums of squares, mean squares, and F -values of U-RM-ANOVA may be exactly replicated in SEM with the use of an orthonormal contrast matrix (as opposed to an orthogonal contrast matrix, which only directly reproduces F -values, see Voelkle, 2007). As an aside, we note that if some contrast matrix \mathbf{C} is orthogonal but not orthonormal, then the rows of \mathbf{C} may be scaled to make a new orthonormal matrix (this may be useful for readers using statistical software that can produce orthogonal contrasts, but not orthonormal contrasts, e.g., we rescaled an orthogonal contrast matrix produced by the statistical software R to obtain \mathbf{C} in Equation 1). Appendix B demonstrates how to rescale the rows of an orthogonal matrix to create an orthonormal matrix.

Taken together, this section alludes to how an orthogonal contrast matrix may be used to evaluate the null hypotheses of main and interaction effects from U-RM-ANOVA (i.e., by forming, and testing whether specific means of π significantly differ from zero). Later in this article we will use this information to estimate π from \mathbf{y} in SEM, and use the tools of SEM to perform the significance tests that reflect main and interaction effects.

Review of Sphericity

Univariate repeated measures ANOVA assumes that the variance covariance matrix of repeated measures conforms to a specific pattern, commonly referred to as “sphericity” or a “spherical matrix”. Therefore, using SEM to test main and interaction effects in the same manner as U-RM-ANOVA requires sphericity to be imposed in SEM, and this section reviews the definition of sphericity in preparation for its inclusion in SEM.

Importantly, the colloquial definitions of sphericity provided in applied statistics textbooks for psychology researchers are often simplifications that do not readily generalize to U-RM-ANOVA designs with two or more within subjects factors. In response, we review the colloquial definitions, describe their potential for misunderstanding, describe the actual definition as originally provided in the statistics literature (Huynh & Feldt, 1970, p. 1587, Theorem 3), and clarify that definition for U-RM-ANOVA designs with more than one within-subjects factor.

Colloquial Definition of Sphericity

In the psychology literature, the colloquial definition of sphericity states that the variances of all pairwise differences between repeated measures are equal (see for example, Field, 1998; Field et al., 2012, pp. 550–552; Lane, 2016; Nimon, 2012). For example, in

a one-way U-RM-ANOVA with three levels (i.e., A1, A2, and A3; not the same as this article's guiding empirical example), the common definition of sphericity prescribes

$$\sigma_{Y_{A1} - Y_{A2}}^2 = \sigma_{Y_{A1} - Y_{A3}}^2 = \sigma_{Y_{A2} - Y_{A3}}^2. \quad (3)$$

The colloquial definition is sufficient for sphericity for designs with one within-subjects factor, but does not clearly generalize to designs with more than one within-subjects factor.

For example, researchers may infer that sphericity for the 2×3 design from the empirical example implies that the 13 unique pairwise differences across the 6 repeated measures must have equal variances, such that

$$\sigma_{Y_{A1B1} - Y_{A1B2}}^2 = \sigma_{Y_{A1B1} - Y_{A1B3}}^2 = \sigma_{Y_{A1B1} - Y_{A2B1}}^2 = \dots = \sigma_{Y_{A2B2} - Y_{A1B3}}^2 = \dots = \sigma_{Y_{A2B2} - Y_{A2B3}}^2. \quad (4)$$

Unfortunately, this constraint is neither required nor implied by sphericity.

A second confusing aspect of the colloquial definition of sphericity concerns the separate tests of sphericity provided by statistical software for each main or interaction effect (i.e., each main and interaction effect receive their own test of sphericity unless the effect has one degree of freedom; see the section [Sphericity for Main and Interaction Effects of U-RM-ANOVA](#) for details). In particular, if the colloquial definition of sphericity were true, then only one test should be needed regardless of the within-subjects design (i.e., the definition refers to pairwise differences rather than main/interaction effects). Therefore, the outputs from statistical software implementing U-RM-ANOVA do not corroborate the colloquial definition.

We want to briefly mention that, in fact, an omnibus test can be constructed to test for sphericity in multiple effects simultaneously. This omnibus test can decrease the Type I error rate of incorrectly rejecting the assumption that sphericity holds. This test will be described in the section [Testing Sphericity Using L-RM-ANOVA](#).

Separate from the colloquial definition, psychology texts often describe compound symmetry as a special case of sphericity, and then describe assumptions in terms of compound symmetry (Field, 1998; Haverkamp & Beauducel, 2017; Maxwell & Delaney, 2004). Even though these texts do not claim equality across compound symmetry and sphericity, the compound symmetry simplification also does not easily generalize to higher-order within-subjects designs.

Original Definition of Sphericity

Huynh and Feldt (1970, p. 1587, Theorem 3) provide the original definition of sphericity. In general (i.e., not specific to the guiding empirical example), a $P \times P$ matrix Σ (e.g., a variance covariance matrix), conforms to sphericity if, and only if,

$$\mathbf{C}\Sigma\mathbf{C}^T = \sigma^2\mathbf{I} \quad (5)$$

wherein \mathbf{C} is a $(P - 1) \times P$ orthogonal contrast matrix, σ^2 is some positive constant, and \mathbf{I} is a $(P - 1) \times (P - 1)$ identity matrix. Therefore, a given matrix (e.g., Σ) adheres to sphericity if (after pre- and post-multiplying by an orthogonal contrast matrix), the diagonal elements are equal and the off-diagonal elements are zero.

However, the definition of sphericity requires more specificity for higher-order U-RM-ANOVAs (i.e., [Huynh & Feldt, 1970](#), definition was for a single repeated measures factor, not for within-subjects designs with more than one repeated measures factor). Similarly, multivariate statistics textbooks describing U-RM-ANOVA often provide the original definition by [Huynh and Feldt \(1970\)](#), without explaining how to extend the definition to obtain separate tests for each main/interaction effect (for example, see [Stevens, 2002](#), p. 421). Hence, we explain the original (1970) definition using the variance covariance matrix of the contrast data (i.e., \mathbf{V}_π) from the empirical example to show how to generalize the definition to multi-factorial designs.

Sphericity for Main and Interaction Effects of U-RM-ANOVA

Following the empirical example, let \mathbf{V}_y refer to the $R \times R$ variance covariance matrix of y , and \mathbf{V}_π refer to the $(R - 1) \times (R - 1)$ variance covariance matrix of π . Then, adhering to the definitions for \mathbf{C} given in [Equation 1](#), it follows that

$$\mathbf{V}_\pi = \mathbf{C}\mathbf{V}_y\mathbf{C}^T \quad (6)$$

and sphericity for each main/interaction effect will correspond to specific constraints within \mathbf{V}_π . In particular, for a given main or interaction effect, sphericity will hold if both (1) the variances in \mathbf{V}_π created from the effect's contrasts are all equal, and (2) the covariances in \mathbf{V}_π created from the effect's contrasts are all zero. To clarify that definition, we explicitly define \mathbf{V}_π from the empirical example as

$$\mathbf{V}_\pi = \begin{bmatrix} \sigma_{\pi_1}^2 & \sigma_{\pi_1\pi_2} & \sigma_{\pi_1\pi_3} & \sigma_{\pi_1\pi_4} & \sigma_{\pi_1\pi_5} \\ \sigma_{\pi_1\pi_2} & \sigma_{\pi_2}^2 & \sigma_{\pi_2\pi_3} & \sigma_{\pi_2\pi_4} & \sigma_{\pi_2\pi_5} \\ \sigma_{\pi_1\pi_3} & \sigma_{\pi_2\pi_3} & \sigma_{\pi_3}^2 & \sigma_{\pi_3\pi_4} & \sigma_{\pi_3\pi_5} \\ \sigma_{\pi_1\pi_4} & \sigma_{\pi_2\pi_4} & \sigma_{\pi_3\pi_4} & \sigma_{\pi_4}^2 & \sigma_{\pi_4\pi_5} \\ \sigma_{\pi_1\pi_5} & \sigma_{\pi_2\pi_5} & \sigma_{\pi_3\pi_5} & \sigma_{\pi_4\pi_5} & \sigma_{\pi_5}^2 \end{bmatrix} \quad (7)$$

and note that the variances and covariances in \mathbf{V}_π that respectively reflect the main effect of A, the main effect of B, and the interaction are $\{\sigma_{\pi_1}^2\}$, $\{\sigma_{\pi_2}^2, \sigma_{\pi_3}^2, \sigma_{\pi_2\pi_3}\}$, and $\{\sigma_{\pi_4}^2, \sigma_{\pi_5}^2, \sigma_{\pi_4\pi_5}\}$. Then, sphericity holds for the main effect of B if both $\sigma_{\pi_2}^2 = \sigma_{\pi_3}^2$ and $\sigma_{\pi_2\pi_3} = 0$, regardless of the remaining elements in \mathbf{V}_π ; sphericity holds for the interaction if both $\sigma_{\pi_4}^2 = \sigma_{\pi_5}^2$ and

$\sigma_{\pi_4\pi_5} = 0$, regardless of the remaining elements in \mathbf{V}_π ; and sphericity is not relevant for the main effect of A because there is only one variance and no covariances (i.e., at least two variances are needed to form an equality constraint between them, and at least one covariance is needed to be set equal to zero).

Thus, the general definition for sphericity in U-RM-ANOVA refers to \mathbf{V}_π (as opposed to variances of all pairwise differences). Once the set of orthogonal contrasts for the main/interaction effects of a specific U-RM-ANOVA design are identified and organized into an $(R - 1) \times R$ matrix \mathbf{C} , sphericity holds for each effects if, and only if, the variances in \mathbf{V}_π that reflect the effect are equal, and the covariances in \mathbf{V}_π that reflect the effect equal zero.

Testing Sphericity, Main Effects, and Interaction Effects in SEM

In the following sub-sections, we demonstrate how to test sphericity as well as main and interaction effects with and without assuming sphericity. We use the open-source R package `lavaan` (Rosseel, 2012) to estimate the models. We furthermore provide the open-source R package `semnova` (Langenberg & Mayer, 2020) which is an interface to `lavaan` and includes user-friendly functions to perform the proposed tests.¹

Latent Repeated Measures ANOVA (L-RM-ANOVA)

The SEM framework can be used to test sphericity because SEMs can (1) estimate \mathbf{V}_π from \mathbf{V}_y , (2) place constraints on the estimated elements in \mathbf{V}_π that conform to sphericity, and (3) perform likelihood ratio tests to quantify statistical significance (analogous to Mauchly's test). This section characterizes estimation of \mathbf{V}_π from \mathbf{V}_y (because the latter two points are commonplace), as a special case of latent repeated measures analysis of variance (L-RM-ANOVA, Langenberg et al., 2020; an extension of the growth components approach given by Mayer et al., 2012). L-RM-ANOVA estimates individual contrasts (i.e., π) as a set of latent variables in SEM. Stated differently, L-RM-ANOVA identifies how to rewrite Equation 2 as an SEM.

To clarify, recall that the SEM measurement model for P manifest variables and Q latent variables may be written as

$$\mathbf{y} = \mathbf{v} + \mathbf{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \quad (8)$$

wherein \mathbf{y} is P dimensional vector of manifest variables, \mathbf{v} is a P dimensional column vector of manifest intercepts, $\mathbf{\Lambda}$ is a $P \times Q$ factor loading matrix, $\boldsymbol{\eta}$ is a Q dimensional

1) The complete code for our examples can be found in the [Supplementary Materials](#) section.

column vector of latent variables; and ϵ is P dimensional column vector of residuals. L-RM-ANOVA rewrites Equation 2 as a special case of Equation 8, wherein π will be estimated as the set of latent variables.

In particular, L-RM-ANOVA sets $P = Q$, $\mathbf{v} = \mathbf{0}$, and $\epsilon = \mathbf{0}$ (i.e., there are as many latent variables as manifest variables, and the latent variables fully account for the means, variances, and covariances of the manifest variables), such that

$$\mathbf{y} = \Lambda\boldsymbol{\eta}. \quad (9)$$

Equation 9 is similar to Equation 2 (i.e., $\boldsymbol{\pi} = \mathbf{C}\mathbf{y}$), when conceptualizing $\boldsymbol{\pi}$ as a set of latent variables (i.e., $\boldsymbol{\pi} = \boldsymbol{\eta}$), and \mathbf{C} as a matrix of loadings ($\mathbf{C} = \Lambda$). Conceptually, multiplying both sides of Equation 2 by the inverse of \mathbf{C} should yield an identical form as the SEM measurement model in Equation 9 (i.e., $\mathbf{y} = \mathbf{C}^{-1}\boldsymbol{\pi}$). However, \mathbf{C} cannot be inverted because it is not a square matrix (currently $R \times (R - 1)$). L-RM-ANOVA augments \mathbf{C} by concatenating a row that contains the constant $1/\sqrt{R}$ to the top of the matrix; creating an invertible matrix \mathbf{C}^* that maintains orthogonality. Following the guiding example,

$$\mathbf{C}^* = \begin{bmatrix} 0.4082 & 0.4082 & 0.4082 & 0.4082 & 0.4082 & 0.4082 \\ -0.4082 & -0.4082 & -0.4082 & 0.4082 & 0.4082 & 0.4082 \\ -0.5000 & 0.0000 & 0.5000 & -0.5000 & 0.0000 & 0.5000 \\ 0.2887 & -0.5774 & 0.2887 & 0.2887 & -0.5774 & 0.2887 \\ 0.5000 & 0.0000 & -0.5000 & -0.5000 & 0.0000 & 0.5000 \\ -0.2887 & 0.5774 & -0.2887 & 0.2887 & 0.5774 & 0.2887 \end{bmatrix}. \quad (10)$$

In general, the concatenated row can be any vector that is pairwise linearly independent to the other rows. However, the way we chose this row, it can be interpreted as an intercept that contributes to each of the manifest variables. The mean of this contrast equals the mean across all of the dependent variables multiplied by the constant R/\sqrt{R} . Replacing \mathbf{C} with \mathbf{C}^* into Equation 2 and solving for \mathbf{y} yields

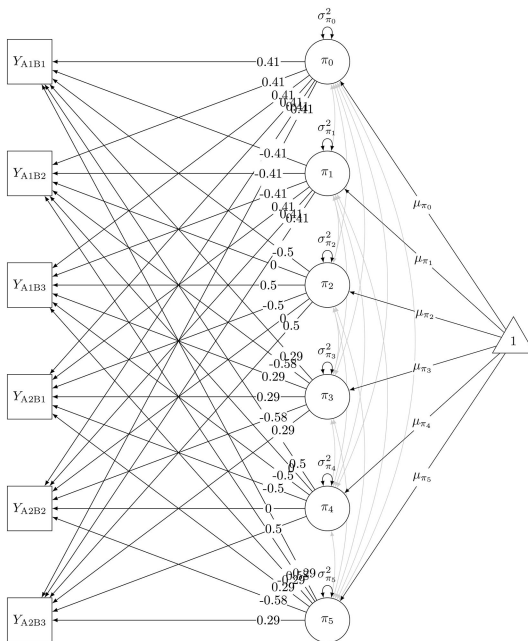
$$\mathbf{y} = \mathbf{C}^{*-1}\boldsymbol{\pi} \quad (11)$$

which follows the structure of the SEM measurement model in Equation 9. Equation 11 demonstrates how L-RM-ANOVA identifies latent contrasts: augmenting the orthogonal contrast matrix \mathbf{C} to create \mathbf{C}^* , and using the inverted version of \mathbf{C}^* as a factor loading matrix in SEM. Estimating Equation 11 as an SEM estimates \mathbf{V}_π , because the means, variances, and covariances of latent variables may be freely estimated. Figure 2 depicts the model prescribed in Equation 11 as applied to the guiding example. Rectangles represent the manifest variables \mathbf{y} and the circles represent the latent contrast variables $\boldsymbol{\pi}$. The weights of the arrows going from $\boldsymbol{\pi}$ to \mathbf{y} can be found in the \mathbf{C}^{*-1} matrix. Importantly, the manifest intercepts and residual variances are forced to equal zero; the means/intercepts,

variances, and covariances of π are freely estimated; and an additional latent variable (i.e., π_0) emerges because C^* contains an extra row relative to C (i.e., the intercept). The latent variables π_1 – π_5 have an identical interpretation as their original interpretation in the section [Review of Orthogonal Contrast Matrices](#), and the new latent variable π_0 indicates a mean across the R repeated measures. Next, this article demonstrates how to test sphericity using L-RM-ANOVA in the SEM framework.

Figure 2

Path Diagram of the SEM Implementing a 2 × 3 (Sentence Type × Grade) Repeated Measures Design Using an Orthonormal Contrast Matrix



Note. Rectangles represent the manifest variables y and the circles represent the latent contrast variables π . The weights of the arrows going from π to y can be found in the C^{*-1} matrix. Intercepts and residual (co)variances of the manifest variables y are set to zero. Intercepts and (co)variances of the contrast variables π are freely estimated.

Testing Sphericity Using L-RM-ANOVA

Sphericity for a given main/interaction effect may be tested by performing a χ^2 -difference test across a model that assumes sphericity for the effect versus a model that does not. Following the guiding example, sphericity for the main effect of B and the interaction may be tested (sphericity is not relevant for the main effect of A because it only has one degree of freedom; see the section [Sphericity for Main and Interaction Effects](#)

of U-RM-ANOVA). Sphericity for the main effect of B prescribes that both $\sigma_{\pi_2}^2 = \sigma_{\pi_3}^2$ and $\sigma_{\pi_2, \pi_3} = 0$; whereas sphericity for the interaction effect prescribes $\sigma_{\pi_4}^2 = \sigma_{\pi_5}^2$, and $\sigma_{\pi_4, \pi_5} = 0$ (see the section [Sphericity for Main and Interaction Effects of U-RM-ANOVA](#) for details). The constrained covariance matrices V_{π_B} and $V_{\pi_{A \times B}}$ are given by:

$$V_{\pi_B} = \begin{matrix} & \pi_0 & \pi_1 & \pi_2 & \pi_3 & \pi_4 & \pi_5 \\ \begin{matrix} \pi_0 \\ \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \\ \pi_5 \end{matrix} & \left[\begin{array}{cccccc} & & & & & \\ & & & & & \\ & & & & & \\ & & & \sigma_B^2 & 0 & \\ & & & 0 & \sigma_B^2 & \\ & & & & & \\ & & & & & \end{array} \right] & , & V_{\pi_{A \times B}} = \begin{matrix} & \pi_0 & \pi_1 & \pi_2 & \pi_3 & \pi_4 & \pi_5 \\ \begin{matrix} \pi_0 \\ \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \\ \pi_5 \end{matrix} & \left[\begin{array}{cccccc} & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \sigma_{A \times B}^2 & 0 \\ & & & & & 0 & \sigma_{A \times B}^2 \end{array} \right] \end{matrix} \end{matrix} \quad (12)$$

where σ_B^2 is the equal variance across π_2 and π_3 , and $\sigma_{A \times B}^2$ is the equal variance across π_4 and π_5 , and omitted cells are freely estimated. In SEM, models with these constraints will be compared against models that do not place any constraints at all. [Table 1](#) presents the results of both Mauchly's test of sphericity and the L-RM-ANOVA based χ^2 -difference tests. The results are virtually identical across the two approaches; providing evidence that sphericity does not hold for either effect.

Table 1

Mauchly's Test and χ^2 -Difference Test for Sphericity

Effect	Mauchly's Test			χ^2 -Difference Test		
	χ^2	df	<i>p</i>	$\Delta\chi^2$	df	<i>p</i>
B	69.81	2	<.001	70.65	2	<.001
A × B	72.41	2	<.001	73.28	2	<.001
B + A × B				112.51	4	<.001

As mentioned earlier, it is also possible to test both sphericity assumptions simultaneously. That is, we can perform an omnibus test that tests for sphericity of in the effect of B and the interaction effect of A and B. This omnibus test can decrease the Type I error rate that can arise due to multiple testing. The results for the omnibus test can also be found in [Table 1](#).

Simulation 1: Comparing Mauchly's Sphericity Test to the SEM Based Test

In the previous sections, we revisited the original definition of sphericity and showed that SEM can be used to test for sphericity in multi-factorial experimental designs. We further compared Mauchly's Sphericity Test to the SEM based test using an empirical example, where both tests showed very similar results. It is, however, important to know the statistical properties across different settings, for instance, sample sizes and degrees of departure from sphericity. In this section, we will conduct a small-scale simulation study to compare both tests. The aim of this study is to guide applied researchers to decide in which situation which test is to be preferred.

Method

We generated data for a 2×3 repeated measures design following the model in [Figure 2](#). We investigated Type 1 error and power of Mauchly's Test and the SEM based test to detect departures from sphericity for the main effect of Factor B which has three levels (i.e., the effect consists of two contrasts). For the data generation, we set the means of all contrasts to zero ($\mu_{\pi_i} = 0$) and the variances to one ($\sigma_{\pi_i}^2 = 1$). We manipulated the degree of departure from sphericity in terms of Mauchly's W , where $W = 1$ indicates no departure and $W = 0$ indicates the largest possible departure. We used four values of W which imply a certain value for the covariances between the contrast variables ($W = 1 \Rightarrow \sigma_{\pi_i, \pi_j} = 0$, $W = 0.8 \Rightarrow \sigma_{\pi_i, \pi_j} = 0.45$, $W = 0.6 \Rightarrow \sigma_{\pi_i, \pi_j} = 0.63$, $W = 0.4 \Rightarrow \sigma_{\pi_i, \pi_j} = 0.77$). All covariances were set to this value, although we would have only had to manipulate σ_{π_2, π_3} and σ_{π_2, π_3} in order to impose different degrees of departure from sphericity on the main effect of B and the interaction effect of A and B (i.e., the other covariances are not important and manipulating them does not do any harm). The four conditions were chosen to cover the full range of possible values of W . We further manipulated the sample size ($N = 30, 40, 50, 60, 70, 80, 90, 100$). The smallest sample size was chosen because it is close to $N = 27$ which is the smallest sample size required to estimate the model (i.e., we estimate 6 means, 6 variances, and 15 covariances). The larger sample sizes were chosen to give a clear picture when the tests reach a power of at least .8. An overview of all conditions is shown in [Table 4](#). We used 1,000 replications for each of the aforementioned conditions. The simulation was performed using the statistical software R ([R Core Team, 2021](#)) in combination with the car package ([Fox & Weisberg, 2019](#)) for Mauchly's test and Mplus ([L.K. Muthén & Muthén, 2017](#)) for the SEM based test. Finally, we chose an alpha level of $\alpha = .05$ for both tests.

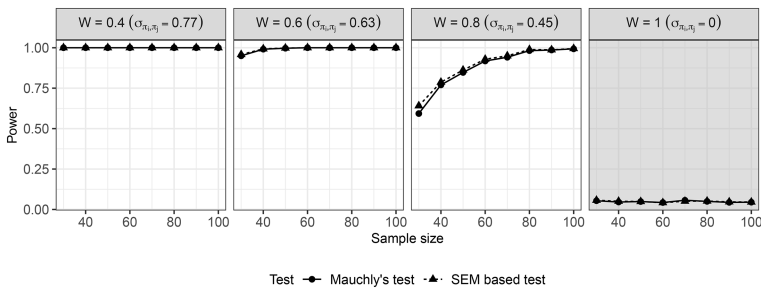
Results and Discussion

The simulation results are shown in [Figure 3](#). The leftmost tile of the figure shows the Type 1 error for both tests when sphericity is *not* violated. Mauchly's Test as well as the

SEM based test show a Type 1 error around the desired 5%. Neither of the tests seem to have an overly inflated Type 1 error. The other tiles show the statistical power of both tests. We can see that the power increases with increasing sample size and increasing departure from sphericity. Both tests show very similar power across all conditions suggesting that SEM is a viable alternative to Mauchly's Test.

Figure 3

Power and Type 1 Error for Mauchly's Sphericity Test and the SEM Based Test as a Function of Sample Size N and Degree of Departure From Sphericity (Mauchly's W)



Note. The first grayed tile ($W = 1$) shows the *Type 1 error*. The other tiles show the *power*.

Testing Main and Interaction Effects of U-RM-ANOVA Using L-RM-ANOVA

A given main or interaction effect may be tested by performing a χ^2 -difference test across two models: a constrained model in which the means belonging to a particular effect are fixed to zero (i.e., conforming to the null hypothesis), and a second unconstrained model in which the means are freely estimated (i.e., conforming to the alternative hypothesis). But, to adhere to the assumptions of U-RM-ANOVA, *both models* used in χ^2 -difference test must impose sphericity.

Following the guiding example, the null hypothesis for the main effect of A prescribes $\mu_{\pi_1} = 0$; the null hypothesis for the main effect of B prescribes both $\mu_{\pi_2} = 0$ and $\mu_{\pi_3} = 0$; and the null hypothesis for interaction effect prescribes both $\mu_{\pi_4} = 0$ and $\mu_{\pi_5} = 0$ (see the section [Review of Orthogonal Contrast Matrices](#) for details). Therefore, testing the main effect of A compares a model that constrains $\mu_{\pi_1} = 0$, versus a model that freely estimates μ_{π_1} (without any extra constraints for the sphericity assumption). Testing the main effect of B compares a model that constrains $\mu_{\pi_2} = 0$ and $\mu_{\pi_3} = 0$, versus a model that freely estimates μ_{π_2} and μ_{π_3} (while *both* models impose sphericity for the main effect of B). Testing the interaction effect compares a model that constrains $\mu_{\pi_4} = 0$ and $\mu_{\pi_5} = 0$, versus a model that freely estimates μ_{π_4} and μ_{π_5} (while *both* models impose sphericity for the interaction effect).

Table 2 displays the relevant test statistics, and their p -values, for the main and interaction effects from U-RM-ANOVA and L-RM-ANOVA. Importantly, the product of an F -value and its numerator degrees of freedom (i.e., df_{num}) is asymptotically equivalent to the χ^2 -difference value from L-RM-ANOVA (i.e., $F_{\text{value}} \times df_{\text{num}} \approx \chi^2 \Leftrightarrow F_{\text{value}} \approx \chi^2 / df_{\text{num}}$, e.g., Fahrmeir et al., 2013; Kohler, 1982; Lu & Zhang, 2010); and the approximate F -values are given from L-RM-ANOVA using that transformation.

Table 2

Results for Main and Interaction Effects

Effect	U-RM-ANOVA				L-RM-ANOVA Sphericity						L-RM-ANOVA			
	F	df_1	df_2	p	p^{GG}	p^{HF}	$\Delta\chi^2$	df_1	$\approx F$	p	$\Delta\chi^2$	df_1	$\approx F$	p
A	371.90	1	167	<.001			196.82	1	196.82	<.001	196.82	1	198.87	<.001
B	733.48	2	334	<.001	<.001	<.001	566.14	2	283.07	<.001	318.45	2	159.22	<.001
A \times B	431.78	2	334	<.001	<.001	<.001	429.04	2	214.52	<.001	249.59	2	124.79	<.001

Note. GG = Greenhouse-Geisser corrected p -value. HF = Huynh-Feldt corrected p -value.

The test statistics and p -values differ across U-RM-ANOVA and L-RM-ANOVA (as opposed to the tests of sphericity from the previous sub-section). The difference across the approaches arises because U-RM-ANOVA bases its test statistics (and p -values) on F -ratios (and F -distributions), whereas L-RM-ANOVA uses χ^2 -differences (and compares those values to χ^2 -distributions). Therefore, discrepancies may arise across the approaches, even though they are designed to test the same hypotheses.

To narrow the gap, the following sub-section (after the excursus) shows how to reproduce both the sums of squares and the F -values from U-RM-ANOVA using the parameter estimates of L-RM-ANOVA. Furthermore, p -values from U-RM-ANOVA may then be more closely reproduced using L-RM-ANOVA by comparing the reproduced F -values to an F distribution.

Excursus: Interpreting Main Effects in the Presence of Interaction Effects

Although not the main focus of this article, we would like to briefly pick up on the discussion about interpreting main effects in the presence of interaction effects. We find it important to note that point estimates of the main effect of sentence type should be interpreted with care. That is, the estimate of the average difference in gaze duration between regular and Landolt sentences across grades is the unweighted average of the conditional effects of sentence type on gaze duration given different grades. This may not be the effect that researchers are interested in. As many researchers have argued in the past, the effect of an independent variable (sentence type) on a dependent variable (gaze duration) is dependent on the moderator (grade) in the presence of an interaction effect (e.g., Aguinis, 2004; Aguinis et al., 2016; Aiken & West, 1991; Busenbark et al., 2021;

Cohen et al., 2003). This phenomenon can also be observed in Figure 1. An alternative approach may be to express the effect of the sentence type on gaze duration as a function of the grade and to look at the conditional effects. This enables us to look at the difference between regular and Landolt sentences in different grades separately. Still, we believe that aggregated or average effects can add valuable additional information in some contexts (see also the discussion in Gräfe et al., 2022).

Calculating Sums of Squares, Mean Squares, and F-Ratios Using L-RM-ANOVA

As noted earlier in this article (section [Review of Orthogonal Contrast Matrices](#)), orthonormal contrast matrices provide the opportunity to reproduce the sums of squares, mean squares, and F -ratios for each effect from U-RM-ANOVA (Voelkle, 2007). This sub-section shows how to reproduce each of these components using L-RM-ANOVA, and Table 3 shows the estimates across U-RM-ANOVA and L-RM-ANOVA. Importantly, all sums of squares, mean squares, and F -ratios are formed from an L-RM-ANOVA that both creates latent contrasts using an orthonormal matrix, and imposes sphericity. The general formulas to calculate the sums of squares are:

$$SS = \sum \mu_{\tau_i}^2 \times N \quad (13)$$

$$RSS = \sum \sigma_{\tau_i}^2 \times N \quad (14)$$

$$MS = SS/df_{\text{num}} \quad (15)$$

$$MRS = RSS/((N - 1) \times df_{\text{num}}) \quad (16)$$

$$F = MS/MSR. \quad (17)$$

Table 3

Sums of Squares

Effect	U-RM-ANOVA					L-RM-ANOVA				
	SS	RSS	MS	MSR	F	SS	RSS	MS	MSR	F
A	184.09	82.67	184.09	0.50	371.90	184.09	82.67	184.09	0.50	371.90
B	210.45	47.92	105.23	0.14	733.48	210.45	47.92	105.23	0.14	733.48
A × B	87.28	33.76	43.64	0.20	431.78	87.28	33.76	43.64	0.20	431.78

The sums of squares for a given effect from U-RM-ANOVA (written as, e.g., SS_A , SS_B , or $SS_{A \times B}$) may be reproduced in L-RM-ANOVA by summing across the squared

means of the effect's latent contrasts, and multiplying the sum by the sample size (Equation 13). Following the guiding example, $SS_A = \mu_{\pi_1}^2 \times N$; $SS_B = (\mu_{\pi_2}^2 + \mu_{\pi_3}^2) \times N$; and $SS_{A \times B} = (\mu_{\pi_4}^2 + \mu_{\pi_5}^2) \times N$ ($N = 169$ in the guiding example).

The residual sums of squares for a given effect from U-RM-ANOVA (written as, e.g., RSS_A , RSS_B , or $RSS_{A \times B}$) may be reproduced in L-RM-ANOVA by summing across the variances of the effect's latent contrasts, and multiplying the sum by the sample size (Equation 14). Following the guiding example, $RSS_A = \sigma_{\pi_1}^2 \times N$, $RSS_B = (\sigma_{\pi_2}^2 + \sigma_{\pi_3}^2) \times N$, and $RSS_{A \times B} = (\sigma_{\pi_4}^2 + \sigma_{\pi_5}^2) \times N$ (again, $N = 169$ in the guiding example).

The mean squares for a given effect from U-RM-ANOVA (written as, e.g., MS_A , MS_B , or $MS_{A \times B}$) may be reproduced in L-RM-ANOVA by dividing the effect's sums of squares by its numerator degrees of freedom (i.e., the number of contrasts that underlies the effect, Equation 15), such that $MS_A = SS_A / df_{A_{\text{num}}}$; $MS_B = SS_B / df_{B_{\text{num}}}$; and $MS_{A \times B} = SS_{A \times B} / df_{A \times B_{\text{num}}}$. In the guiding example, $df_{A_{\text{num}}} = 1$, $df_{B_{\text{num}}} = 2$, and $df_{A \times B_{\text{num}}} = 2$.

The mean squares of the residuals for a given effect from U-RM-ANOVA (written as, e.g., MSR_A , MSR_B , or $MSR_{A \times B}$) may be reproduced in L-RM-ANOVA by dividing each effect's residual sums of squares by the product of $N - 1$ and the effect's numerator degrees of freedom (Equation 16). Following the guiding example, we have $MSR_A = RSS_A / ((N - 1) \times df_{A_{\text{num}}})$; $MSR_B = RSS_B / ((N - 1) \times df_{B_{\text{num}}})$; and $MSR_{A \times B} = RSS_{A \times B} / ((N - 1) \times df_{A \times B_{\text{num}}})$.

F -values from U-RM-ANOVA may be reproduced in L-RM-ANOVA by dividing the effect's mean squares by its mean squared residuals (Equation 17). Following the guiding example leads to $F_A = MS_A / MSR_A$, $F_B = MS_B / MSR_B$, and $F_{A \times B} = MS_{A \times B} / MSR_{A \times B}$. As shown in Table 3, all values are virtually identical across U-RM-ANOVA and L-RM-ANOVA; and therefore researchers may compute F and p -values that closely match those from U-RM-ANOVA using L-RM-ANOVA.

Finally, we would like to note that a major advantage of SEM is that sums of squares can also be calculated in the presence of missing values. Parameter estimates can be obtained through full information maximum likelihood, which are then used to calculate sums of squares as shown above. The above calculations can further be extended to mixed within- and between-subjects designs with any number of factors. We refer the interested reader to Langenberg et al. (2022), which includes instructions in the appendix to calculate F -values and the effect size measure η_p^2 for larger within- and between-subjects designs.

Testing Main and Interaction Effects Without the Assumption of Sphericity Using L-RM-ANOVA

In contrast to U-RM-ANOVA, L-RM-ANOVA may test main/interaction effects without the assumption of sphericity. The model comparison procedure described above (i.e., χ^2 -difference tests described in the section Testing Main and Interaction Effects of U-RM-ANOVA Using L-RM-ANOVA) can relax sphericity by estimating all elements of V_π in

both models (i.e., removing the constraints that describe sphericity). The right part of [Table 2](#) provides estimates of main and interaction effects from L-RM-ANOVA when estimated without the assumption of sphericity.

Currently, U-RM-ANOVA relies on post-hoc corrections to relax sphericity (e.g., Greenhouse-Geisser and Huynh-Feldt corrections), which computationally (and conceptually) differ from direct estimation of V_{π} as performed via L-RM-ANOVA. [Table 2](#) also provides p -values associated with these common post-hoc corrections.

Imposing the assumption of sphericity can increase power of hypothesis tests in U-RM-ANOVA. This is true for any type of assumption imposed to statistical models as fewer parameters need to be estimated if the assumptions is true. However, if the assumption, in fact, does not hold, an increased Type I error rate may arise (e.g., [Haverkamp & Beauducel, 2017](#)).

Simulation 2: Comparing U-RM-ANOVA and L-RM-ANOVA With and Without Sphericity

In the previous two sections, we compared hypothesis tests of U-RM-ANOVA and L-RM-ANOVA with and without the assumption of sphericity. Both approaches yield very similar F -values and p -values. It remains to be answered whether any of the approaches outperforms the others in terms of statistical properties. For instance, L-RM-ANOVA is estimated through maximum likelihood which is known to have an inflated Type 1 error in small samples (e.g., [Green & Babyak, 1997](#); [Hu et al., 1992](#); [Muthén & Kaplan, 1985](#); [Raykov & Widaman, 1995](#)). U-RM-ANOVA, in contrast, is said to have a power advantage but should also have an inflated Type 1 error when sphericity does not hold. In this section, we will examine the statistical properties of the aforementioned approaches across several settings. We have two main hypotheses: (1) We expect L-RM-ANOVA to have an inflated Type 1 error when testing main and interaction effects in small samples as compared to RM-ANOVA because it is estimated through maximum likelihood which relies on asymptotic theory, and (2) we also expect the models that assume sphericity to have an inflated Type 1 error when testing main and interaction effects and larger bias of effect size estimates when sphericity does not hold.

Method

We generated data for a 2×3 repeated measures design following the model in [Figure 2](#). We investigated Type 1 error, power, bias and root mean squared error (RMSE) of multivariate repeated measures ANOVA (RM-ANOVA), U-RM-ANOVA (with and without Greenhouse-Geisser and Huynh-Feldt corrections), L-RM-ANOVA (with and without assuming sphericity) for the test of the main effect of the Factor B which has three levels (i.e., the effect consisted of two contrasts). We manipulated the degree of departure from sphericity W (i.e., we again set the variances to one and only manipulated the covariances), sample size N , and the effect size η_p^2 . In particular, we again used

four degrees of departure from sphericity ($W = 1 \Rightarrow \sigma_{\pi_i, \pi_j} = 0$, $W = 0.8 \Rightarrow \sigma_{\pi_i, \pi_j} = 0.45$, $W = 0.6 \Rightarrow \sigma_{\pi_i, \pi_j} = 0.63$, $W = 0.4 \Rightarrow \sigma_{\pi_i, \pi_j} = 0.77$), and eight sample sizes, ($N = 30, 40, 50, 60, 70, 80, 90, 100$). We further used four effect sizes ($\eta_p^2 = 0, 0.01, 0.06, 0.14$). η_p^2 is a common effect size measure for repeated measures ANOVA (e.g. Keselman et al., 1998; Maxwell et al., 2008; Olejnik & Algina, 2000; Perugini et al., 2018; Steiger, 2004), where $\eta_p^2 = 0$ indicates no effect is present and the other three choices represent a small, medium, and large effect according to Cohen (1988). We imposed a particular effect size by setting the means of the two regarding contrast variables (μ_{π_2} and μ_{π_3}) to a certain value. Since means, variances and the covariance of the contrast variables contribute to the effect size, the two means were chosen in a way that accounted for the degree of departure from sphericity (i.e., the two means were different for different W s even if the effect size was the same). An overview of all conditions is shown in Table 4.

Table 4

Conditions Used in the Two Simulation Studies

Simulation study	Conditions		Population parameters			
	η_p^2	W	μ_{π_0}, μ_{π_1} ^a	$\mu_{\pi_2}, \mu_{\pi_3}, \mu_{\pi_4}, \mu_{\pi_5}$ ^b	$\sigma_{\pi_1}^2$	σ_{π_i, π_j}
1, 2	0	0.4	0.00	0.00	1	0.77
1, 2	0	0.6	0.00	0.00	1	0.63
1, 2	0	0.8	0.00	0.00	1	0.45
1, 2	0	1.0	0.00	0.00	1	0
2	0.01	0.4	0.10	0.09	1	0.77
2	0.01	0.6	0.10	0.09	1	0.63
2	0.01	0.8	0.10	0.09	1	0.45
2	0.01	1.0	0.10	0.07	1	0
2	0.06	0.4	0.25	0.24	1	0.77
2	0.06	0.6	0.25	0.23	1	0.63
2	0.06	0.8	0.25	0.21	1	0.45
2	0.06	1.0	0.25	0.18	1	0
2	0.14	0.4	0.40	0.38	1	0.77
2	0.14	0.6	0.40	0.36	1	0.63
2	0.14	0.8	0.40	0.34	1	0.45
2	0.14	1.0	0.40	0.29	1	0

Note. The first simulation study used only the first four conditions. The second simulation study used all conditions. We used eight different samples sizes ($N = 30, 40, 50, 60, 70, 80, 90, 100$) which have been omitted to reduce the size of the table.

^aMeans of contrast variables that belong to an effect with *one* degree of freedom (i.e., intercept and main effect of A). ^bMeans of contrast variables that belong to an effect with *two* degrees of freedom (i.e., main effect of B and interaction effect of A and B).

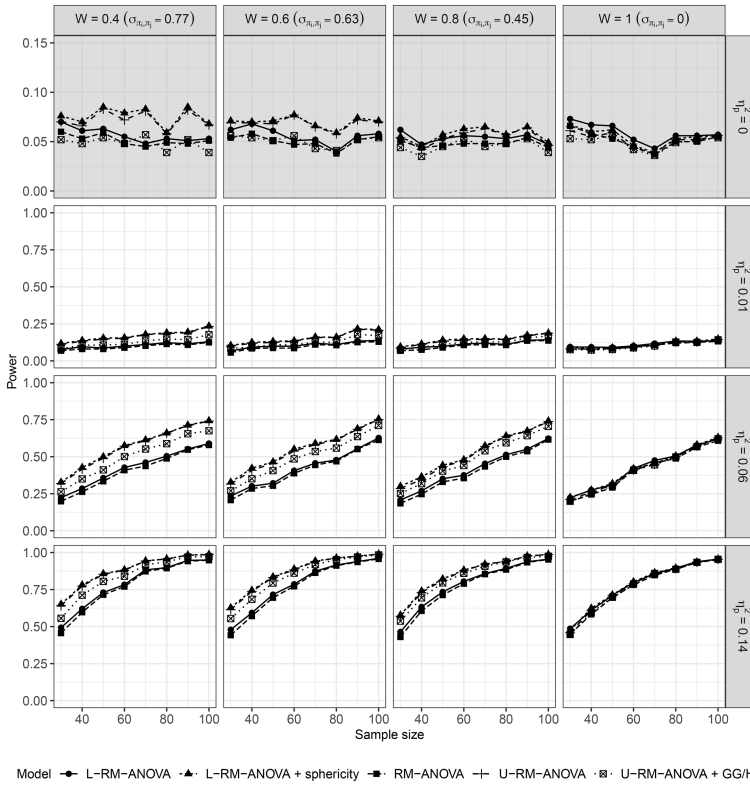
We used 1,000 replications for each of the aforementioned conditions. The simulation was performed using the statistical software R (R Core Team, 2021) in combination with the car package (Fox & Weisberg, 2019) for RM-ANOVA and U-RM-ANOVA, and Mplus (Muthén & Muthén, 2017) for the SEM based models. Finally, we chose an alpha level of $\alpha = .05$ for all of the performed hypothesis tests.

Results and Discussion

Power — Hypothesis tests were based on an F -test and the sums of squares for RM-ANOVA models, and based on a likelihood ratio test for SEMs. The Greenhouse-Geisser and the Huynh-Feldt corrections showed virtually identical results, which is why we will summarize both corrections under U-RM-ANOVA + GG/HF. We thus compared the statistical power and Type 1 error of five models, namely L-RM-ANOVA, L-RM-ANOVA + sphericity, RM-ANOVA, U-RM-ANOVA, and U-RM-ANOVA + GG/HF. The results are shown in Figure 4, where the first row shows Type 1 error ($\eta_p^2 = 0$) and the other rows show power ($\eta_p^2 > 0$). As hypothesized, the SEM based models showed a slightly inflated Type 1 error of up to 7.3% for the small sample size condition $N = 30$ and when the simulated effect size was zero (as shown in the first row of the figure). The Type 1 error was further inflated of up to 8.5% for all models that mistakenly assumed sphericity when the assumption did not hold (first row, the three right-hand tiles). The Type 1 error decreased with increasing sample size for the SEM based models, but seemed to be constant across sample sizes for univariate models that incorrectly assume sphericity. Furthermore, power increased with sample size and effect size for all models. The multivariate approaches (RM-ANOVA and L-RM-ANOVA) were unaffected from departure from sphericity in terms of power. The univariate models (U-RM-ANOVA and L-RM-ANOVA + sphericity) showed higher power by up to 15.8% as compared to multivariate models, particularly when sphericity was strongly violated $W = 0.4$ and the sample size was small $N = 30$. We argue that this presumed power “advantage” is bought from the inflated Type 1 error and should not be trusted. The corrected univariate model (U-RM-ANOVA + GG/HF) also shows a slight power advantage (by 6.3% with $W = 0.4$ and $N = 30$) which we think can be trusted as the model does not show Type 1 error inflation.

Figure 4

Power and Type 1 Error for RM-ANOVA, U-RM-ANOVA, L-RM-ANOVA and L-RM-ANOVA Assuming Sphericity as a Function of Effect Size η_p^2 , Sample Size N , and Degree of Departure From Sphericity (Mauchly's W)



Note. The first grayed row ($\eta_p^2 = 0$) shows the Type 1 error. The other rows show the power.

As mentioned in the beginning of the previous paragraph, hypotheses are tested by different statistical tests in RM-ANOVA models and SEMs. Those tests can differ in terms of power and Type 1 error rate. However, it is also possible to derive the sums of squares and an F -test based on the point estimates of means, variances and covariances from SEM (see [Calculating Sums of Squares, Mean Squares, and F-Ratios Using L-RM-ANOVA](#)). The resulting test would have the same statistical properties as the F -test of RM-ANOVA. We limited comparison, however, to the classical tests (i.e., F -tests for RM-ANOVA and likelihood ratio test for SEM) because they are most common in the two frameworks.

Bias — Relative and absolute bias of the estimated effect size $\hat{\eta}_p^2$ was identical across the univariate models (L-RM-ANOVA + sphericity and U-RM-ANOVA), and also across the multivariate models (L-RM-ANOVA and RM-ANOVA). This pattern is not surprising

because exact sums of squares (which effect size estimates rely on) can exactly be derived in SEM (see [Calculating Sums of Squares, Mean Squares, and F-Ratios Using L-RM-ANOVA](#)). The results are shown in Figure 5, where the first row shows the

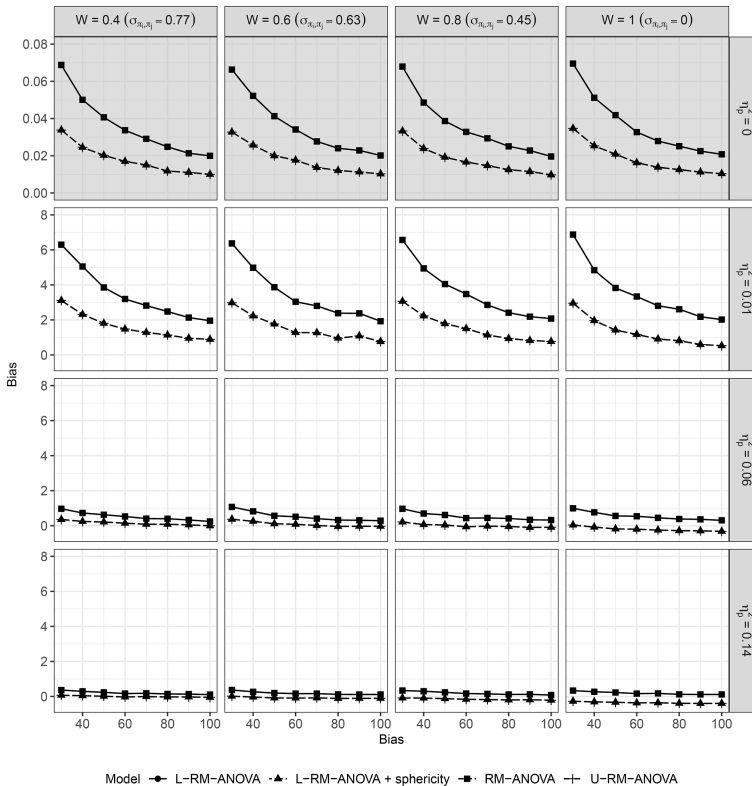
$$\text{absolute bias: } \frac{1}{n} \sum_{i=1}^n \hat{\eta}_{p_i}^2 - \eta_p^2$$

(we cannot divide by $\eta_p^2 = 0$) and the other rows show the

$$\text{relative bias: } \frac{1}{n} \sum_{i=1}^n \frac{\hat{\eta}_{p_i}^2 - \eta_p^2}{\eta_p^2}.$$

Figure 5

Bias for RM-ANOVA, U-RM-ANOVA, L-RM-ANOVA and L-RM-ANOVA Assuming Sphericity as a Function of Effect Size η_p^2 , Sample Size N, and Degree of Departure From Sphericity (Mauchly's W)



Note. The first grayed row ($\eta_p^2 = 0$) shows the absolute bias. The other rows show the relative bias.

In general, relative and absolute bias decreased with increasing sample size and relative bias decreased with effect size. Violations from sphericity did not seem to affect relative bias. Univariate models had a smaller bias as compared to multivariate models. As apposed to our expectation, departures from sphericity did not seem to affect bias for neither the multivariate or the univariate models.

Root Mean Squared Error — As for bias, the relative and absolute RMSE of the estimated effect size $\hat{\eta}_p^2$ was identical across the univariate models (L-RM-ANOVA + sphericity and U-RM-ANOVA), and also across the multivariate models (L-RM-ANOVA and RM-ANOVA). The results are shown in [Figure 6](#), where the first row shows the

$$\text{absolute RMSE: } \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{\eta}_{p_i}^2 - \eta_p^2)^2}$$

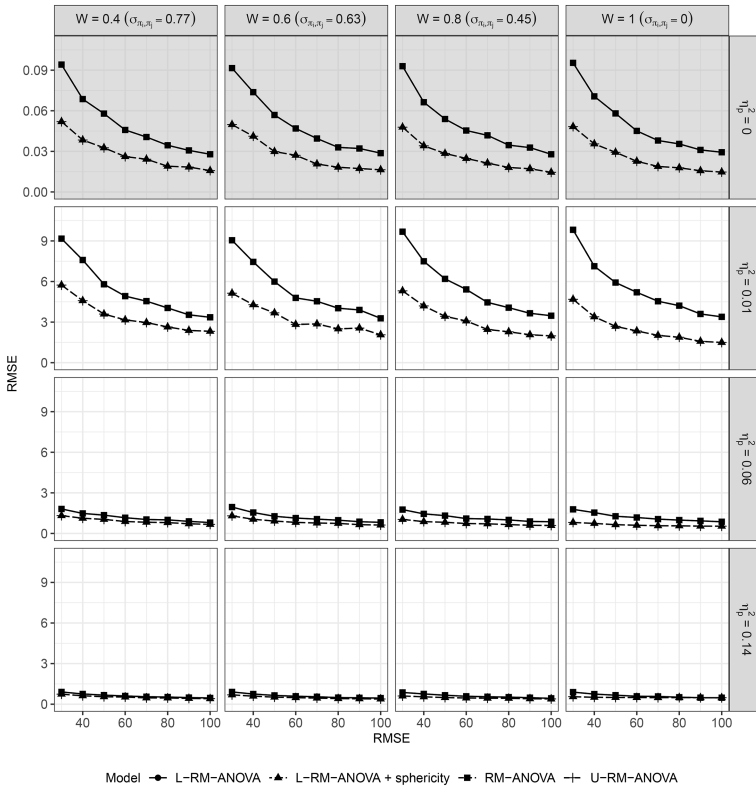
(we cannot divide by $\eta_p^2 = 0$) and the other rows show the

$$\text{relative RMSE: } \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{\eta}_{p_i}^2 - \eta_p^2)^2}}{\eta_p^2}.$$

In general, relative and absolute RMSE decreased with increasing sample size and relative bias decreased with increasing effect size. Again, univariate models showed a smaller RMSE as compared to multivariate models. And also as opposed to our expectation, departures from sphericity did not seem to affect the RMSE for neither the multivariate or the univariate models.

Figure 6

Root Mean Squared Error (RMSE) for RM-ANOVA, U-RM-ANOVA, L-RM-ANOVA and L-RM-ANOVA Assuming Sphericity as a Function of Effect Size η_p^2 , Sample Size N , and Degree of Departure From Sphericity (Mauchly's W)



Note. The first grayed row ($\eta_p^2 = 0$) shows the *absolute* RMSE. The other rows show the *relative* RMSE.

Testing Sphericity, Main Effects, and Interaction Effects for L-RM-ANOVAs With Measurement Models

Traditional (U-)RM-ANOVA assumes that the outcome variable can be observed across experimental conditions. The outcome of interest, however, oftentimes includes questionnaire items, test scores, reaction times, and accuracies that serve as indicators to measure an underlying psychological construct, such as cognitive processes, attention, traits, or attitudes. Underlying constructs, however, cannot be measured directly in many cases and indicators suffer from measurement error. Latent variable models can be used to explicitly model measurement error. This section describes how sphericity,

main effects, and interaction effects may be tested in extensions of L-RM-ANOVA that include measurement models (conceptually similar to second order growth curves; see Langenberg et al., 2020). We will use the two manifest measures mean gaze duration and mean total viewing to measure the latent construct “reading ability” and re-analyze the data.

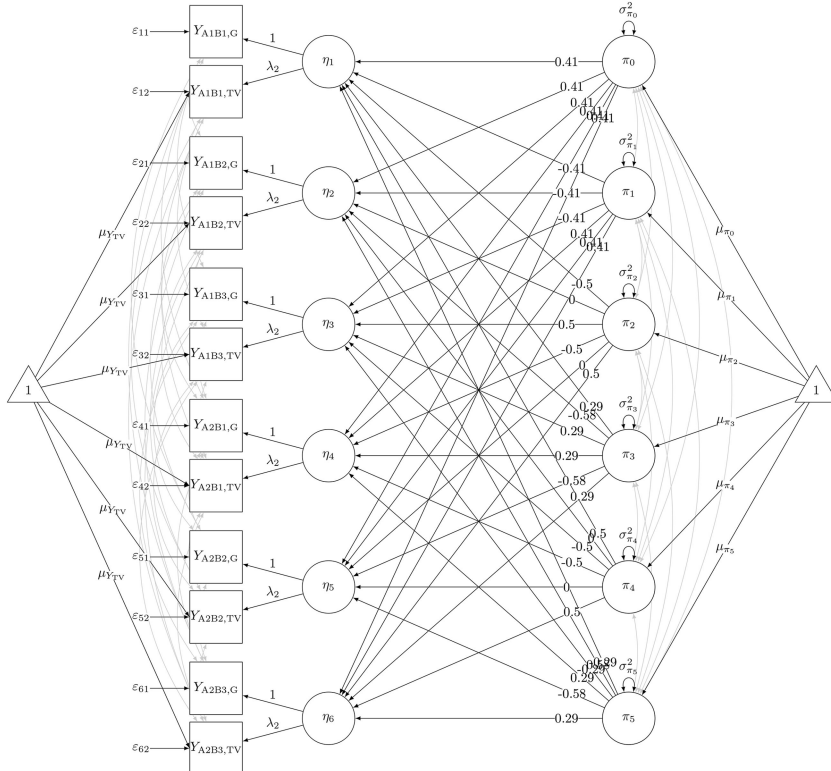
Figure 7 extends the guiding example to include two manifest measures (mean gaze duration and mean total viewing duration, see Figure 1) at each of the six measurement conditions; forming six latent common factors (η_1 – η_6) that are transformed into six latent contrasts (π_0 – π_5). Rectangles represent manifest variables and circles represent latent variables. Following standards for longitudinal models with multiple indicators per measurement occasion (e.g., Newsom, 2015, p. 42), residual covariances are estimated between the same manifest indicator across the six measurement conditions (depicted as gray double-headed arrows). The η variables explain common variance in the manifest variables in each of the conditions. However, residual covariances between manifest variables across conditions may occur. These correlations can be accounted for by including residual covariances between mean gaze duration across conditions and between mean total viewing duration, respectively.

Oftentimes, such designs are inappropriately analyzed by averaging across the two indicators, so that traditional methods can be used (e.g., RM-ANOVA). Averaging across indicators, however, can lead to ignoring other random factors, such as stimuli, and can introduce substantial bias (Judd et al., 2017). As a solution, linear mixed models (LMM; Fitzmaurice et al., 2011; Laird & Ware, 1982) are able to include all of the measures and to estimate the model. LMMs, however, assume that the two indicators are parallel measurements. A parallel measurement model assumes that loadings equal one ($\lambda_2 = 1$), the intercepts equal zero ($\mu_{Y_{TV}} = 0$), and the residual variances are equal ($\text{Var}(\varepsilon_{ij}) = \text{Var}(\varepsilon_{kl})$). This condition, however, does not necessarily hold in the given example and neither in many other examples from psychological research, such as test scores or questionnaire items. In SEM, on the contrary, the assumption can be relaxed and the more general congeneric measurement model can be used (as was used in Figure 7).

Inclusion of measurement models across repeated measures (e.g., second-order growth curves) further necessitates consideration of (and tests for) measurement invariance (e.g., Newsom, 2015, Ch. 2). However, given this article’s focus on sphericity and main/interaction effects, the didactic nature of this section (i.e., we do not aim to explicitly test psychological theories), and available literature on measurement invariance (Newsom, 2015; Pitts et al., 1996; Widaman et al., 2010), this article assumes that readers have knowledge of measurement invariance and does not discuss the topic in detail (see Langenberg et al., 2020, for measurement invariance in L-RM-ANOVA). Instead, we note that data in the guiding example adhered to a model with strong measurement invariance (CFI = .946, TLI = .914, RMSEA = .115, 90% CI RMSEA = [.099, .132]), which facilitates comparisons across the means/intercepts of η_1 – η_6 . That is, loadings are con-

Figure 7

Path Diagram of the SEM Implementing a 2 × 3 (Sentence Type × Grade) Repeated Measures Design Using an Orthonormal Contrast Matrix and a Measurement Model With the Manifest Variables Mean Gaze Duration and Mean Total Viewing Duration



Note. Rectangles represent manifest variables and circles represent latent variables. Intercepts of mean gaze duration in each condition are set to zero. Intercepts of mean total viewing duration are estimated but constrained to be equal. Residual variances of the manifest variables *y* are freely estimated. Residual covariances among the *y* of the same type of sentence and variable are estimated but constrained to be equal. Intercepts and (co)variances of the contrast variables *π* are freely estimated.

strained to be equal across time (i.e., the first loading is fixed to 1 and second loading is the same λ₂) and so are the intercepts of the manifest variables (i.e., the intercept of the first indicator is fixed to 0 and the second indicator is constrained to be equal).

As is standard for SEM estimation more generally (and departing from prior analyses that only used complete data), this section analyzes all available data (*N* = 268) via full information maximum likelihood (FIML). Analogous to measurement invariance, we

assume that readers have sufficient knowledge of FIML and its application in SEM to account for missing data (Enders, 2013).

If strong measurement invariance holds, then sphericity of each main/interaction effect may be tested in an identical manner as L-RM-ANOVA without measurement models. Sphericity may still be tested via χ^2 -difference tests across models that constrain/relax the appropriate variances and covariances of \mathbf{V}_π (see the section [Testing Sphericity Using L-RM-ANOVA](#)). In the guiding example, sphericity for the main effect of B fails ($\Delta\chi^2 = 67.44$, $df = 2$, $p < .001$), and sphericity for the interaction effect also fails ($\Delta\chi^2 = 60.88$, $df = 2$, $p < .001$).

Main and interaction effects may still be tested via χ^2 -difference tests across models that constrain/relax the appropriate means of the π latent contrasts. And those comparisons may be done under the assumption of sphericity (i.e., both models in the comparison include the constraints that conform to sphericity), or not (i.e., both models in the comparison fully estimate \mathbf{V}_π). Although this article does not intend to compare SEM to LMM, we would like to point out that imposing sphericity on the variances of latent contrast variables in an SEM is conceptually similar to constraining the covariance matrix of the random effects in a LMM (for similarities between SEM and LMM in the context of growth curves, see, e.g., Rovine & Molenaar, 1998; see also Newsom, 2002). When assuming sphericity, there is evidence for a main effect of A ($\Delta\chi^2 = 329.54$, $df = 1$, $p < .001$), B ($\Delta\chi^2 = 859.88$, $df = 2$, $p < .001$), and the interaction ($\Delta\chi^2 = 624.81$, $df = 2$, $p < .001$). When relaxing sphericity, there is evidence for a main effect of A ($\Delta\chi^2 = 329.54$, $df = 1$, $p < .001$), B ($\Delta\chi^2 = 505.53$, $df = 2$, $p < .001$), and the interaction ($\Delta\chi^2 = 360.06$, $df = 2$, $p < .001$).

Conclusions and Future Directions

This article identified and exemplified the direct connection between U-RM-ANOVA and SEM: L-RM-ANOVA. More specifically, latent contrasts may be formed by using the inverse of an orthogonal contrast matrix as a factor loading matrix, sphericity corresponds to specific constraints on the variances and covariances of those latent contrasts (i.e., specific elements in \mathbf{V}_π), and tests of main/interaction effects correspond to the significance of latent contrast means. As shown in the examples, sphericity may be imposed, relaxed, and tested in the SEM context via χ^2 -difference tests, and results mirror those from Mauchly's test (see the section [Testing Sphericity Using L-RM-ANOVA](#)). And, although the χ^2 -difference tests of SEM do not exactly match the F -tests from U-RM-ANOVA, they do test the same hypotheses, and sums of squares and exact F -values may be reproduced in L-RM-ANOVA via an orthonormal contrast matrix. Finally, taking full advantage of the SEM framework, the L-RM-ANOVA approach can include measurement models and accommodate missing data via FIML when testing for sphericity, main

effects, and interaction effects. Therefore, this article serves to fill the gap of how U-RM-ANOVA is a special case of SEM.

Two simulation studies were performed: (1) examining the statistical properties of the SEM based sphericity test, and (2) comparing properties of (U-)RM-ANOVA and L-RM-ANOVA with and without sphericity. The first simulation study shows that Mauchly's test and the SEM based test yield virtually identical power and Type 1 error rates. The second simulation study shows that RM-ANOVA and L-RM-ANOVA have similar statistical properties. However, L-RM-ANOVA has a slightly inflated Type 1 error of about 7% for a sample size of $N = 30$, which approaches the desired 5% for larger sample sizes. The univariate approaches also show an inflated Type 1 error of up to 8% when sphericity is violated. Greenhouse-Geisser and Huynh-Feldt corrected hypothesis tests from U-RM-ANOVA, furthermore, perform best in terms of power. Although the power advantage is rather small, power was larger than for multivariate RM-ANOVA, while the Type 1 error was not inflated as in the case of the uncorrected U-RM-ANOVA.

This article also helped illuminate the definition of sphericity. In contrast to the colloquial definition (i.e., equal variances across all pairwise differences; which only holds for the within-subjects design with one factor) this article emphasized the original definition of sphericity via matrix algebra, showed how that definition may be generalized to within-subjects designs with two or more factors, and illustrated how that definition may be implemented/tested in L-RM-ANOVA both with and without measurement models. It is our hope that researchers familiar with SEM—but either lack clear understanding of sphericity, or adhere to the colloquial definition—can use this article to gain a clearer understanding of sphericity.

Although L-RM-ANOVA can be extended to mixed designs (i.e., those with both within- and between-subjects factors), we narrowed the scope of the article to within-subjects designs to emphasize identification and tests of sphericity, and because L-RM-ANOVA with mixed-designs has been described elsewhere (see Langenberg et al., 2020, for examples of mixed designs). Nevertheless, we briefly mention the two extensions for incorporating between-subjects factors into L-RM-ANOVA. First, coded versions of between-subjects factors (e.g., dummy codes, effect codes) may be used to predict the latent contrasts. Second, a multiple group SEM can estimate the L-RM-ANOVA model for each intersection of the between subjects factors. The first approach must assume that V_{π} is equal across the intersections of the between-subjects factors, whereas the multiple group approach can test/relax that assumption.

We discuss two future directions which build on the knowledge generated in this article. First, the L-RM-ANOVA approach can be generalized to non-normal and/or non-continuous dependent variables (e.g., log-normal, dichotomous, or ordinal measures). Researchers implementing experimental designs likely measure such variables. SEM has extensions available for non-normal dependent variables (e.g., Finney & DiStefano, 2013). For instance, analyzing error rates requires a binomial distribution, and reaction times of-

ten follow a skewed distribution (e.g., log-normal). Stated differently, L-RM-ANOVA can capitalize on SEMs extension for non-normal and non-continuous dependent variables, which researchers will likely find useful. Second, the L-RM-ANOVA approach can be compared to linear mixed models. As was described in the section [Testing Sphericity, Main Effects, and Interaction Effects for L-RM-ANOVAs With Measurement Models](#), L-RM-ANOVA allows for relaxing the assumption of a parallel measurement model which, in contrast, is essential for LMM. It would be interesting to see how both approaches perform if this assumption is violated. Following the first future direction, L-RM-ANOVA and LMMs could also be compared for non-normal outcomes.

In conclusion, this article identified and demonstrated the missing link that connects U-RM-ANOVA to SEM (via L-RM-ANOVA), and provided researchers a clear definition of sphericity. We hope that this article enables applied researchers to use SEM in practice (especially in cases that warrant measurement models), and motivates quantitative researchers to continue building on the L-RM-ANOVA framework.

Funding: This work was supported by the German Research Foundation under Grant MA 7702/1-2.

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors have declared that no competing interests exist.

Related Versions: This paper is partially based on the doctoral dissertation of the corresponding author Benedikt Langenberg, which is available on the website of the Universität Bielefeld's Universitätsbibliothek at: <https://doi.org/10.4119/unibi/2963576>.

Data Availability: Data is freely available at [Supplementary Materials](#).

Supplementary Materials

The supplementary materials provided are the preprint and code. They can be accessed at the [Index of Supplementary Materials](#) below).

Index of Supplementary Materials

Langenberg, B., Helm, J. L., Günther, T., & Mayer, A. (2022). *Supplementary materials to "Understanding, testing, and relaxing sphericity of repeated measures ANOVA with manifest and latent variables using SEM"* [Preprint, code]. OSF. <https://osf.io/87jnj/>

References

Aguinis, H. (2004). *Regression analysis for categorical moderators*. Gilford Press.

- Aguinis, H., Edwards, J. R., & Bradley, K. J. (2016). Improving our understanding of moderation and mediation in strategic management research. *Organizational Research Methods, 20*(4), 665–685. <https://doi.org/10.1177/1094428115627498>
- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. SAGE.
- Busenbark, J. R., Graffin, S. D., Campbell, R. J., & Lee, E. Y. (2021). A marginal effects approach to interpreting main effects and moderation. *Organizational Research Methods, 25*(1), 147–169. <https://doi.org/10.1177/1094428120976838>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. <https://doi.org/10.4324/9780203771587>
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Routledge.
- Enders, C. K. (2013). Analyzing structural equation models with missing data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling: A second course* (2nd ed., pp. 493–520). Information Age Publishing.
- Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression*. Springer. <https://doi.org/10.1007/978-3-642-34333-9>
- Field, A. (1998). A bluffer's guide to... sphericity. *Newsletter of the Mathematical, Statistical and Computing Section of the British Psychological Society, 6*(1), 13–22.
- Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. SAGE.
- Finney, S. J., & DiStefano, C. (2013). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller, (Eds.), *Structural Equation Modeling: A second course* (2nd ed., pp. 439–492). Information Age Publishing.
- Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis* (2nd ed.). John Wiley & Sons.
- Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage.
- Gräfe, L., Hahn, S., & Mayer, A. (2022). On the relationship between ANOVA main effects and average treatment effects. *Multivariate Behavioral Research*. <https://doi.org/10.1080/00273171.2022.2068122>
- Green, S. B., & Babyak, M. A. (1997). Control of Type I errors with multiple tests of constraints in Structural Equation Modeling. *Multivariate Behavioral Research, 32*(1), 39–51. https://doi.org/10.1207/s15327906mbr3201_2
- Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*(2), 95–112. <https://doi.org/10.1007/BF02289823>
- Haverkamp, N., & Beauducel, A. (2017). Violation of the sphericity assumption and its effect on Type I error rates in repeated measures ANOVA and multi-level linear models (MLM). *Frontiers in Psychology, 8*, 1–12. <https://doi.org/10.3389/fpsyg.2017.01841>
- Heim, S., von Tongeln, F., Hillen, R., Horbach, J., Radach, R., & Günther, T. (2018). Reading without words or target detection? A re-analysis and replication fMRI study of the Landolt paradigm. *Brain Structure and Function, 223*(7), 3447–3461. <https://doi.org/10.1007/s00429-018-1698-x>

- Hillen, R., Günther, T., Kohlen, C., Eckers, C., van Ermingen-Marbach, M., Sass, K., Scharke, W., Vollmar, J., Radach, R., & Heim, S. (2013). Identifying brain systems for gaze orienting during reading: fMRI investigation of the Landolt paradigm. *Frontiers in Human Neuroscience*, 7, Article 384. <https://doi.org/10.3389/fnhum.2013.00384>
- Hu, L.-t., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, 112(2), 351–362. <https://doi.org/10.1037/0033-2909.112.2.351>
- Huynh, H., & Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact f-distributions. *Journal of the American Statistical Association*, 65(332), 1582–1589. <https://doi.org/10.2307/2284340>
- Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, 68(1), 601–625. <https://doi.org/10.1146/annurev-psych-122414-033702>
- Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68(3), 350–386. <https://doi.org/10.3102/00346543068003350>
- Kohler, D. F. (1982). The relation among the likelihood ratio-, Wald-, and Lagrange multiplier tests and their applicability to small samples. *Rand Paper Series*, P-6756, 1–10.
- Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4), 963–974. <https://doi.org/10.2307/2529876>
- Lane, D. M. (2016). The assumption of sphericity in repeated-measures designs: What it means and what to do when it is violated. *Quantitative Methods for Psychology*, 12(2), 114–122. <https://doi.org/10.20982/tqmp.12.2.p114>
- Langenberg, B., Helm, J. L., & Mayer, A. (2020). Repeated measures ANOVA with latent variables to analyze interindividual differences in contrasts. *Multivariate Behavioral Research*, 57(1), 2–19. <https://doi.org/10.1080/00273171.2020.1803038>
- Langenberg, B., Helm, J. L., & Mayer, A. (2022). Bayesian analysis of multi-factorial experimental designs using SEM. Manuscript submitted for publication.
- Langenberg, B., & Mayer, A. (2020). semnova: Latent repeated measures ANOVA (Version 0.1-6). [Computer software]. <https://cran.r-project.org/package=semnova>
- Lu, Y., & Zhang, G. (2010). The equivalence between likelihood ratio test and f-test for testing variance component in a balanced one-way random effects model. *Journal of Statistical Computation and Simulation*, 80(4), 443–450. <https://doi.org/10.1080/00949650802695664>
- Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Lawrence Erlbaum Associates.
- Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, 59(1), 537–563. <https://doi.org/10.1146/annurev.psych.59.103006.093735>

- Mayer, A., Steyer, R., & Mueller, H. (2012). A general approach to defining latent growth components. *Structural Equation Modeling*, *19*(4), 513–533.
<https://doi.org/10.1080/10705511.2012.713242>
- McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 561–614). Plenum Press.
- McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *60*(1), 577–605.
<https://doi.org/10.1146/annurev.psych.60.110707.163612>
- McArdle, J. J., & Epstein, D. B. (1987). Latent growth curves within developmental structural equation models. *Child Development*, *58*(1), 110–133. <https://doi.org/10.2307/1130295>
- McArdle, J. J., & Hamagami, F. (2001). Latent difference score structural models for linear dynamic analyses with incomplete longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change*, (pp. 139–175). American Psychological Association.
<https://doi.org/10.1037/10409-005>
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. <https://doi.org/10.1007/BF02294825>
- Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*(2), 171–189. <https://doi.org/10.1111/j.2044-8317.1985.tb00832.x>
- Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.
- Newsom, J. T. (2002). A multilevel structural equation model for dyadic data. *Structural Equation Modeling*, *9*(3), 431–447. https://doi.org/10.1207/S15328007SEM0903_7
- Newsom, J. T. (2015). *Longitudinal structural equation modeling: A comprehensive introduction*. Routledge.
- Nimon, K. F. (2012). Statistical assumptions of substantive analyses across the general linear model: A mini-review. *Frontiers in Psychology*, *3*, Article 322. <https://doi.org/10.3389/fpsyg.2012.00322>
- Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*(3), 241–286.
<https://doi.org/10.1006/ceps.2000.1040>
- Perugini, M., Gallucci, M., & Costantini, G. (2018). A practical primer to power analysis for simple experimental designs. *International Review of Social Psychology*, *31*(1), Article 20.
<https://doi.org/10.5334/irsp.181>
- Pitts, S. C., West, S. G., & Tein, J. Y. (1996). Longitudinal measurement models in evaluation research: Examining stability and change. *Evaluation and Program Planning*, *19*(4), 333–350.
[https://doi.org/10.1016/S0149-7189\(96\)00027-4](https://doi.org/10.1016/S0149-7189(96)00027-4)
- R Core Team. (2021). *R: A language and environment for statistical computing*. [Computer software]. R Project for Statistical Computing.

- Raykov, T. (1999). Are simple change scores obsolete? An approach to studying correlates and predictors of change. *Applied Psychological Measurement, 23*(2), 120–126.
<https://doi.org/10.1177/01466219922031248>
- Raykov, T., & Widaman, K. F. (1995). Issues in applied structural equation modeling research. *Structural Equation Modeling: A Multidisciplinary Journal, 2*(4), 289–318.
<https://doi.org/10.1080/10705519509540017>
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–20. <https://doi.org/10.18637/jss.v048.i02>
- Rovine, M. J., & Molenaar, P. C. M. (1998). A nonstandard method for estimating a linear growth model in LISREL. *International Journal of Behavioral Development, 22*(3), 453–473.
<https://doi.org/10.1080/016502598384225>
- Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9*(2), 164–182.
<https://doi.org/10.1037/1082-989x.9.2.164>
- Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Lawrence Erlbaum Associates.
- Steyer, R., Eid, M., & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research, 2*(1), 21–33.
- UCLA Statistical Consulting Group. (2011). R library contrast coding systems for categorical variables.
<https://stats.idre.ucla.edu/r/library/r-library-contrast-coding-systems-forcategorical-variables/>
- Voelkle, M. C. (2007). Latent growth curve modeling as an integrative approach to the analysis of change. *Psychology Science, 49*(4), 375–414.
- Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives, 4*(1), 10–18. <https://doi.org/10.1111/j.1750-8606.2009.00110.x>

Appendices

Appendix A

R Code to Create an Orthogonal Contrast Matrix

```

> # create a data frame containing all experimental conditions
> idata <- expand.grid(A = c("A1", "A2"), B = c("B1", "B2", "B3"))
> idata
  A B
1 A1 B1
2 A2 B1
3 A1 B2
4 A2 B2
5 A1 B3
6 A2 B3
> # create a formula that describes the effects to be analyzed
> idesign <- ~A*B
> idesign
~A * B
> # create the B matrix (the inverse of the contrast matrix C)
> # using the model.matrix() function
> B_matrix <- model.matrix(
+   idesign,
+   idata,
+   contrasts.arg = list(A = "contr.poly", B = "contr.poly")
+ )
> B_matrix
      (Intercept)      A.L      B.L      B.Q      A.L:B.L      A.L:B.Q
1      1 -0.7071068 -7.071068e-01  0.4082483  5.000000e-01 -0.2886751
2      1  0.7071068 -7.071068e-01  0.4082483 -5.000000e-01  0.2886751
3      1 -0.7071068 -7.850462e-17 -0.8164966  5.551115e-17  0.5773503
4      1  0.7071068 -7.850462e-17 -0.8164966 -5.551115e-17 -0.5773503
5      1 -0.7071068  7.071068e-01  0.4082483 -5.000000e-01 -0.2886751
6      1  0.7071068  7.071068e-01  0.4082483  5.000000e-01  0.2886751
attr(,"assign")
[1] 0 1 2 2 3 3
attr(,"contrasts")
attr(,"contrasts")$A
[1] "contr.poly"
attr(,"contrasts")$B
[1] "contr.poly"

> # the C matrix is the inverse of the B matrix
> C_matrix <- solve(B_matrix)
> C_matrix
      1      2      3      4      5      6
(Intercept) 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667
A.L      -0.2357023 0.2357023 -0.2357023 0.2357023 -0.2357023 0.2357023
B.L      -0.3535534 -0.3535534 0.0000000 0.0000000 0.3535534 0.3535534
B.Q      0.2041241 0.2041241 -0.4082483 -0.4082483 0.2041241 0.2041241
A.L:B.L    0.5000000 -0.5000000 0.0000000 0.0000000 -0.5000000 0.5000000
A.L:B.Q   -0.2886751 0.2886751 0.5773503 -0.5773503 -0.2886751 0.2886751

```

Appendix B

R Code to Create an Orthonormal Contrast Matrix

```

> # the C matrix is an orthogonal contrast matrix
> C_matrix
      1      2      3      4      5      6
(Intercept) 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667
A.L          -0.2357023 0.2357023 -0.2357023 0.2357023 -0.2357023 0.2357023
B.L          -0.3535534 -0.3535534 0.0000000 0.0000000 0.3535534 0.3535534
B.Q           0.2041241 0.2041241 -0.4082483 -0.4082483 0.2041241 0.2041241
A.L:B.L       0.5000000 -0.5000000 0.0000000 0.0000000 -0.5000000 0.5000000
A.L:B.Q      -0.2886751 0.2886751 0.5773503 -0.5773503 -0.2886751 0.2886751
>
> # the C matrix is orthogonal if CC^T is a diagonal matrix
> round(C_matrix %*% t(C_matrix), 3)
      (Intercept)  A.L B.L B.Q A.L:B.L A.L:B.Q
(Intercept)      0.167 0.000 0.0 0.0      0      0
A.L              0.000 0.333 0.0 0.0      0      0
B.L              0.000 0.000 0.5 0.0      0      0
B.Q              0.000 0.000 0.0 0.5      0      0
A.L:B.L          0.000 0.000 0.0 0.0      1      0
A.L:B.Q          0.000 0.000 0.0 0.0      0      1
>
> # scale each row by the square root of its sum of squares
> C_matrix <- t(apply(C_matrix, 1, function(row) row / sqrt(sum(row^2))))
> C_matrix
      1      2      3      4      5      6
(Intercept) 0.4082483 0.4082483 0.4082483 0.4082483 0.4082483 0.4082483
A.L          -0.4082483 0.4082483 -0.4082483 0.4082483 -0.4082483 0.4082483
B.L          -0.5000000 -0.5000000 0.0000000 0.0000000 0.5000000 0.5000000
B.Q           0.2886751 0.2886751 -0.5773503 -0.5773503 0.2886751 0.2886751
A.L:B.L       0.5000000 -0.5000000 0.0000000 0.0000000 -0.5000000 0.5000000
A.L:B.Q      -0.2886751 0.2886751 0.5773503 -0.5773503 -0.2886751 0.2886751
>
> # the matrix is orthoNORMAL if (B^T)B is the identity matrix
> round(C_matrix %*% t(C_matrix), 3)
      (Intercept)  A.L B.L B.Q A.L:B.L A.L:B.Q
(Intercept)      1  0  0  0      0      0
A.L              0  1  0  0      0      0
B.L              0  0  1  0      0      0
B.Q              0  0  0  1      0      0
A.L:B.L          0  0  0  0      1      0
A.L:B.Q          0  0  0  0      0      1

```



Methodology is the official journal of the European Association of Methodology (EAM).



leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.