

How to Use Residual Dynamic Structural Equation Modeling to Study Individual Differences and Intraindividual Variability in Experimental Factorial Designs: A Tutorial

Benedikt Langenberg^a, Jonathan L. Helm^b, Connor J. McCabe^c, Thomas Günther^d, and Axel Mayer^e

^aMaastricht University, The Netherlands; ^bSan Diego State University, USA; ^cUniversity of Washington, USA; ^dRWTH Aachen University, Germany; ^eBielefeld University, Germany

ABSTRACT

This article demonstrates the application of residual dynamic structural equation modeling (RDSEM) for analyzing custom contrasts in experimental factorial designs. Previous applications of RDSEM have often focused on ecological momentary assessment and daily diary data. However, RDSEM was explicitly developed for intensive longitudinal data more generally, including settings with very short time intervals between observations. Beyond these types of studies, RDSEM is also well suited for analyzing data from laboratory studies such as eye-tracking or reaction time experiments. We compare three analytic approaches, namely analysis of variance, linear mixed models, and RDSEM, emphasizing the unique advantages of RDSEM. Although often applied to momentary assessment data, RDSEM proves highly effective for experimental analysis, offering the ability to integrate both time-varying and time-invariant covariates, model autoregressive effects, and capture interindividual differences in residual variances / intraindividual variability. These strengths arise from RDSEM's integration of time-series, multilevel, and latent variable modeling, all implemented through Bayesian estimation.

KEYWORDS

Residual dynamic structural equation modeling; experimental designs; interindividual differences; interindividual variability

Technological advances enable researchers to collect a vast amount of intensive data when conducting experiments with factorial designs in the social and behavioral sciences. These advances present new challenges to researchers when selecting appropriate methods for data analysis. Dynamic structural equation modeling (DSEM; Asparouhov et al., 2018) is a flexible tool that has been proposed in recent years for analyzing intensive longitudinal data. DSEM can be applied to longitudinal data with hours or days between assessments (e.g., ecological momentary assessment data, Smyth & Stone, 2003; daily diary studies, Bolger et al., 2003). DSEM is also well-suited for analyzing experiments with factorial designs that involve a time-series component as well (e.g., longitudinal data collected in a laboratory context, which may have seconds between consecutive assessments). DSEM introduces several advantages to the analysis of

experiments, namely, (1) a time-series component allowing for autoregressive effects, (2) multilevel modeling, (3) a latent variable framework, (4) computational feasibility through Bayesian estimation, (5) estimation of interindividual differences in residual variances, and (6) latent group-mean centering. This article aims to illustrate those advantages in the context of an experiment from reading research (although the advantages would generalize to any experimental context that collects intensive data). Intensive eye-tracking data will be used to examine the development of reading skills in children from grade one to four. DSEM will further be compared to results from analysis of variance (ANOVA) and linear mixed models (LMM) to demonstrate how DSEM improves upon these more traditional methodologies.

In order to compare DSEM with ANOVA and LMM, this article builds upon the recently published

latent repeated measures analysis of variance (L-RM-ANOVA) approach (Langenberg et al., 2024, 2022), which translates traditional ANOVA into the structural equation modeling (SEM) framework. When estimated using Bayesian methods, L-RM-ANOVA is equivalent to a Bayesian multivariate ANOVA. L-RM-ANOVA introduces many advantages of SEM to the analysis of data from experiments, including, but not limited to, a more comprehensive examination of interindividual differences, a convenient way of specifying custom contrasts, and the provision of a measurement model (i.e., latent variables). However, L-RM-ANOVA has been developed for single-level data only and does not allow for modeling dynamic time-series processes. These features are crucial when analyzing data that contain intensively repeated measurements over time. For instance, in eye-tracking experiments, the reading time for a sentence may be influenced by the difficulty level of a previous word. Furthermore, because observations are nested within participants, it is important to examine the extent to which main and interaction effects differ across individuals. This article integrates the advantages of L-RM-ANOVA into DSEM, thereby enabling researchers to explore interindividual differences in custom contrasts from factorial designs, as well as to examine intraindividual variability in dynamic¹ processes. Specifically, this article utilizes residual dynamic structural equation modeling (RDSEM, Asparouhov et al., 2018), a variant of DSEM that models autoregressive effects between residuals. Below we will refer to L-RM-ANOVA simply as ANOVA for the sake of brevity when discussing effect testing in single-level factorial designs. We will use the term *traditional* ANOVA when specifically referring to the original statistical method. It is important to note that, technically, ANOVA is not a model but a method for testing hypotheses that involve at least two parameters (i.e., numerator degrees of freedom of the *F*-test). This method is based on the general linear model, typically estimated using least squares.

This article is structured as follows: First, this article introduces the motivating example of an eye-tracking experiment including a 2×3 factorial within-subject design. Second, this article shows how to specify a set of custom contrasts which represent (i.e., test directly) hypotheses of substantive interest, while incorporating both fixed effects (i.e., a common, typically average, effect

across individuals) as well as random effects (i.e., interindividual differences that vary around the fixed effect) for the custom contrasts. Third, this article estimates and compares results from the ANOVA, a linear mixed model, and RDSEM. Although ANOVA operates on aggregated data and does not allow for time-varying covariates, the other models make use of the full dataset. The RDSEM will further include a set of time-invariant and time-varying covariates, an autoregressive effect, and interindividual differences in residual variances (which cannot be incorporated into ANOVA; time-varying covariates are common in LMMs but modeling error structures is admittedly complex and rarely done). Finally, the article draws general conclusions, provides further advice for applied researchers, and identifies future directions for research.

Motivating example

This section introduces the motivating example of this article and briefly formulates several example research questions that will be analyzed and answered in the remainder of this article. The example data stem from an eye-tracking experiment that investigated reading efficiency in $N = 267$ children in grades one, two, and four. The main research question is how children's reading efficiency improves over grades in elementary school. This research question will also guide the contrast coding and scaling, and the model specification in the following sections. The "Landolt" paradigm (e.g., Heim et al., 2018; Hillen et al., 2013) was used to measure reading efficiency. The Landolt paradigm involves reading two types of sentences: regular sentences and Landolt sentences, in which the characters of the words are replaced by circles. For Landolt sentences, some of the circles in the words are replaced by the character "C" and participants are asked to read through the words and discover the "C"s.² The idea behind this paradigm is that Landolt sentences involve only oculomotor skills, whereas regular sentences involve both oculomotor skills and linguistic processing (including semantic, phonetic and morphosyntactic processing). The difference in, for instance, reading times can be perceived as a measure of reading efficiency.

The two experimental within-subject factors were sentence type (regular sentence vs. Landolt sentence) and grade (grade one vs. two vs. four). In each grade, children read up to 41 sentences (some children read fewer sentences; median = 36, IQR = (36, 41)) with up to 10 words

¹Throughout this article we use the term "dynamic" following the terminology of Mplus, where it refers to autoregressive dependencies between successive observations. Other frameworks use the label differently, for example continuous-time models (e.g., Driver et al., 2017) where dynamics are defined through relations among derivatives of latent processes. We emphasize this distinction to avoid confusion for readers from different methodological traditions.

²Landolt sentences have the same structure and number of characters and words as regular sentences but only consist of circles and "C". Example: Regular sentence: The mother thinks Paul's big drum is loud
Landolt sentence: OOO OOCOOO OOOOOO OOOOO OOO OOOO OO OCOO.

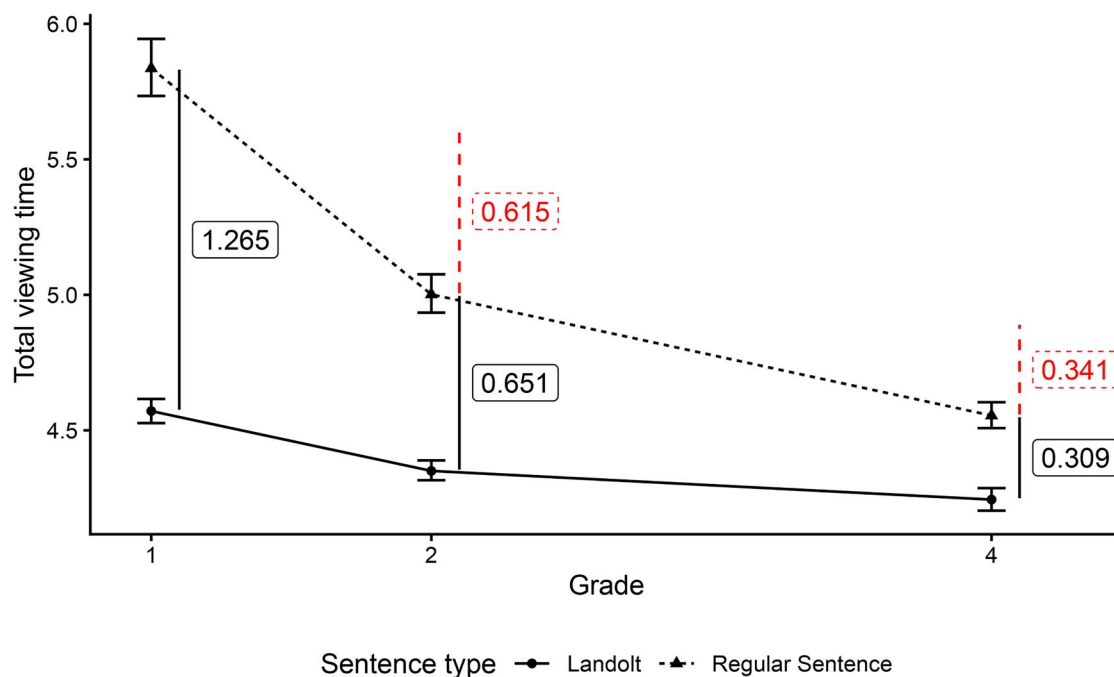


Figure 1. Model-implied fixed effects of the RDSEM (see Section “Model 3: RDSEM”) for the six experimental conditions. Error bars indicate credibility intervals. The solid vertical bars indicate the difference between Landolt and regular sentences (i.e., reading efficiency) for each of the grades. The dashed vertical bars indicate the change in reading efficiency from grade one to two and two to four, respectively.

per sentence (sentences differed in length; median 7, IQR = (5, 9)), resulting in up to 1,575 measurements per child (median = 1,341, IQR = (1,239, 1,424)). Children in the sample were recruited from different schools, so additional clustering effects at the school level may be present. Because school membership was not the focus of the present analysis and the number of children per school was limited, these potential dependencies could not be modeled explicitly. This limitation should be kept in mind when interpreting the results. Several measures were collected such as total viewing time³ (TVT). Figure 1 shows the average viewing time per word across the $2 \times 3 = 6$ experimental conditions. The viewing time was divided by the number of characters of the word and then log-transformed as the data was substantially skewed.

Development of reading efficiency

In this article, we will investigate several example research questions. We will first examine the development of reading efficiency over grades. In Figure 1, the black bars indicate the average difference in viewing time between regular and Landolt sentences (i.e., reading efficiency) for each of the three grades. The vertical dashed red bars indicate the average change in reading

efficiency (i.e., difference of differences) from grade one to two and from two to four. We will use the identical contrasts in the model specification section as they match the substantive research question. The larger the change, the more efficient readers became from one grade to the next. We will examine both average changes across grades (i.e., fixed effects) and the variability of these changes across participants (i.e., random effects). The data further includes several time-invariant covariates: (1) children’s IQ, (2) attention problems *via* the child behavior checklist, (CBCL; Achenbach & Rescorla, 2001), (3) reading difficulty (RD), the average percentile below the 25th percentile for reading speed (*Salzburger Lese- und Rechtschreibtests* (Salzburg Reading and Spelling Test), SLRT-II; Moll & Landerl, 2010), and (4) reading comprehension (*Ein Lese-Verständnistest für Erst- bis Sechstklässler* (A Reading Comprehension Test for First to Sixth Graders), ELFE 1–6; Lenhard & Schneider, 2006). All time-invariant covariates were measured before or at the beginning of the first grade. We will use these covariates in this article to explain and predict the change in reading efficiency.

Time-varying covariates

The second research question addresses the use of time-varying covariates. Specifically, we aim to examine how the initial landing position (LP) within a word influences the viewing time. Prior research has established that

³Total viewing time is the sum of all gaze times on a word while reading a sentence, including looking back on a word. Total viewing time will be abbreviated “viewing time” in the remainder of this article.

experienced readers tend to land near the center of a word during reading (Radach & McConkie, 1998), and landing closer to a word's center is expected to result in shorter viewing times. To quantify this, the landing position was measured as the absolute distance (in characters) between the first fixation and the center of the word. This distance was then standardized by dividing it by the total word length, yielding a variable that ranges from 0 (indicating a landing exactly in the center of the word) to 1 (indicating a landing on either the first or last character).

Heterogeneity of residual variances

The third question concerns interindividual differences in intraindividual variability. Individual residual variance is a measure of intraindividual variability: it describes how an individual's viewing time fluctuates around their mean. This variability may, however, be affected by different person characteristics (i.e., time-invariant covariates such as IQ, CBCL, RD). That is, participants with reading difficulties may show higher intraindividual variability. We will examine how IQ, CBCL, and RD affect this variability.

Autoregressive effects

Lastly, we will investigate the degree to which viewing time of the current word is affected by the viewing time of the previous word. For example, viewing time may be longer when the previous word is more difficult (e.g., due to cognitive fatigue). Furthermore, these autoregressive effects may vary across individuals, and may be predicted by time-invariant variables (e.g., children with higher IQ may have a weaker autoregressive effect if they are able to process more difficult words without cognitive fatigue).

Prior models for intensive longitudinal data in experimental contexts

Traditional ANOVA

Traditional ANOVA is capable of estimating average changes in reading efficiency across grades (i.e., between-person effects in the present design; see Section "Development of Reading Efficiency"), and predicting these changes through time-invariant variables (Langenberg et al., 2022). However, traditional ANOVA cannot model the within-person process (e.g., the autoregressive effect), thereby precluding examination of between-person differences in the within-person process.

Linear mixed models

LMM (for introductions to linear mixed models, see, e.g., Fitzmaurice et al., 2011; Laird & Ware, 1982) is an

extension of ANOVA that allows estimation of within-person effects, as it distinguishes between within- and between-subject variability. LMMs are able to additionally incorporate time-varying covariates at the within level, such as the landing position or autoregression (see Sections below "Time-Varying Covariates" and "Autoregressive Effects"). However, including time-varying covariates and autoregressive effects in multilevel models requires careful consideration of different centering methods (e.g., grand-mean centering or cluster-mean centering; Hamaker & Grasman, 2014). The choice of centering can substantially change the interpretation of the model parameters. This is especially true for factorial designs, where the interpretation of the intercept (and thus the autoregressive effect) further depends on the choice of factor coding. Furthermore, research has shown that, in some cases, autoregressive effects should be modeled between residuals; otherwise, biased variance estimates may result ((e.g., Sivo et al., 2005; Kwok et al., 2007). This article will demonstrate how to estimate and interpret autoregressive effects between residuals in experiments with custom contrasts using latent cluster-mean centering.

There is also literature on interindividual differences in intraindividual variability (e.g., Hoffman, 2007). A variant of the LMM known as the location scale model (LCM) is an example of this, allowing for the modeling of within-subject variability as a function of time or other cluster-level covariates (e.g., Hedeker et al., 2008; Williams et al., 2019, 2020). This type of model is typically applied to longitudinal data, such as momentary assessment data (e.g., Hedeker et al., 2008, 2012; for an implementation, see Hedeker, 2013). LCMs can also include autoregressive effects. However, to our knowledge existing literature does not include LCMs with modeled autoregressive effects among residuals; instead, it has used the previous measurement to predict the current measurement or the residual variance of the current measurement.⁴

⁴A closely related recent development is the mixed-effects location scale (MELS) framework (Nestler, 2022; Nestler & Blozis, 2024), which extends location-scale modeling by jointly capturing individual differences in mean levels (location), intraindividual variability (scale), and serial dependence parameters (e.g., autocorrelation) within a unified mixed-effects specification. MELS separates within- and between-person components, allows time-varying and time-constant predictors, and can model interindividual differences in intraindividual variability as well as autocorrelation parameters. Notably, MELS has also been extended to a latent-variable formulation (Nestler & Blozis, 2024), which allows outcomes to be represented at the measurement level. In this framework, associations among person-specific effects (e.g., between a mean slope and a scale component) are typically expressed through correlations among random effects, rather than through structural regressions among slopes. Recent extensions of the MELS framework further incorporate autoregressive residuals (Nestler, 2022; Nestler & Blozis, 2024).

DSEM and RSEM models

In DSEM models, autoregressive relations are specified for the outcome itself (within level), whereas in RDSEM models the autoregressive relations are specified for the residuals after accounting for the structural part of the model (Asparouhov et al., 2018; Asparouhov & Muthén, 2020a, 2020b). This residual-based specification in RDSEM separates the structural relations from the time-series component and keeps the focus on contemporaneous relations among variables, whereas DSEM models structural relations that extend across time points (Asparouhov & Muthén, 2020). In terms of application, the choice should follow the substantive question: when interest lies in dynamics of the outcome across time, DSEM is appropriate; when interest lies in residual dynamics after conditioning on predictors, RDSEM is appropriate (Asparouhov & Muthén, 2020). In our experimental setting, the focus is on contemporaneous effects between reading-related variables, making RDSEM a natural choice. One could also consider including a time trend across sentences (e.g., children becoming faster as the experiment progresses), although the presence of such a trend alone should not determine model selection. If both specifications are plausible, model adequacy of the DSEM and RSEM models can be examined using Bayesian information criteria such as the deviance information criterion (DIC) under the same time-scale settings, and posterior predictive checks. Simulation studies show that DIC can identify the correct modeling framework in comparative settings (Asparouhov & Muthén, 2020). For our motivating example, we will therefore also briefly compare the DIC of DSEM and RDSEM. RDSEM has further been shown to be more robust to certain misspecifications of unevenly spaced observations. This robustness alone should not drive model choice; instead, times should be modeled as accurately as possible and model selection should combine substantive arguments with comparative fit (Asparouhov & Muthén, 2020a).

RDSEM also offers additional advantages over existing approaches. RDSEM models slopes as latent variables that have means and variances (i.e., fixed and random effects), allowing for the regression of slopes on one another, unlike LMM, which does not support a structural model among slopes. Modeling random effects as latent variables further allows for the estimation of the proportion of variance explained by covariates. Although this can also be achieved in LMM through more complex parameterization, RDSEM naturally provides this feature. Lastly, RDSEM enables the estimation of correlations between random effects, such as an autoregressive

effect (capturing relationships between residuals) and the effect of an experimental factor. In this tutorial paper, we focus on RDSEM because of its clear illustration of the modeling of experimental designs with time-varying covariates and autoregressive processes.

In the remainder of this article, the three proposed models will be estimated to examine the development and change of reading efficiency. The models build upon and extend each other step by step, allowing for the research questions from the previous section to be addressed.

Specifying custom contrasts

This section provides a brief introduction on how to specify linear hypotheses in SEM. This approach formulates hypotheses as contrasts (or differences) between experimental conditions of within-subject factors and is derived from the latent growth components approach (Mayer et al., 2012, 2013). The idea can be perceived as a generalization of latent growth curves, since growth curves are, in fact, also differences between variables. The dependent variable under the experimental conditions is denoted by y and contrasts between those are labeled π . In our motivating example, the experimental conditions are the six combinations of sentence type and grade of the variable viewing time. Hypotheses are specified by expressing the contrasts as a function of the dependent variables using a contrast matrix C :

$$\pi = Cy. \quad (1)$$

The contrast matrix can contain any set of contrasts, as long as it is invertible. In practice, this means that the matrix must include the same number of contrasts as there are experimental conditions, and that the contrasts must be linearly independent (i.e., none of them can be expressed as a combination of the others). Researchers often prefer an orthogonal contrast matrix because orthogonal contrasts are uncorrelated with one another, which makes their interpretation more straightforward. Orthogonality simply means that the rows of the contrast matrix are independent in a geometric sense and do not overlap in the information they represent

For our example, we use a set of specific contrasts. Recall, we are interested in reading efficiency and change of efficiency over time. For that reason, we create contrasts for (1) an intercept (i.e., grand mean), (2) average reading efficiency across all grades; (3) change in average viewing time from grade one to two, (4) change in average viewing time from grade two to four, (5) change in reading efficiency from grade one to two, (6) change in reading efficiency from grade two to four. The latter two contrasts represent the main research

focus. Some literature (e.g., Rosenthal & Rosnow, 1985; Abelson & Prentice, 1997) recommends to scale the contrasts (i.e., normalize so the sum of squared coefficients equals one), as this may facilitate comparisons between contrasts. However, we decided not to scale coefficients because of their substantial interpretation.

The contrast matrix for these contrasts is given by:

$$\mathbf{C} = \begin{array}{cc} \text{Factor} & \text{Sentence type} \\ & \text{Grade} \end{array} \begin{array}{cc} \text{S} & \text{L} \\ \text{G1} & \text{G1} \end{array} \begin{array}{cc} \text{S} & \text{L} \\ \text{G2} & \text{G2} \end{array} \begin{array}{cc} \text{S} & \text{L} \\ \text{G4} & \text{G4} \end{array} \begin{array}{c} \text{Intercept} \\ \text{D} \\ \text{T}_1 \\ \text{T}_2 \\ \text{D} : \text{T}_1 \\ \text{D} : \text{T}_2 \end{array} \begin{pmatrix} 0.167 & 0.167 & 0.167 & 0.167 & 0.167 & 0.167 \\ -0.333 & 0.333 & -0.333 & 0.333 & -0.333 & 0.333 \\ 0.5 & 0.5 & -0.5 & -0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 & -0.5 & -0.5 \\ -1 & 1 & 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 1 & -1 \end{pmatrix}. \quad (2)$$

We can see that the first row represents an intercept which is the grand mean across the six conditions. Choosing the grand mean as the first contrast simplifies interpretation of model parameters later in this manuscript (see Section “Model 3: RDSEM”). The second row represents the difference between the

Hypotheses usually make statements about means and variances of the contrasts π . To estimate those means and variances, we need to express the dependent variables y as a function of the contrasts π . This can be achieved by solving the equation for y and inverting the contrast matrix $\mathbf{C}$:

$$\mathbf{B}^* = \mathbf{C}^{-1} = \begin{array}{cc} \text{Factor} & \text{Sentence type} \\ & \text{Grade} \end{array} \begin{array}{cc} \text{Intercept} & \text{D} \\ \text{T}_1 & \text{T}_2 \\ \text{D} : \text{T}_1 & \text{D} : \text{T}_2 \end{array} \begin{pmatrix} 1 & -0.5 & 0.667 & 0.333 & -0.333 & -0.167 \\ 1 & 0.5 & 0.667 & 0.333 & 0.333 & 0.167 \\ 1 & -0.5 & -0.333 & 0.333 & 0.167 & -0.167 \\ 1 & 0.5 & -0.333 & 0.333 & -0.167 & 0.167 \\ 1 & -0.5 & -0.333 & -0.667 & 0.167 & 0.333 \\ 1 & 0.5 & -0.333 & -0.667 & -0.167 & -0.333 \end{pmatrix}. \quad (3)$$

mean of all Landolt conditions and the mean of all regular sentence conditions. The third and fourth row represent the difference of the mean of Landolt and regular sentences between grade one and two, and

The inverted contrast matrix is often referred to as $\mathbf{B}^*$ (Langenberg et al., 2022) and the resulting equation is given by:

$$\mathbf{y}_i = \begin{pmatrix} \text{TVT}_{\text{L,G1}} \\ \text{TVT}_{\text{S,G1}} \\ \text{TVT}_{\text{L,G2}} \\ \text{TVT}_{\text{S,G2}} \\ \text{TVT}_{\text{L,G4}} \\ \text{TVT}_{\text{S,G4}} \end{pmatrix} = \begin{pmatrix} 1 & -0.5 & 0.667 \\ 1 & 0.5 & 0.667 \\ 1 & -0.5 & -0.333 \\ 1 & 0.5 & -0.333 \\ 1 & -0.5 & -0.333 \\ 1 & 0.5 & -0.333 \end{pmatrix} \mathbf{B}^* \begin{pmatrix} \pi_i \\ \beta_{0,i} \\ \beta_{\text{D},i} \\ \beta_{\text{T}_1,i} \\ \beta_{\text{T}_2,i} \\ \beta_{\text{D:T}_1,i} \\ \beta_{\text{D:T}_2,i} \end{pmatrix} \quad (4)$$

We can see that this equation looks very similar to regular regression with categorical predictors. The power, however, of this approach is the very flexible and convenient way of specifying contrasts as differences between experimental conditions of the dependent variables. After we derived the contrasts, we now need save them in the data file. The resulting long-format file for child 2 looks like this:

id	trial	word	grade	sentence	tvw	iq	cbcl	rd	lp	c_d	c_t1	c_t2	c_dt1	c_dt2
<int>	<dbl>	<dbl>	<fct>	<fct>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2	1	1	1	sentence	9.58	87.9	51	-0.5	0.147	0.5	0.667	0.333	0.333	0.167
2	1	1	1	landolt	7.54	87.9	51	-0.5	0.767	-0.5	0.667	0.333	-0.333	-0.167
2	1	1	2	sentence	8.49	87.9	51	-0.5	0.473	0.5	-0.333	0.333	-0.167	0.167
2	1	1	2	landolt	7.32	87.9	51	-0.5	0.487	-0.5	-0.333	0.333	0.167	-0.167
2	1	1	4	sentence	7.43	87.9	51	-0.5	0.66	0.5	-0.333	-0.667	-0.167	-0.333
2	1	1	4	landolt	6.69	87.9	51	-0.5	0.107	-0.5	-0.333	-0.667	0.167	0.333

We see the six rows that stem from the same subject (“id” equals 2), representing the first word (“word”; word index) of the first sentence (i.e., sentence index; “trial”) in the six experimental conditions. In each of the three years, children had to read 41 Landolt and regular sentences. “sentence” (sentence type: regular vs. Landolt) and “grade” (school grade: 1, 2, or 4) represent the six experimental conditions, which are coded through the six variables beginning with “c_”. These contrast variables were derived from the above contrast matrix. The variable “tvw” (log total viewing time per character) represents the dependent variable. “iq” (intelligence quotient), “cbcl” (child behavior checklist, attention problems score), and “rd” (reading difficulty) are the aforementioned time-invariant covariates, and “lp” (landing position within a word) is the time-varying covariate. The resulting dataset is now ready for the proposed modeling in the following sections.

Inverting the contrast matrix, creating and saving the contrast variables can be cumbersome. We provide a convenience R (R Core Team, 2021) function in the [online supplemental material](#) “8 Function for Creating Custom Contrasts” that automates the creation of the contrasts. It takes a data frame in R and a contrast matrix as inputs, and will augment the data set by the inverse of the contrast matrix.

Model 1: ANOVA

The following sections contain the three proposed models that are able to answer the questions raised in the introduction. The models are each estimated in Mplus (Muthén & Muthén, 1918-2017). We used 2,000 iterations (i.e., Bayesian updates) for each of the models. The results of fixed effects, that is intercepts and regression coefficients, can be found in [Table 1](#).

The results of variances, that is variability of random effects, can be found in [Table 2](#). All time-invariant covariates are grand mean centered (i.e., each covariate is centered around the sample mean) to preserve the interpretation of the intercept. We will use Bayesian estimation and Mplus (Muthén & Muthén, 1918-2017) for the model to ensure comparability with each of the other models presented in this manuscript. The software code for all models is publicly available at <https://osf.io/7wdfv/and> in the [online supplemental materials](#) (materials “Mplus Code for Model 1: ANOVA”, “Mplus Code for Model 2: Linear Mixed Model”, and “Mplus Code for Model 3: RDSEM”). Bayesian credible intervals are reported at the 95% level, reflecting the central range of the posterior distribution.

The first model that we consider is a “simple” ANOVA model based on L-RM-ANOVA. The motivating example constitutes a hierarchical structure, with observations (i.e., words) nested in participants. This structure will be ignored by averaging viewing time within each of the experimental conditions. This will yield six observations of viewing time for each of the $N = 267$ participants. The following equation expresses the dependent variable viewing time as a function of the six contrast variables that we created in the previous section and the time-invariant covariates:

$$\begin{array}{c} \mathbf{y}_i \\ \left(\begin{array}{c} TVT_{L,G1} \\ TVT_{S,G1} \\ TVT_{L,G2} \\ TVT_{S,G2} \\ TVT_{L,G4} \\ TVT_{S,G4} \end{array} \right) \end{array} = \begin{array}{c} \mathbf{v} \\ \left(\begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} \right) \end{array} + \begin{array}{c} \mathbf{\Lambda} \\ \left(\begin{array}{ccc|ccc} 1 & -0.5 & 0.667 & 0.333 & -0.333 & -0.167 \\ 1 & 0.5 & 0.667 & 0.333 & 0.333 & 0.167 \\ 1 & -0.5 & -0.333 & 0.333 & 0.167 & -0.167 \\ 1 & 0.5 & -0.333 & 0.333 & -0.167 & 0.167 \\ 1 & -0.5 & -0.333 & -0.667 & 0.167 & 0.333 \\ 1 & 0.5 & -0.333 & -0.667 & -0.167 & -0.333 \end{array} \right) \end{array} \begin{array}{c} \boldsymbol{\pi}_i \\ \left(\begin{array}{c} \beta_{0,i} \\ \beta_{D,i} \\ \beta_{T_1,i} \\ \beta_{T_2,i} \\ \beta_{D:T_1,i} \\ \beta_{D:T_2,i} \end{array} \right) \end{array} + \begin{array}{c} \boldsymbol{\varepsilon}_i \\ \left(\begin{array}{c} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} \right) \end{array} \quad (5)$$

$$\begin{array}{c} \boldsymbol{\pi}_i \\ \left(\begin{array}{c} \beta_{0,i} \\ \beta_{D,i} \\ \beta_{T_1,i} \\ \beta_{T_2,i} \\ \beta_{D:T_1,i} \\ \beta_{D:T_2,i} \end{array} \right) \end{array} = \begin{array}{c} \boldsymbol{\alpha} \\ \left(\begin{array}{c} \gamma_{00} \\ \gamma_{10} \\ \gamma_{20} \\ \gamma_{30} \\ \gamma_{40} \\ \gamma_{50} \end{array} \right) \end{array} + \begin{array}{c} \boldsymbol{\Gamma} \\ \left(\begin{array}{ccc} \gamma_{01} & \gamma_{02} & \gamma_{03} \\ \gamma_{11} & \gamma_{12} & \gamma_{13} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} \\ \gamma_{31} & \gamma_{32} & \gamma_{33} \\ \gamma_{41} & \gamma_{42} & \gamma_{43} \\ \gamma_{51} & \gamma_{52} & \gamma_{53} \end{array} \right) \end{array} \begin{array}{c} \mathbf{x}_i \\ \left(\begin{array}{c} IQ_i^c \\ CBCL_i^c \\ RD_i^c \end{array} \right) \end{array} + \begin{array}{c} \boldsymbol{\zeta}_i \\ \left(\begin{array}{c} \zeta_{0,i} \\ \zeta_{1,i} \\ \zeta_{2,i} \\ \zeta_{3,i} \\ \zeta_{4,i} \\ \zeta_{5,i} \end{array} \right) \end{array} \quad (6)$$

The β 's in the model represent the contrast variables as defined in the previous section. The γ 's are the means of the contrast variables (i.e., the fixed effects). Although this is not a multilevel approach, we still model random effects (i.e., interindividual differences in contrasts / effects) through $\boldsymbol{\zeta}$ (Rovine & Molenaar, 1998, 2000, 2001; Bauer, 2003), where $\text{Var}(\boldsymbol{\zeta}) = \boldsymbol{\Psi}$, a variance-covariance matrix with diagonal elements τ_{jj} . In this formulation, residual variances at the within level are not estimated because, after aggregating to condition means, each person contributes only a single observation per condition. Thus, the within-person residual variance is absorbed into the variance of the contrasts. As a consequence, the residuals are not separately identifiable in this specification.

We now examine selected contrasts in more detail. $\beta_{0,i}$ is the intercept, that is the mean of viewing time across all six conditions. The average of this mean across participants (i.e., the fixed effect) is $\hat{\gamma}_{00} = 4.69$ [4.65, 4.72], where the values in brackets represent the 95% credibility interval of the posterior distribution. This means that participants needed a viewing time of 4.69 units per character on the log-transformed milliseconds scale to read a word, or 109 ms on the original scale. In the remainder of this article, we will only discuss findings on the log-transformed scale for the sake of simplicity. The variance of the deviations from the average across participants (i.e., the variability of the random effect) is $\hat{\tau}_{00} = 0.06$ [0.05, 0.08]. From those values, we can illustrate how viewing time differs across children by taking the average effect and adding and subtracting one standard deviation of the random effect.

This yields the expected viewing time for participants who are one standard deviation below or above average: $\hat{\gamma}_{00} \pm \sqrt{\hat{\tau}_{00}} = 4.69 \pm 0.25 = [4.44, 4.94]$.⁵

The two contrasts of primary interest are $\beta_{D:T_1,i}$ and $\beta_{D:T_2,i}$. The mean change in reading efficiency from grade one to two is $\hat{\gamma}_{40} = 0.56$ [0.49, 0.63] and from grade two to four is $\hat{\gamma}_{50} = 0.30$ [0.25, 0.35]. In ANOVA, omnibus hypothesis tests of main effects of a factor are often of interest, with the null hypothesis $H_0 : \gamma_{40} = \gamma_{50} = 0$, meaning there is no change in reading efficiency from grade one to two, or from grade two to four. In Bayesian models, we cannot use an F -test to compare two (or more) model parameters against zero, nor can we use a Wald or Likelihood Ratio Test. However, a test can be constructed based on the draws from the posterior, which is asymptotically equivalent to a Wald test (Asparouhov & Muthén, 2020). Mplus calculates the median of the draws for each parameter and the variance-covariance matrix of the draws. A Wald test based on this information yields $\chi^2(2) = 453.24$, $p < .001$, providing strong evidence for a change in reading efficiency from grade one to two or from grade two to four, or both.

⁵This calculation illustrates variability across children rather than uncertainty of the estimate. It shows the expected viewing times for two hypothetical participants who lie one standard deviation below and above the sample average. It is therefore not a credibility interval. If desired, credibility intervals for these derived values could be obtained by repeating the calculation for each posterior draw and summarizing the resulting distributions (e.g., using the 2.5% and 97.5% percentiles).

Table 1. Parameter estimates of the fixed effects for the three models ANOVA, linear mixed model, and RDSEM.

		ANOVA		Mixed Model		RDSEM	
		Est.	95% CI	PDO	Est.	95% CI	PDO
Intercepts							
TVT_{wi}	γ_{00}	4.687* (4.649, 4.724)		83%	4.696* (4.658, 4.734)		12%
$\beta_{D,i}$	γ_{10}	0.627* (0.570, 0.689)		14%	0.544* (0.493, 0.598)		0%
$\beta_{T_1,i}$	γ_{20}	0.507* (0.467, 0.548)		17%	0.454* (0.419, 0.489)		7%
$\beta_{T_2,i}$	γ_{30}	0.254* (0.221, 0.289)		42%	0.230* (0.201, 0.258)		15%
$\beta_{D:T_1,i}$	γ_{40}	0.563* (0.494, 0.631)		24%	0.485* (0.427, 0.541)		5%
$\beta_{D:T_2,i}$	γ_{50}	0.302* (0.249, 0.352)		21%	0.240* (0.195, 0.285)		5%
$\beta_{LP,i}$	γ_{60}				0.160* (0.149, 0.171)		76%
ε_{wi}	$\exp(\omega_{70})$						
φ_i	γ_{80}				0.154* (0.144, 0.164)		93%
Regressions							
TVT_{wi}							
IQ_i^c	γ_{01}	−0.002 (−0.005, 0.001)		86%	−0.002 (−0.005, 0.001)		92%
$CBCL_i^c$	γ_{02}	−0.004 (−0.011, 0.003)		91%	−0.003 (−0.010, 0.004)		93%
RD_i^c	γ_{03}	0.347* (0.263, 0.431)		86%	0.362* (0.274, 0.447)		85%
$\beta_{D,i}$							
IQ_i^c	γ_{11}	−0.002 (−0.007, 0.002)		92%	−0.002 (−0.006, 0.002)		87%
$CBCL_i^c$	γ_{12}	−0.003 (−0.014, 0.008)		94%	−0.003 (−0.012, 0.007)		93%
RD_i^c	γ_{13}	0.501* (0.375, 0.622)		54%	0.424* (0.305, 0.537)		53%
$\beta_{T_1,i}$							
IQ_i^c	γ_{21}	−0.001 (−0.004, 0.003)		70%	−0.002 (−0.004, 0.001)		77%
$CBCL_i^c$	γ_{22}	−0.002 (−0.010, 0.006)		80%	−0.004 (−0.010, 0.002)		82%
RD_i^c	γ_{23}	0.076 (−0.007, 0.167)		88%	0.086* (0.008, 0.160)		91%
$\beta_{T_2,i}$							
IQ_i^c	γ_{31}	−0.000 (−0.003, 0.002)		95%	−0.000 (−0.003, 0.002)		92%
$CBCL_i^c$	γ_{32}	0.004 (−0.002, 0.010)		89%	0.005 (−0.001, 0.010)		93%
RD_i^c	γ_{33}	0.089* (0.008, 0.165)		64%	0.056 (−0.009, 0.123)		69%
$\beta_{D:T_1,i}$							
IQ_i^c	γ_{41}	−0.004 (−0.009, 0.001)		90%	−0.004 (−0.009, 0.000)		93%
$CBCL_i^c$	γ_{42}	0.010 (−0.003, 0.023)		90%	0.009 (−0.001, 0.020)		95%
RD_i^c	γ_{43}	0.208* (0.065, 0.364)		82%	0.237* (0.115, 0.365)		85%
$\beta_{D:T_2,i}$							
IQ_i^c	γ_{51}	−0.002 (−0.006, 0.002)		80%	−0.003 (−0.006, 0.000)		90%
$CBCL_i^c$	γ_{52}	−0.006 (−0.015, 0.003)		82%	−0.008 (−0.016, 0.000)		72%
RD_i^c	γ_{53}	0.177* (0.060, 0.291)		77%	0.146* (0.048, 0.244)		85%
$\beta_{LP,i}$							
IQ_i^c							
$CBCL_i^c$							
RD_i^c							
ε_{wi}							
IQ_i^c	$\exp(\omega_{71})$						
$CBCL_i^c$	$\exp(\omega_{72})$						
RD_i^c	$\exp(\omega_{73})$						
φ_i							
IQ_i^c	γ_{81}						
$CBCL_i^c$	γ_{82}						
RD_i^c	γ_{83}						

Note. Bold red font indicates changes of more than 10% in the median of the posterior distribution between neighboring models. “PDO” (posterior distribution overlap) indicates the overlap of the posterior distribution of neighboring models in percentage. The asterisk indicates a significant result using a 95% credibility interval.

Our focus is on interindividual differences rather than hypothesis tests for main and interaction effects. Furthermore, the previous χ^2 test is significant at $p < .001$ and will most likely not differ across ANOVA, LMM, and RDSEM. For these reasons, we will focus on the variances of the contrasts. The variance of this change from grade one to two is $\hat{\tau}_{44} = 0.18$ [0.14, 0.24] and $\hat{\tau}_{55} = 0.11$ [0.09, 0.14] from grade two to four, respectively. Following the same logic as above, one standard deviation below and above the average change is obtained as $\hat{\gamma}_{40} \pm \sqrt{\tau_{44}} = [0.14, 0.99]$ for the change from grade one to two and $\hat{\gamma}_{50} \pm \sqrt{\tau_{55}} = [-0.03, 0.63]$ for

the change from grade two to four, illustrating the range of developmental changes that can be expected for children who lie one standard deviation below or above the average. We can see that the change from grade two to four is smaller. This is not surprising based on prior theory and research since we expect viewing time in regular sentences to converge against viewing time in Landolt sentences. We can also estimate the effect of the time-invariant covariates on the change of reading efficiency. Although IQ and CBCL did not have a significant effect, RD does have an effect. For children that suffer from RD, the change in reading efficiency will be larger with

Table 2. Parameter estimates of the random effects for the three models ANOVA, linear mixed model, and RDSEM.

		ANOVA			Mixed Model			RDSEM	
		Est.	95% CI	PDO	Est.	95% CI	PDO	Est.	95% CI
Residual Variances (within)									
TVT_{wi}	τ				0.292*	(0.290, 0.294)			
Residual Variances									
TVT_{wi}	τ_{00}	0.062*	(0.050, 0.079)	68%	0.068*	(0.055, 0.086)	64%	0.062*	(0.050, 0.078)
$\beta_{D,i}$	τ_{11}	0.152*	(0.122, 0.190)	23%	0.114*	(0.091, 0.144)	22%	0.152*	(0.123, 0.192)
$\beta_{T_1,i}$	τ_{22}	0.064*	(0.052, 0.082)	19%	0.047*	(0.037, 0.059)	22%	0.063*	(0.050, 0.078)
$\beta_{T_2,i}$	τ_{33}	0.047*	(0.039, 0.060)	30%	0.037*	(0.030, 0.047)	34%	0.046*	(0.037, 0.060)
$\beta_{D:T_1,i}$	τ_{44}	0.182*	(0.143, 0.237)	14%	0.126*	(0.098, 0.160)	27%	0.166*	(0.132, 0.211)
$\beta_{D:T_2,i}$	τ_{55}	0.107*	(0.087, 0.135)	17%	0.077*	(0.060, 0.098)	22%	0.105*	(0.082, 0.135)
$\beta_{LP,i}$	τ_{66}				0.003*	(0.002, 0.004)	23%	0.002*	(0.001, 0.003)
ε_{wi}	τ_{77}							0.025*	(0.021, 0.032)
φ_i	τ_{88}				0.003*	(0.003, 0.005)	76%	0.004*	(0.003, 0.005)
Rsquared (within)									
TVT_{wi}					0.400	(0.397, 0.403)		0.425	(0.420, 0.429)
Rsquared									
TVT_{wi}		0.310	(0.198, 0.412)		0.177	(0.106, 0.253)		0.176	(0.106, 0.249)
$\beta_{D,i}$		0.279	(0.175, 0.376)		0.149	(0.080, 0.223)		0.150	(0.088, 0.225)
$\beta_{T_1,i}$		0.037	(0.005, 0.113)		0.031	(0.006, 0.078)		0.019	(0.003, 0.057)
$\beta_{T_2,i}$		0.069	(0.015, 0.163)		0.024	(0.004, 0.062)		0.028	(0.006, 0.065)
$\beta_{D:T_1,i}$		0.123	(0.037, 0.238)		0.071	(0.028, 0.129)		0.050	(0.015, 0.099)
$\beta_{D:T_2,i}$		0.082	(0.019, 0.174)		0.053	(0.015, 0.116)		0.040	(0.009, 0.093)
$\beta_{LP,i}$								0.091	(0.030, 0.189)
ε_{wi}								0.047	(0.015, 0.095)
φ_i								0.020	(0.003, 0.058)
Residual Correlations									
TVT_{wi}									
$\beta_{D,i}$	ρ_{01}	0.616*	(0.509, 0.707)	75%	0.582*	(0.464, 0.679)	73%	0.621*	(0.510, 0.707)
$\beta_{T_1,i}$	ρ_{02}	0.371*	(0.227, 0.504)	96%	0.364*	(0.206, 0.504)	89%	0.387*	(0.240, 0.515)
$\beta_{T_2,i}$	ρ_{03}	0.425*	(0.281, 0.548)	86%	0.401*	(0.239, 0.532)	90%	0.420*	(0.281, 0.548)
$\beta_{D:T_1,i}$	ρ_{04}	0.296*	(0.144, 0.434)	53%	0.393*	(0.238, 0.527)	58%	0.308*	(0.152, 0.443)
$\beta_{D:T_2,i}$	ρ_{05}	0.462*	(0.329, 0.584)	76%	0.419*	(0.275, 0.557)	72%	0.470*	(0.331, 0.585)
$\beta_{D,i}$									
$\beta_{T_1,i}$	ρ_{12}	0.261*	(0.110, 0.399)	81%	0.302*	(0.145, 0.441)	82%	0.263*	(0.103, 0.409)
$\beta_{T_2,i}$	ρ_{13}	0.443*	(0.301, 0.560)	97%	0.444*	(0.298, 0.574)	96%	0.435*	(0.290, 0.560)
$\beta_{D:T_1,i}$	ρ_{14}	0.337*	(0.188, 0.475)	61%	0.414*	(0.256, 0.544)	65%	0.348*	(0.195, 0.481)
$\beta_{D:T_2,i}$	ρ_{15}	0.313*	(0.156, 0.457)	88%	0.287*	(0.121, 0.446)	87%	0.313*	(0.162, 0.457)
$\beta_{T_1,i}$									
$\beta_{T_2,i}$	ρ_{23}	0.005	(-0.161, 0.183)	88%	0.029	(-0.147, 0.207)	86%	-0.000	(-0.172, 0.169)
$\beta_{D:T_1,i}$	ρ_{24}	0.531*	(0.406, 0.641)	47%	0.436*	(0.282, 0.566)	54%	0.517*	(0.382, 0.629)
$\beta_{D:T_2,i}$	ρ_{25}	0.095	(-0.077, 0.253)	66%	0.168	(-0.007, 0.346)	73%	0.107	(-0.069, 0.279)
$\beta_{T_2,i}$									
$\beta_{D:T_1,i}$	ρ_{34}	0.132	(-0.038, 0.304)	33%	0.296*	(0.127, 0.449)	35%	0.127	(-0.049, 0.303)
$\beta_{D:T_2,i}$	ρ_{35}	0.206*	(0.049, 0.344)	52%	0.092	(-0.075, 0.274)	53%	0.202*	(0.039, 0.347)
$\beta_{D:T_1,i}$									
$\beta_{D:T_2,i}$	ρ_{45}	-0.120	(-0.280, 0.052)	66%	-0.038	(-0.228, 0.145)	74%	-0.102	(-0.275, 0.078)

Note. Bold red font indicates changes of more than 10% in the median of the posterior distribution between neighboring models. "PDO" (posterior distribution overlap) indicates the overlap of the posterior distribution of neighboring models in percentage. The asterisk indicates a significant result using a 95% credibility interval.

$\gamma_{43} = 0.21 [0.07, 0.36]$ for grade one to two and $\gamma_{53} = 0.18 [0.06, 0.29]$ for grade two to four, respectively. A possible explanation could be that these children had a considerably higher viewing time for regular sentences in grade one in the first place. This explanation is also supported by the correlations ρ_{04} and ρ_{05} of the intercept of TVT_{wi} and the two change variables $\beta_{D:T_1,i}$ and $\beta_{D:T_2,i}$. The correlations are both positive, meaning that a child that had a high intercept (i.e., average viewing time across conditions), is likely to have a large change in reading efficiency.

Model 2: Mixed model

The second model that we consider is a linear mixed model. To maximize comparability of the results will also use Bayesian multilevel SEM in Mplus to estimate the model. The second model does not aggregate measurements within the six conditions, but makes use of the full dataset. This yields between 496 to 1,575 observations per participant (i.e., we had differing numbers of observations). In the Mplus framework, slopes (i.e., contrasts) in multilevel models are modeled as latent variables. The means of these latent

variables represent fixed effects, whereas their variances represent the variability of random effects. As described earlier, using a multilevel SEM has several advantages over the ANOVA model. First, we can obtain an estimate of the within residual variance. Second, we can utilize covariates to predict slopes (which represented are latent variables). That is, we estimate the effect of the covariates on the contrasts (i.e., this is an interaction effect). By regressing the latent slopes on the covariates, we also yield an estimate of the explained variance (i.e., R^2) of the contrasts. This is an indicator of how much of the interindividual differences in the contrasts are explained by the covariates.

In the multilevel SEM framework, we have to deviate from the previous notation. We no longer use a multivariate but a univariate model. We will make use of a slightly different notation:

$$\begin{aligned} TVT_{wi} = & \beta_{0,i} + \beta_{D,i}C_{D,wi} + \beta_{T_1,i}C_{T_1,wi} + \beta_{T_2,i}C_{T_2,wi} \\ & + \beta_{D:T_1,i}C_{D:T_1,wi} + \beta_{D:T_2,i}C_{D:T_2,wi} + \beta_{LP,i}LP_{wi} \\ & + \varphi_i TVT_{(w-1)i} + \varepsilon_{wi} \end{aligned} \quad (7)$$

$$\beta_{0,i} = \gamma_{00} + \gamma_{01}IQ_i^c + \gamma_{02}CBCL_i^c + \gamma_{03}RD_i^c + u_{0,i} \quad (8)$$

$$\beta_{D,i} = \gamma_{10} + \gamma_{11}IQ_i^c + \gamma_{12}CBCL_i^c + \gamma_{13}RD_i^c + u_{1,i} \quad (9)$$

$$\beta_{T_1,i} = \gamma_{20} + \gamma_{21}IQ_i^c + \gamma_{22}CBCL_i^c + \gamma_{23}RD_i^c + u_{2,i} \quad (10)$$

$$\beta_{T_2,i} = \gamma_{30} + \gamma_{31}IQ_i^c + \gamma_{32}CBCL_i^c + \gamma_{33}RD_i^c + u_{3,i} \quad (11)$$

$$\beta_{D:T_1,i} = \gamma_{40} + \gamma_{41}IQ_i^c + \gamma_{42}CBCL_i^c + \gamma_{43}RD_i^c + u_{4,i} \quad (12)$$

$$\beta_{D:T_2,i} = \gamma_{50} + \gamma_{51}IQ_i^c + \gamma_{52}CBCL_i^c + \gamma_{53}RD_i^c + u_{5,i} \quad (13)$$

$$\beta_{LP,i} = \gamma_{60} + u_{6,i} \quad (14)$$

$$\varphi_i = \gamma_{80} + u_{8,i} \quad (15)$$

Equation (7) expresses viewing time for child i and word w as a function of the previously defined contrasts. The β 's represent the random slopes. Equations (8) to (15) represent the latent slopes. We can see that the random slopes $\beta_{j,i}$ (i.e., j represents one of the six contrasts) are a function of an intercept γ_{jk} (the fixed effect, with $k \in \{0, 1, 2, 3\}$), a random effect $u_{j,i}$, and the three time-invariant covariates. Note that we do not actually estimate the random effects $u_{j,i}$ (i.e., each

participant's deviation from the intercept), but we only estimate their variance τ_{jj} .

We next examine the two contrasts of interest $\beta_{D:T_1,i}$ (Equation 12) and $\beta_{D:T_2,i}$ (Equation 13). In (12), the estimated mean change in reading efficiency from grade one to two is $\hat{\gamma}_{40} = 0.49$ [0.43, 0.54] and its variance is $\hat{\tau}_{44} = 0.13$ [0.10, 0.16]. In (13), the estimated mean change in reading efficiency from grade two to four is $\hat{\gamma}_{50} = 0.24$ [0.20, 0.29] and its variance is $\hat{\tau}_{55} = 0.08$ [0.06, 0.10]. One standard deviation below and above the average change is given by $\hat{\gamma}_{40} \pm \sqrt{\hat{\tau}_{44}} = [0.13, 0.84]$ for grade one to two and $\hat{\gamma}_{50} \pm \sqrt{\hat{\tau}_{55}} = [-0.04, 0.52]$. Furthermore, the multilevel model offers additional insights. The explained interindividual differences (R^2) of the contrasts drop significantly with $R^2(\beta_{D:T_1,i}) = 0.07$ [0.03, 0.13] and $R^2(\beta_{D:T_2,i}) = 0.05$ [0.02, 0.12] (changes of more than 10% are indicated by bold red font in results tables). A likely reason is that we now explicitly estimate the within residual variance ($\hat{\tau} = 0.29$ [0.29, 0.29]), which in the ANOVA model was absorbed into the variances of the contrasts. By separating within-person variability from between-person variability, less variance remains at the level of the random effects, which reduces the proportion of interindividual differences in the contrasts that can be explained by the covariates. At the same time, a substantial proportion of variance in the outcome at the observation level is explained overall, as indicated by $R^2(TVT_{wi}) = 0.40$ [0.40, 0.40]. However, the model does not allow us to directly examine whether and how the covariates predict the within residual variance itself—RDSEM, however, will give us the answer.

Model 3: RDSEM

The third model the RDSEM. RDSEM is a sophisticated analytical technique tailored for analyzing intensive longitudinal multilevel data, tracing the trajectory of both observed and latent variables across time (Asparouhov et al., 2018). As noted above, noteworthy advantages of RDSEM encompass: (1) integration of a time-series element permitting autoregressive influences, (2) delineation of between- and within-person effects via multilevel modeling, (3) adoption of a latent variable framework, (4) computational tractability facilitated by Bayesian estimation, (5) interindividual differences in within residual variances, and (6) utilization of latent group-mean centering. Diverging from conventional SEM, RDSEM is purposefully crafted for intensive longitudinal data sets and is adept at capturing intricate interrelationships among variables across temporal domains, rendering it well-suited for ecological

momentary and mobile assessment data (e.g., Trull & Ebner-Priemer, 2014; Conner & Barrett, 2012; Hamaker & Wichers, 2017; Mehl et al., 2023).

In this section, we will extend the previous models, allowing us to answer questions defined in the beginning of the article. In particular, the RDSEM model extends the LMM by an autoregressive effect which can vary across individuals, time-varying covariates, and interindividual differences in the within residual variance. The number of observations per participant was again 496 to 1,575.⁶ The model equation is given by:

$$\begin{aligned} TVT_{wi} = & \beta_{0,i} + \beta_{D,i}C_{D,wi} + \beta_{T_1,i}C_{T_1,wi} + \beta_{T_2,i}C_{T_2,wi} \\ & + \beta_{D:T_1,i}C_{D:T_1,wi} + \beta_{D:T_2,i}C_{D:T_2,wi} + \beta_{LP,i}LP_{wi} \\ & + \varepsilon_{wi} \end{aligned} \quad (16)$$

$$\beta_{0,i} = \gamma_{00} + \gamma_{01}IQ_i^c + \gamma_{02}CBCL_i^c + \gamma_{03}RD_i^c + u_{0,i} \quad (17)$$

$$\beta_{D,i} = \gamma_{10} + \gamma_{11}IQ_i^c + \gamma_{12}CBCL_i^c + \gamma_{13}RD_i^c + u_{1,i} \quad (18)$$

$$\beta_{T_1,i} = \gamma_{20} + \gamma_{21}IQ_i^c + \gamma_{22}CBCL_i^c + \gamma_{23}RD_i^c + u_{2,i} \quad (19)$$

$$\beta_{T_2,i} = \gamma_{30} + \gamma_{31}IQ_i^c + \gamma_{32}CBCL_i^c + \gamma_{33}RD_i^c + u_{3,i} \quad (20)$$

$$\beta_{D:T_1,i} = \gamma_{40} + \gamma_{41}IQ_i^c + \gamma_{42}CBCL_i^c + \gamma_{43}RD_i^c + u_{4,i} \quad (21)$$

$$\beta_{D:T_2,i} = \gamma_{50} + \gamma_{51}IQ_i^c + \gamma_{52}CBCL_i^c + \gamma_{53}RD_i^c + u_{5,i} \quad (22)$$

$$\beta_{LP,i} = \gamma_{60} + \gamma_{61}IQ_i^c + \gamma_{62}CBCL_i^c + \gamma_{63}RD_i^c + u_{6,i} \quad (23)$$

$$\varepsilon_{wi} = \varphi_i\varepsilon_{(w-1)i} + \delta_{wi} \quad (24)$$

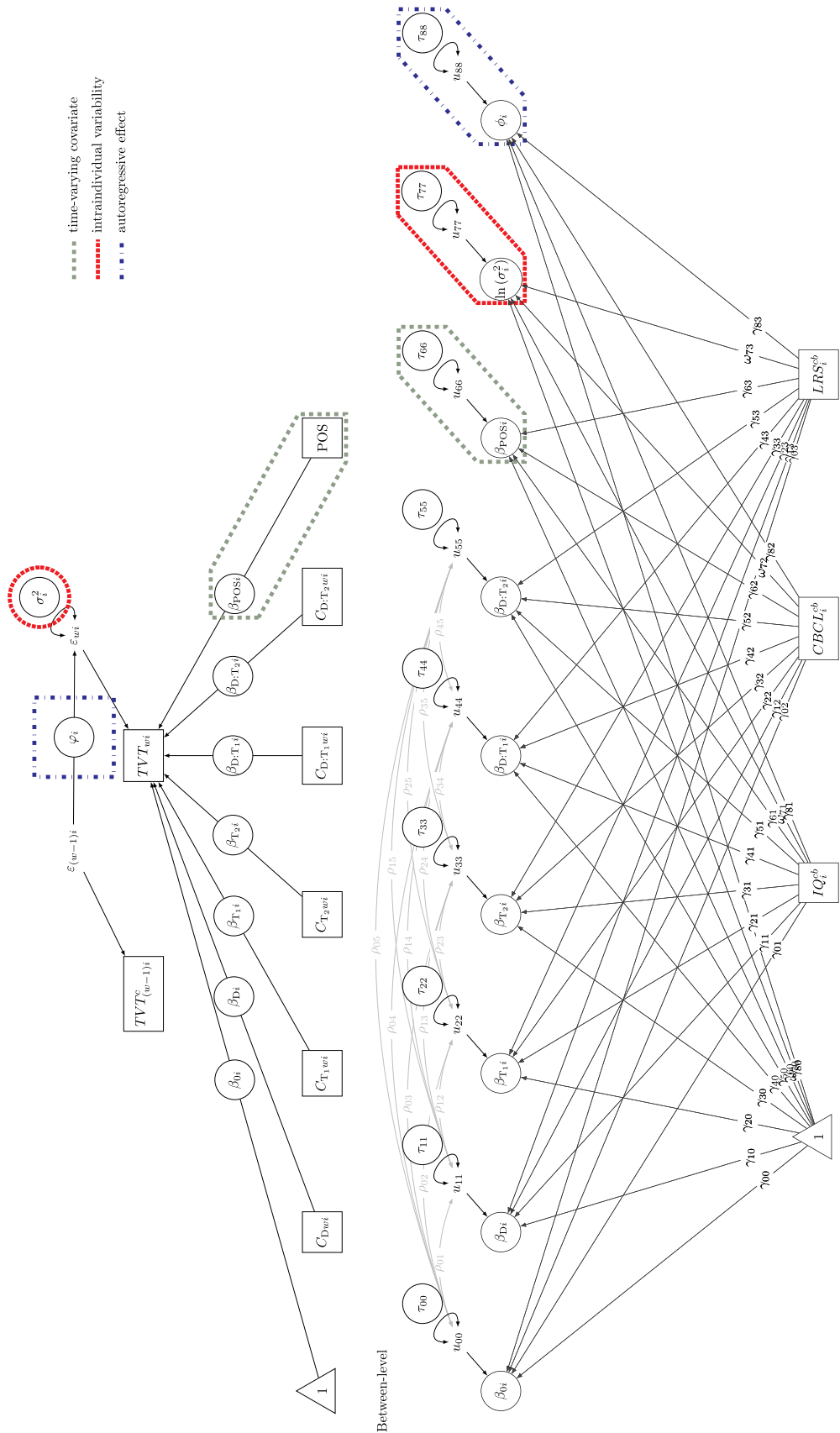
$$\sigma^2(\delta_{wi}) = \exp(\omega_{70} + \omega_{71}IQ_i^c + \omega_{72}CBCL_i^c + \omega_{73}RD_i^c + u_{7,i}) \quad (25)$$

$$\varphi_i = \gamma_{80} + \gamma_{81}IQ_i^c + \gamma_{82}CBCL_i^c + \gamma_{83}RD_i^c + u_{8,i} \quad (26)$$

⁶To prevent the first word of a sentence from being regressed on the last word of the previous sentence, we constructed a time index variable that was used with the Mplus TINTERVAL option. Within each child, observations were sorted by grade, sentence type, trial, and word. We then created a time index variable (tint) that increased by one unit for each successive word. At the beginning of a new trial, instead of increasing the index by one, we increased it by two. This creates a gap in the time index between the last word of one sentence and the first word of the next sentence. When the TINTERVAL option detects such a gap, Mplus internally inserts an additional time point with missing values. This artificial observation contains missing values for all time-varying variables, including the outcome and all contrast variables.

Although a more thorough discussion of the extensions will follow in subsequent sections, we would like to briefly highlight the modifications in the above equation. We can observe three changes in Equation (16). First, we introduced the time-varying covariate landing position through $\beta_{LP,i}LP_{wi}$ (see Equation 23). Second, we express the residual as a function of an autoregressive effect from the previous word's residual, $\varphi_i\varepsilon_{(w-1)i}$, and a residual δ_{wi} (see Equations 24 and 26). Third, we express the variance of the residual $\sigma^2(\delta_{wi})$ as a function of several covariates and further add a random effect to it $u_{7,i}$ (see also Equation 25). Figure 2 represents the full RDSEM, where colored boxes indicate the three extensions. Figure 1 includes the model-implied fixed and random effects of the RDSEM for the six experimental conditions.

A further issue concerns the specification of the time structure and the handling of missing data. In Mplus, the `tint` (time interval) option is used to define the distance between successive measurements. This option directly affects the estimate of the autoregressive parameter because the lag is scaled by the specified time interval. However, it does not affect the estimates of the experimental contrasts, since the structural part of the model is separated from the autoregressive component. This separation is a particular advantage of using RDSEM rather than DSEM in our application (see Section “Prior Models for Intensive Longitudinal Data in Experimental Contexts”), because we do not operate on a physical time scale but rather on sequences of words in a sentence. In this context, the autoregressive effect is interpreted as the dependency between subsequent words, rather than between observations separated by a fixed amount of time. At the same time, the RDSEM framework can naturally be applied to data with truly time-based spacing between measurements, where the interpretation of the autoregressive effect changes accordingly. Regarding missing data, it is important to note that the issue of missingness is not unique to our application but arises in the general RDSEM framework as well. In Mplus, missing data are treated as unknown parameters estimated within the Bayesian framework, which increases the number of estimated parameters. The underlying assumption is that data are missing at random (MAR), meaning that missingness may depend on observed variables but not on the missing values themselves (Muthén & Muthén, 1998-2017). The validity of this assumption must be carefully evaluated in each empirical application.



(Mohan & Pearl, 2021; Potthoff et al., 2006; Enders, 2022).

For the RDSEM, the fixed and the random effects of the two contrasts of interest (Equations 21 and 22) change only slightly compared with LMM. The mean change in reading efficiency from grade one to two is $\hat{\gamma}_{40} = 0.61 [0.54, 0.69]$ and its variance is $\hat{\tau}_{44} = 0.17 [0.13, 0.21]$. The mean change in reading efficiency from grade two to four is $\hat{\gamma}_{50} = 0.34 [0.29, 0.40]$ and its variance is $\hat{\tau}_{55} = 0.11 [0.08, 0.14]$. One standard deviation below and above the average change is given by $\hat{\gamma}_{40} \pm \sqrt{\hat{\tau}_{44}} = [0.21, 1.02]$ for grade one to two and $\hat{\gamma}_{50} \pm \sqrt{\hat{\tau}_{55}} = [0.02, 0.67]$. The point estimates of the fixed and the variance of the random effects do not change materially as compared to the LMM (bold red font in Table 1 and 2 indicate a change of more than 10% in the posterior median). A model comparison based on the deviance information criterion (DIC) indicated better fit for RDSEM relative to DSEM.

Time-varying covariates (landing position)

The first extension answers the question whether the initial landing position affects viewing time. We test this by incorporating the time-varying covariate LP_{wi} into the model. This is not possible in the ANOVA approach. The extension can only be incorporated in multilevel models, such as LMM and RDSEM. Previous research indicates that skilled readers tend to fixate closer to the center of a word on their first gaze (i.e., even though each word may be viewed multiple times, only the initial gaze is considered; Radach & McConkie, 1998). We thus expect that the landing position will have a positive association with reading efficiency. We also allow for interindividual differences (random effects) in the slope $\beta_{LP,i}$ across children. The mean of the slope is given by $\hat{\gamma}_{60} = 0.16 [0.15, 0.17]$ and its variance by $Var(u_{6,i}) = \hat{\tau}_{66} = 0.002 [0.001, 0.003]$. Because LP_{wi} ranges from 0 (center) to 1 (edge), the mean slope $\hat{\gamma}_{60} = 0.16$ represents the expected difference in viewing time between a perfectly centered fixation and one at the beginning or end of a word. One standard deviation below and above the average effect is given by $\hat{\gamma}_{60} \pm \sqrt{\hat{\tau}_{66}} = [0.11, 0.20]$, meaning that for some participants, the effect will be closer to 0.11 and for some, it will be closer to 0.20. In addition, we can calculate the amount of explained variance in the residual variance with $R^2(\beta_{LP,i}) = 0.09 [0.03, 0.19]$.

Heterogeneity in residual variances

The second extension answers whether some of the person characteristics (i.e., cluster-level covariates) affect the intraindividual variability. We first partition the residual into a part attributable to an autoregressive effect from the previous word (see the following section) and a remaining error term δ_{wi} . δ_{wi} , in turn, is a function of the time-invariant covariates. We will use a different letter (ω) to refer to these regression coefficients because we need to apply the exponential function (i.e., the function $\exp(x) = e^x$, which ensures that the predicted residual variance is strictly positive) to obtain an interpretable quantity of the coefficients. The intercept of the residual variance is given by $\exp \hat{\omega}_{70} = 0.30 [0.30, 0.31]$. Since the covariates in this equation are grand-mean centered, the value can be interpreted as the average within residual variance across participants. While IQ and CBCL do not affect the residual variance, RD does have an effect with $\exp \hat{\omega}_{73} = 1.10 [1.04, 1.16]$. If we subtract 1 from the estimate and multiply by 100 $((1.10 - 1) \times 100 = 10\%)$, we get the percentage increase in the residual variance. Children with RD will have a residual variance that is 10% larger as compared to children without RD. Lastly, we allow for a random effect in the residual variance. That is, we are modeling “between-person variance in the within-person residual variance”. The random effect can be interpreted as the interindividual differences in the within residual variance that cannot be explained by the covariates. The variability of the random effect is given by $Var(u_{7,i}) = \hat{\tau}_{77} = 0.03 [0.02, 0.03]$. The interpretation is not straightforward, because we cannot simply take the exponential function of this. We can, however, calculate the residual variance for a person that is one standard deviation above or below the average while setting the covariates to zero (i.e., setting the covariates to the sample average): $\exp \hat{\omega}_{70} \pm \sqrt{\hat{\tau}_{77}} = [0.26, 0.35]$. We can also express the change in within residual variance in terms of percentage by $\exp \sqrt{\hat{\tau}_{77}} = 1.17$, meaning that a person’s within residual variance increases by 17% if this person is one standard deviation above average. Lastly, we can calculate the amount of explained variance in the residual variance with $R^2(\sigma^2(\delta_{wi})) = 0.04 [0.02, 0.10]$.

Autoregressive effect

The third extension answers whether the viewing time of the previous word ($TVT_{(w-1)i}$) will affect reading of the current word. We can investigate this by incorporating an autoregressive (lagged) effect of viewing

time into the model. There are two ways of modeling this effect in Mplus. First, we could grand-mean center viewing time of the previous word and use it as a predictor. Second, we estimate the residual of the previous word ($\varepsilon_{(w-1)i}$) and use that to predict the residual of the current word. Recall, RDSEM differs from regular DSEM in the sense that autoregressive effects are not estimated for the dependent variable itself but for the residual of the dependent variable. RDSEM provides a superior approach if a time trend is part of the model. We did not include a trend but we will use of this approach in this manuscript because it generalizes more easily to models that include such a trend. The relevant line can be found in Equations (24) and (26). The mean of the autoregressive effect is $\hat{\gamma}_{80} = 0.16$ [0.14, 0.17] and its variance is $\hat{\tau}_{88} = 0.004$ [0.003, 0.005]. The variance of the slope is negligibly small, so we will focus on the interpretation of its mean γ_{80} . If the residual of the previous word exceeds the average viewing time by one unit, the residual of the current word will be 0.16 units larger. This could happen if, for instance, the previous word was particularly difficult and it took the children longer to process the word. Recall that viewing time was standardized by the number of characters of the words. Thus, it is not word length that accounts for changes in viewing time, but other factors such as difficulty or familiarity.

We can further improve the interpretation of γ_{80} by calculating the expected autoregressive effect ($\varphi_i \varepsilon_{(w-1)i}$) if the previous word's viewing time was one standard deviation above average. For a child with average IQ, CBCL, and RD (i.e., setting these variables to zero because they are centered), the average standard deviation of the residual is $SD(\varepsilon_{wi} | IQ_i^c = CBCL_i^c = RD_i^c = 0) \approx 0.56$.⁷ Remember, the average autoregressive effect (i.e., fixed effect) is $\hat{\gamma}_{80} = 0.16$. Therefore, if the previous word's residual was one standard deviation above average, the residual of the current word will be $0.16 \cdot 0.56 = 0.09$ units higher.

Lastly, we regressed the autoregressive effect on the three time-invariant covariates, which all did not have a significant effect. This is in line with the rather small explained variance of the autoregressive effect $R^2(\varphi_i) = 0.02$ [0.003, 0.059].

Posterior distribution overlap across models

Another contribution of this tutorial is to illustrate the use of posterior distribution overlap (PDO) to

compare results across models. Tables 1 and 2 compared the results across the three models. PDO quantifies the similarity of two posterior distributions as the proportion of shared area. A PDO close to 100% indicates nearly identical posteriors, while smaller values indicate differences in central tendency, uncertainty, or both. This is important because point estimates or credible intervals alone can conceal meaningful differences in the full distributions. Figure 3 visualizes the overlap of each of the fixed effect parameters for the LMM and the RDSEM estimates.

The PDO may especially be valuable for applied researchers because it highlights how different modeling choices can influence parameter estimates and, in turn, the outcome of an analysis. Ultimately, the aim of the analysis is to make substantive statements about developmental processes, not only to report statistical significance. If the PDO between two reasonable models is very low, the resulting substantive statements may differ. For example, one model may suggest that the change in reading efficiency from grade one to two is large and highly certain, whereas another model yields a posterior that barely overlaps with the first. In such a case, an applied researcher might draw different conclusions about the magnitude and even the direction of developmental change. In this sense, PDO provides a tool for sensitivity analysis. It should not be used for model selection. Model choice should be guided by substantive arguments and, where available, by fit indices. We note that RDSEM lacks many conventional fit indices, with the exception of the DIC. The practical value of PDO is to shift attention away from p -values and toward the magnitude of effects and the certainty around them, which is what matters for substantive interpretation.

ANOVA vs. LMM

The tables indicate marked differences for several fixed effects when moving from ANOVA (aggregated condition means) to LMM (full data). This is also indicated by the percentage of overlap of the PDO. For the two developmental contrasts that are central to this article, the PDO is $PDO_{\gamma_{40}} = 24\%$ for the change from grade one to two ($\beta_{D:T_1,i}$) and $PDO_{\gamma_{50}} = 21\%$ for the change from grade two to four ($\beta_{D:T_2,i}$). Other fixed effects show similar patterns: $PDO_{\gamma_{10}} = 14\%$ ($\beta_{D_1,i}$), $PDO_{\gamma_{20}} = 17\%$ ($\beta_{T_1,i}$), and $PDO_{\gamma_{30}} = 42\%$ ($\beta_{T_2,i}$). The intercept shows higher overlap with $PDO_{\gamma_{00}} = 83\%$. These values indicate that using all observations in a multilevel framework can meaningfully change fixed-effect estimates of the developmental contrasts relative to aggregated ANOVA; in contrast, the overall level of the outcome is more stable.

⁷The stationary variance of the AR(1) residual process is $\sigma^2(\delta_{wi})/(1 - \varphi_i^2)$. With $\sigma^2(\delta_{wi}) = \exp(\hat{\omega}_{70}) = 0.303$ and $\varphi_i = \hat{\gamma}_{80} = 0.155$ (evaluated at centered covariates), this gives $\text{Var}(\varepsilon_{wi}) \approx 0.310$, so $SD \approx 0.56$.

LMM vs. RDSEM

Introducing residual autoregression and person-specific residual variances in RDSEM changes the fixed developmental contrasts further. In the tables, the overlap between LMM and RDSEM is very small for the two developmental contrasts: $\text{PDO}_{\gamma_{40}} = 5\%$ and $\text{PDO}_{\gamma_{50}} = 5\%$. As a comparison, when comparing the means of two samples, Cohen's $d = 0.5$ would equal an overlap of around 80% assuming equal variances. This indicates a considerable difference in either the point estimate or certainty of the means. As shown in Table 1 (Figure 3), several other fixed effects also exhibit small overlaps across these models (for example, $\text{PDO}_{\gamma_{10}} = 0\%$, $\text{PDO}_{\gamma_{20}} = 7\%$, $\text{PDO}_{\gamma_{30}} = 15\%$), whereas some parameters show higher stability, such as the landing-position slope with $\text{PDO}_{\gamma_{60}} = 76\%$ and the autoregressive effect with $\text{PDO}_{\gamma_{80}} = 93\%$ (Table 1). This pattern indicates that modeling residual autoregression primarily alters the fixed developmental contrasts, whereas parameters that capture local dynamics (for example, γ_{60} and γ_{80}) are much more similar across LMM and RDSEM.

Figure 3 confirms the table-based conclusions. Using the posterior draws underlying the figures, the fixed-effect posteriors for the developmental contrasts show very small overlaps and sizable median shifts. For γ_{40} we observe $\Delta\text{Median}_{\gamma_{40}} = 27\%$ with $\text{PDO}_{\gamma_{40}} \approx 5\%$, and for γ_{50} we observe $\Delta\text{Median}_{\gamma_{50}} = 42\%$ with $\text{PDO}_{\gamma_{50}} \approx 5\%$.

Figure 4 provides a visualization of the differences between LMM and RDSEM in the posterior effect sizes. We derived a naive η_p^2 for each contrast j as $\eta_p^2(\beta_{j,i}) = \frac{\gamma_{j0}^2}{\gamma_{j0}^2 + \tau_{jj}}$, which summarizes the size of the fixed effect relative to the total between-person variability for that contrast.⁸ Using this definition, Figure 4 shows that $\eta_p^2(\beta_{D:T_1,i})$ has $\Delta\text{Median} = 7\%$ with $\text{PDO} = 58\%$, whereas $\eta_p^2(\beta_{D:T_2,i})$ has $\Delta\text{Median} = 23\%$ with $\text{PDO} = 36\%$. These results make clear that even if two models would both yield small p -values for a contrast, they can still yield different standardized effect sizes once the full posteriors are considered. For an applied researcher, these findings signal that statements about the size of developmental change are sensitive to whether residual dynamics are modeled.

Summary

In sum, PDO and Δ Median together provide a transparent sensitivity analysis across modeling choices.

⁸We call this naive because there is vast literature about effect size estimates in multilevel models and RDSEM is a special and rather complicated case. This version of η_p^2 may not be the most sensitive effect size measure but one that is easy to compute and intuitively interpretable.

First, several fixed effects, and in particular the developmental contrasts γ_{40} and γ_{50} , change materially in magnitude when moving from ANOVA to LMM, and change even more when moving from LMM to RDSEM, as evidenced by low PDO values and substantial Δ Median in the figures. Second, parameters capturing local dynamics, such as γ_{60} and γ_{80} , show high overlap across models. For applied researchers, high PDO and small Δ Median suggest that different methods support the same substantive conclusions, whereas low PDO with large Δ Median indicates that substantive conclusions may change with the modeling approach. PDO should therefore be used to assess robustness of conclusions across different modeling approaches, not to select models.

Discussion and concluding remarks

This study illustrated the application of RDSEM to intensive repeated measures data collected in laboratory settings, using the example of a reading efficiency experiment with children. Previous applications of RDSEM have often focused on ecological momentary assessment or daily diary data. However, both DSEM and RDSEM were explicitly designed for intensive longitudinal data, including contexts with short intervals between observations (e.g., seconds, minutes, or hours). More recently, mobile sensing approaches have expanded the scope of intensive longitudinal data collection, for example through continuous sensor-based measurements of behavior and physiology (Mehl et al., 2023). Our study illustrates that RDSEM can also be applied in laboratory experiments with temporal proximity of milliseconds, highlighting the flexibility of the framework across different time scales and application domains.

Comparison of results

ANOVA

When analyzing the reading data, ANOVA provided only aggregate-level estimates, summarizing the average effects of sentence type (Landolt vs. regular), grade level and their interaction on viewing times. ANOVA also provides estimates of interindividual differences (i.e., random effects), represented by variances of contrast variables. Although ANOVA detected general trends and variability in reading efficiency, such as improvement across grades and how improvement differs between children, it did not account for individual variability or dynamic relationships between consecutive observations (e.g., how reading the previous word affects reading the current one). By focusing on group averages,

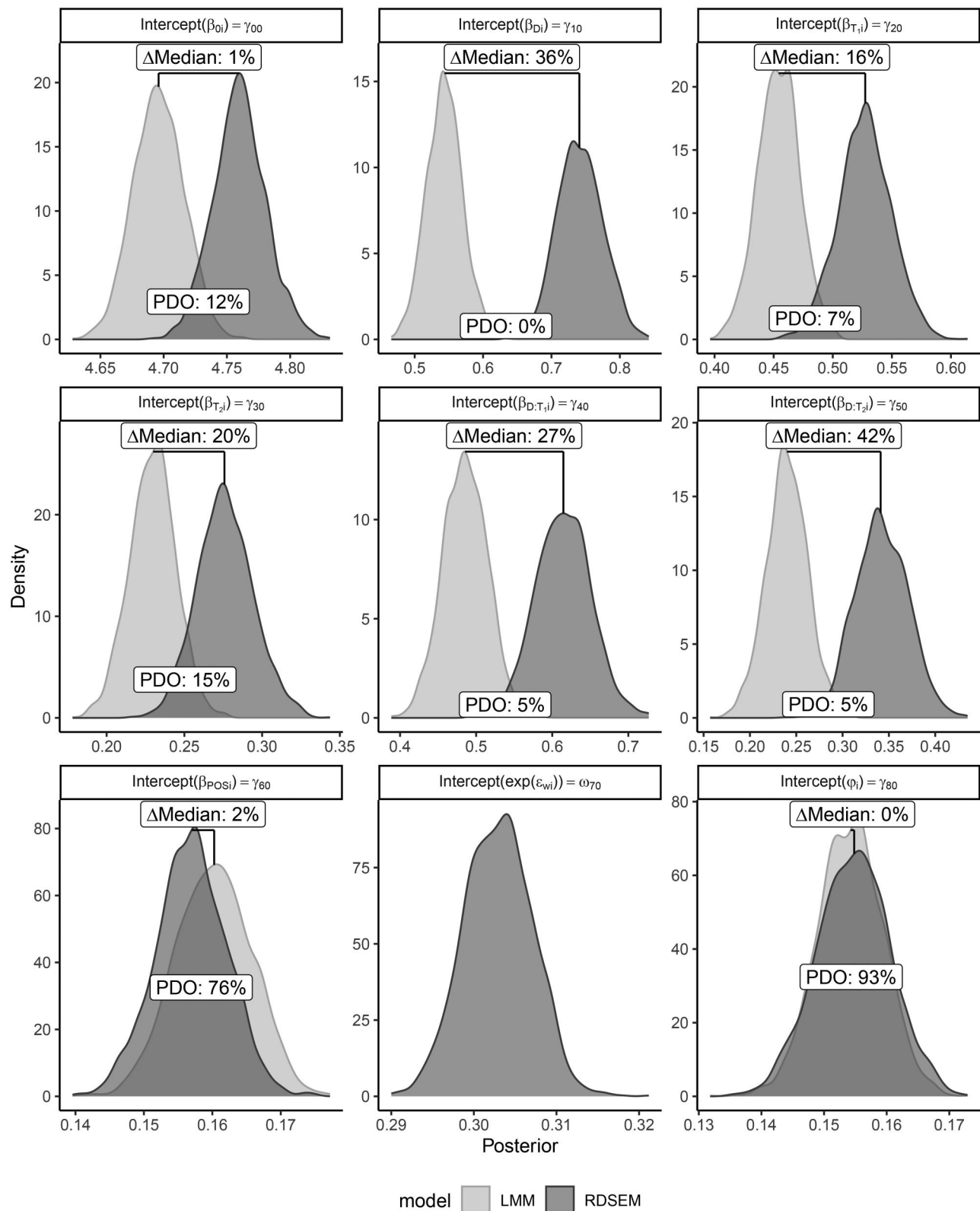


Figure 3. Posterior distributions of fixed effects in the LMM (light gray) and RDSEM (dark gray). For each parameter, the change in posterior medians (Δ Median) and the percentage of distributional overlap (PDO) are reported.

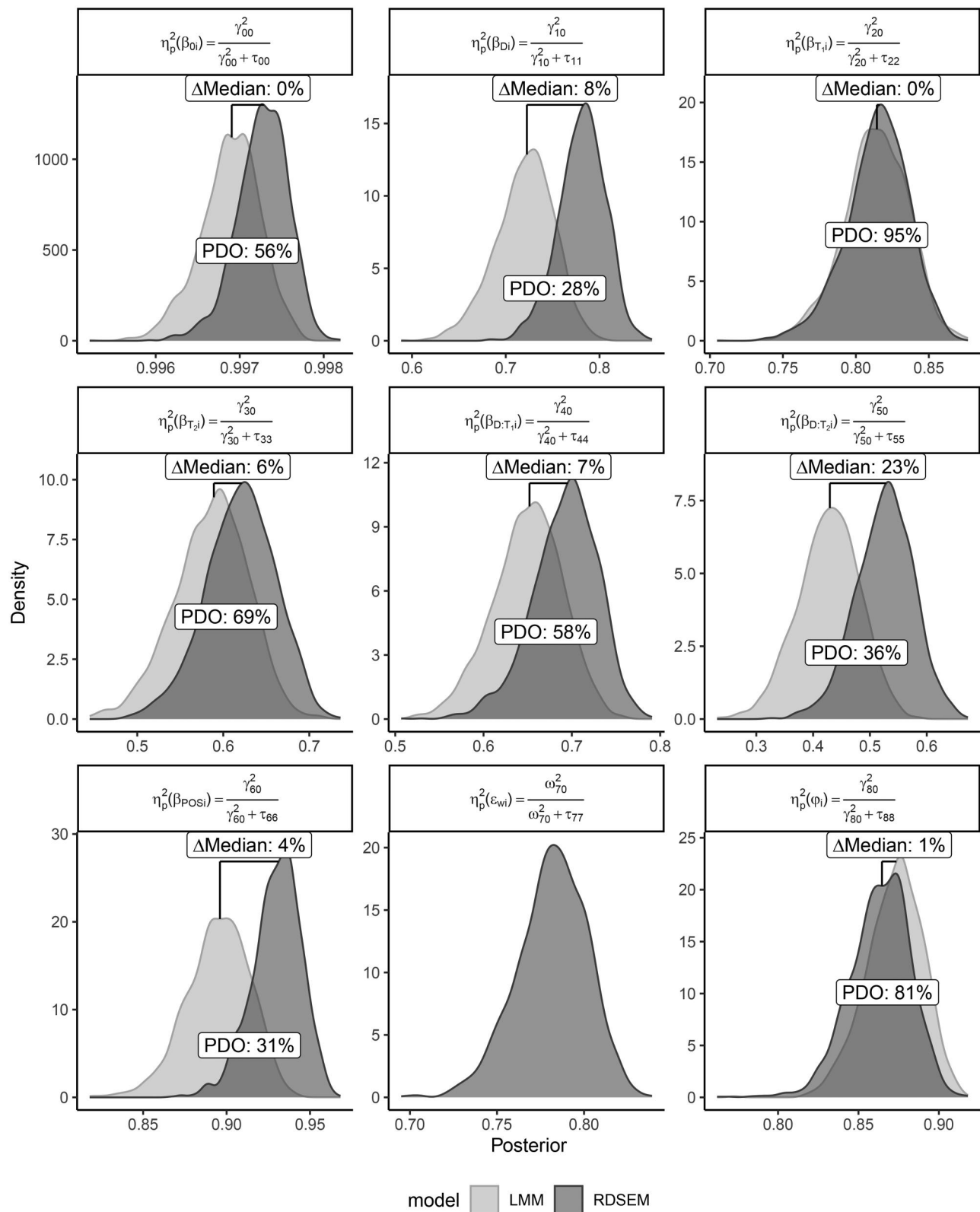


Figure 4. Posterior distributions of naive effect size measure η_p^2 in the LMM (light gray) and RDSEM (dark gray). For each parameter, the change in posterior medians (Δ Median) and the percentage of distributional overlap (PDO) are reported.

ANOVA simplifies the data structure, but overlooks potentially important within-person processes. Additionally, ANOVA is not able to distinguish between interindividual differences and residual variance (which is absorbed by the variances of the contrasts) because it is only a single-level approach.

LMM

LMM improved on ANOVA by distinguishing between variance of random effects and residual variance, allowing for a more nuanced estimation of individual differences. This approach showed that while children generally improved in reading efficiency over time, there was significant variability in how much each child improved. However, LMM still treated the data in a way that did not fully capture the complexity of the reading process, particularly when it came to handling time-varying predictors and modeling the influence of previous words on current reading behavior.

RDSEM

RDSEM provided the most comprehensive analysis by building on LMM's ability to estimate both fixed and random effects and adding the capacity to model time-varying covariates and autoregressive processes. RDSEM revealed several important dynamics in the data that were not captured by the other models. For instance, it showed that a child's viewing time for a word was influenced by the difficulty of the previous word, highlighting an autoregressive effect. This insight into the temporal dependency between words was not captured by either ANOVA or LMM. Although autoregressive effects may be modeled through LMMs, RDSEM adds to this by allowing for modeling autoregressive effects between residuals. Moreover, RDSEM modeled how landing position (i.e., where a child first fixates on a word) affected viewing time on a word-by-word basis, adding another layer of detail to the analysis.

In addition to these findings, RDSEM allowed us to explore inter-individual differences in intra-individual variability. Children with reading difficulties exhibited greater variability in their reading performance across trials. This result indicated that such children's viewing times fluctuate more from word to word, potentially signaling greater difficulty in maintaining consistent reading strategies. Such insights cannot be obtained from ANOVA or LMM.

RDSEM stands out for its ability to model random effects not only in main and interaction effects but also in autoregressive effects and residual variances. Previous work, as mentioned in Section "Prior Models

for Intensive Longitudinal Data in Experimental Contexts", has explored models incorporating random effects in autoregressive processes and residual variances independently. However, to our knowledge, no prior research has combined these features into a single, unified framework. RDSEM is the first model to naturally incorporate all of these elements—inter-individual variability, time-varying covariates, autoregressive effects, and random effects in residual variances—within a comprehensive modeling structure. This integration allows for a deeper and more accurate understanding of the dynamic processes within individuals, offering unparalleled insight into both fixed and random effects in experimental data.

Although LMMs can in principle be extended to include autoregressive error structures (e.g., Chi & Reinsel, 1989) as well, in such cases the autoregression is imposed directly on the observed-level errors. In contrast, RDSEM defines autoregression between residuals after accounting for predictors, which avoids conflating systematic effects with serial dependence and at the same time allows random effects to be modeled in a latent-variable framework. To keep the comparison transparent and to perform the comparison within a single statistical software, we therefore relied on the standard LMM specification in Mplus in this tutorial and highlighted the additional contributions of RDSEM when modeling autoregressive processes.

Generalization

The benefits of RDSEM extend far beyond this reading experiment. Its ability to model dynamic processes, time-varying covariates, and individual variability makes it highly applicable to a wide range of experimental settings. In cognitive neuroscience, RDSEM could model how brain activity measured *via* EEG or fMRI evolves over the course of a cognitive task, capturing both within-person fluctuations in neural responses and how these are influenced by previous stimuli or task demands (e.g., McCormick & Kievit, 2023). RDSEM allows researchers to capture the dynamic interplay between time-varying factors and individual differences in ways that traditional methods cannot. Given these advantages, we recommend that researchers consider RDSEM for analyzing experimental data where such dynamic and individualized patterns are of interest.

Another advantage of RDSEM (and of Mplus) is its flexibility to handle multivariate outcomes. For example, in the motivating study, the viewing times for Landolt and regular sentences could be treated as separate variables if their residual variances differ

significantly. Although this possibility was not explored in the current manuscript, it presents an interesting direction for future research. Moreover, the dependent variables do not need to be continuous, as Mplus allows for the inclusion of both continuous and categorical outcomes. This feature could be particularly useful when examining how error rates (i.e., binary outcomes) interact with viewing time.

TINTERVAL option in Mplus

In the present application, we used the `TINTERVAL` option in Mplus to ensure that the autoregressive residual process does not carry over across independent sentences. This required the construction of a time index variable that introduces gaps between trials. When such a gap is detected, Mplus internally inserts a missing time point in the data. At this artificial time point, all time-varying variables are treated as missing and are estimated within the Bayesian framework as unknown quantities. This behavior has important implications. In standard DSEM, where lagged relations may involve both outcomes and time-varying predictors, inserted missing values can affect not only autoregressive parameters but also time-varying covariates. In such cases, Mplus samples missing covariate values from their posterior distribution. If those covariates represent experimentally defined contrasts, this may lead to a loss of their intended interpretation, particularly when lagged predictors or higher-order lags are specified.

In contrast, our RDSEM specification uses contemporaneous regression of the outcome on the contrast variables and models autoregression only at the residual level. The artificial time points introduced by `TINTERVAL` contain missing values for both the outcome and all contrast variables. Because no observed data exist at these time points, they do not contribute directly to the estimation of fixed effects for the contrasts. Instead, they serve only to interrupt the autoregressive residual chain between sentences. For this reason, the contrast coding for all observed time points remains unchanged and interpretable. This reasoning applies to the RDSEM specification used in the present tutorial, which includes contemporaneous regression of the outcome on the contrast variables and a first-order autoregressive residual process. When lagged predictors or higher-order autoregressive effects (e.g., AR(2) or higher) are specified, the artificial time points introduced by `TINTERVAL` may interact with the dynamic structure in more complex ways. In such cases, researchers

should verify that the intended interpretation of time-varying predictors is preserved.

To verify this behavior, we exported the expanded data structure and confirmed that all inserted rows contained only missing values for the outcome and contrast variables, while all observed rows retained their original contrast coding. In addition, we generated artificial data under known parameter values and estimated the RDSEM model using the same `TINTERVAL` procedure. The true generating parameters were recovered with good accuracy for a single example with sufficiently large sample size. The R code used to generate the data, the dataset used in the manuscript, and the Mplus input and output files are available on the OSF repository.

Future directions

Future research should explore quantifying effect sizes of experimental factors in RDSEM. This is not straightforward in hierarchical data due to the presence of autoregressive effects and random effects in the within-subject residual variance. Additionally, the model could be expanded to account for multiple random factors in the design, such as stimuli in cognitive psychology experiments. Another key area for development is providing more comprehensive guidance on specifying different types of contrasts in RDSEM, as the choice of contrast can significantly affect the interpretation of parameters, particularly autoregressive effects and residual variances. For instance, when the intercept is defined as a reference category rather than the overall average, the interpretation of these effects can shift, making it essential to clarify these distinctions in future tutorials. Another promising direction would be to use RDSEM for multilevel mediation analyses. For example, in our motivating study one could test whether children's IQ influences reading efficiency indirectly through their landing position on words (IQ $\rightarrow$ landing position $\rightarrow$ total viewing duration). Such models would allow researchers to examine not only the direct effects of stable individual characteristics, but also their indirect effects through time-varying processes. As in all mediation analyses, care must be taken to rule out the effects of potential confounding variables, a challenging task (VanderWeele, 2015; MacKinnon, 2008). Lastly, an important goal would be to offer open-source software for estimating the models proposed in this article. Currently, we rely on commercial software like Mplus, and while the code is available in the [online supplemental materials](#), an Mplus license is required. Open-source tools would broaden access to these advanced methodological techniques,

allowing for wider use across the research community. There are attempts to developing open-source software for modeling intensive longitudinal data. These include the *brms* R package (Bürkner, 2017) and the recent *mlts* R package (Kosłowski et al., 2024), which both rely on the Stan modeling language (Stan Development Team, 2023a, 2023b).

Beyond their advantages, both LMMs and RDSEMs also rely on specific assumptions that should be considered when interpreting results. LMMs assume, for example, normally distributed random effects and homoscedastic residuals, and results can be sensitive to centering choices when time-varying predictors are included. In our RDSEM analyses estimated with Bayesian methods in *Mplus*, time-varying predictors are handled using latent group-mean centering. This approach decomposes each predictor into a within-person component and a latent between-person component, thereby separating within- and between-level effects at the model level. In contrast to observed group-mean centering, the latent centering approach accounts for measurement error in cluster means and helps avoid biases that can arise with small or unbalanced cluster sizes and with missing data. The contrast variables in the present study were entered in their original coding and were not manually centered. Their effects are estimated at the within level of the model, while between-person variability is captured at the between level through the latent decomposition. Time-invariant covariates were grand mean centered to facilitate interpretation of the intercept.

The broader literature discusses several issues that are relevant when specifying and interpreting categorical predictors in multilevel models. For example, with unbalanced data, coding schemes can imply weighted versus unweighted mean comparisons, which may affect substantive interpretation (Cohen et al., 2003). In multilevel settings, failure to properly separate within- and between-cluster components may lead to conflated or biased estimates, particularly when cluster sizes are small or unequal (Enders, 2013; Yaremych et al., 2023). For categorical predictors and interactions with continuous variables, interpretation depends directly on the chosen coding scheme and centering strategy (West et al., 1996). The latent centering framework implemented in Bayesian RDSEM addresses many of these concerns by explicitly modeling within- and between-level components and by accommodating missing data under standard model assumptions. Because our focus is on the implementation of custom contrasts within RDSEM, we limit ourselves to this clarification of how coding and centering operate in the present modeling

context and refer readers to the cited literature for a more general treatment.

RDSEM adds further assumptions that are specific to time-series modeling. In particular, RDSEM assumes stationarity, meaning that autoregressive effects remain constant across time, and homoscedasticity, meaning that the innovation variance is constant across both time and experimental conditions. In our application, this implies that the experimental manipulations are modeled as influencing only the mean level of the outcome but not the temporal dynamics or the residual variance. Yet, it is possible that conditions also affect these aspects, which would represent a violation of the assumptions if equality is imposed across conditions. One possible extension to address this limitation would be to formulate a multivariate RDSEM in which each experimental condition is represented as a separate variable. Stationarity and homoscedasticity would still be assumed within each condition, but parameters such as the residual variance or autoregressive effect could be allowed to differ across conditions, while others could remain constrained to equality. This approach would relax the across-condition constraints and provide a more flexible framework for capturing condition-specific dynamics.

In conclusion, RDSEM offers significant advantages over traditional methods like ANOVA and LMM. By capturing both within-person dynamics and between-person variability, RDSEM provides a deeper and more comprehensive understanding of experimental data, making it a powerful tool for researchers in various fields.

Article information

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.

Funding: This work was supported by Grant MA 7702/1-2 from the German Research Foundation (DFG) and Pathway to Independence Award R00AA030052 (PI:

McCabe) from the National Institute on Alcohol Abuse and Alcoholism (NIAAA, NIH)

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.

Acknowledgments: The ideas and opinions expressed in this manuscript are those of the authors alone, and endorsement by the authors' institutions or the funding agencies is not intended and should not be inferred.

Funding

The author(s) reported there is no funding associated with the work featured in this article. This work was supported by Grant MA 7702/1-2 from the German Research Foundation (DFG) and Pathway to Independence Award R00AA030052 (PI: McCabe) from the National Institute on Alcohol Abuse and Alcoholism (NIAAA, NIH)

References

- Abelson, R. P., & Prentice, D. A. (1997). Contrast tests of interaction hypotheses. *Psychological Methods*, 2(4), 315–328. <https://doi.org/10.1037/1082-989X.2.4.315>
- Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for ASEBA school-age forms & profiles: An integrated system of multi-informant assessment*. University of Vermont, Research Center for Children, Youth & Families.
- Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(3), 359–388. <https://doi.org/10.1080/10705511.2017.1406803>
- Asparouhov, T., & Muthén, B. (2020a). Advances in Bayesian model fit evaluation for structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 28(1), 1–14. <https://doi.org/10.1080/10705511.2020.1764360>
- Asparouhov, T., & Muthén, B. (2020b). Comparison of models for the analysis of intensive longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(2), 275–297. <https://doi.org/10.1080/10705511.2019.1626733>
- Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, 28(2), 135–167. <https://doi.org/10.3102/10769986028002135>
- Williams, D. R., Martin, S. R., Liu, S., & Rast, P. (2020). Bayesian multivariate mixed-effects location scale modeling of longitudinal relations among affective traits, states, and physical activity. *European Journal of Psychological Assessment: Official Organ of the European Association of Psychological Assessment*, 36(6), 981–997. <https://doi.org/10.1027/1015-5759/a000624>
- Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, 54(1), 579–616. <https://doi.org/10.1146/annurev.psych.54.101601.145030>
- Bürkner, P. (2017). brms: An R package for bayesian multi-level models using Stan. *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Chi, E. M., & Reinsel, G. C. (1989). Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association*, 84(406), 452–459. <https://doi.org/10.1080/01621459.1989.10478790>
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Routledge. <https://doi.org/10.4324/9780203774441>
- Conner, T. S., & Barrett, L. F. (2012). Trends in ambulatory self-report: The role of momentary experience in psychosomatic medicine. *Psychosomatic Medicine*, 74(4), 327–337. <https://doi.org/10.1097/psy.0b013e3182546f18>
- Driver, C. C., Oud, J. H. L., & Voelkle, M. C. (2017). Continuous time structural equation modeling with r package ctsem. *Journal of Statistical Software*, 77(5), 1–35. <https://doi.org/10.18637/jss.v077.i05>
- Enders, C. K. (2013). Centering predictors and contextual effects. In *The sage handbook of multilevel modeling* (pp. 89–108). SAGE Publications Ltd. <https://doi.org/10.4135/9781446247600.n6>
- Enders, C. K. (2022). *Applied missing data analysis* (2nd ed.). The Guilford Press.
- Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis* (2nd ed.). John Wiley & Sons, Inc.
- Hamaker, E. L., & Grasman, R. P. P. P. (2014). To center or not to center? Investigating inertia with a multilevel autoregressive model. *Frontiers in Psychology*, 5, 1492. <https://doi.org/10.3389/fpsyg.2014.01492>
- Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science*, 26(1), 10–15. <https://doi.org/10.1177/0963721416666518>
- Hedeker, D., & Nordgren, R. (2013). MIXREGLS: A program for mixed-effects location scale analysis. *Journal of Statistical Software*, 52(12), 1–38. <https://doi.org/10.18637/jss.v052.i12>
- Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2008). An application of a mixed-effects location scale model for analysis of ecological momentary assessment (ema) data. *Biometrics*, 64(2), 627–634. <https://doi.org/10.1111/j.1541-0420.2007.00924.x>
- Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2012). Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. *Statistics in Medicine*, 31(27), 3328–3336. <https://doi.org/10.1002/sim.5338>
- 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 & 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, 384. <https://doi.org/10.3389/fnhum.2013.00384>
- Hoffman, L. (2007). Multilevel models for examining individual differences in within-person variation and covariation over time. *Multivariate Behavioral Research*, 42(4), 609–629. <https://doi.org/10.1080/00273170701710072>
- Koslowski, K., Münch, F., Koch, T., & Holtmann, J. (2024). mlts: Multilevel latent time series models with R and Stan (Version 1.0.0) [Computer software].
- Kwok, O.-m., West, S. G., & Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. *Multivariate Behavioral Research*, 42(3), 557–592. <https://doi.org/10.1080/00273170701540537>
- Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4), 963–974. <https://doi.org/10.2307/2529876>
- Langenberg, B., Helm, J. L., & Mayer, A. (2022). Repeated measures ANOVA with latent variables to analyze inter-individual 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. (2024). Bayesian analysis of multi-factorial experimental designs using SEM. *Multivariate Behavioral Research*, 59(4), 716–737. <https://doi.org/10.1080/00273171.2024.2315557>
- Lenhard, W., & Schneider, W. (2006). *ELFE 1-6. Ein leseverständnistest für erst- bis sechstklässler*. Hogrefe.
- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis* (1st ed.). Erlbaum.
- Mayer, A., Geiser, C., Infurna, F. J., & Fiege, C. (2013). Modelling and predicting complex patterns of change using growth component models: An application to depression trajectories in cancer patients. *European Journal of Developmental Psychology*, 10(1), 40–59. <https://doi.org/10.1080/17405629.2012.732721>
- Mayer, A., Steyer, R., & Mueller, H. (2012). A general approach to defining latent growth components. *Structural Equation Modeling: A Multidisciplinary Journal*, 19(4), 513–533. <https://doi.org/10.1080/10705511.2012.713242>
- McCormick, E. M., & Kievit, R. A., Cambridge Centre for Ageing and Neuroscience (2023). Poorer white matter microstructure predicts slower and more variable reaction time performance: Evidence for a neural noise hypothesis in a large lifespan cohort. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 43(19), 3557–3566. <https://doi.org/10.1523/jneurosci.1042-22.2023>
- Mehl, M. R., Eid, M., Wrzus, C., Harari, G. M., & Ebner-Priemer, U. W. (Eds.). (2023). *Mobile sensing in psychology: Methods and applications*. The Guilford Press.
- Mohan, K., & Pearl, J. (2021). Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534), 1023–1037. <https://doi.org/10.1080/01621459.2021.1874961>
- Moll, K., & Landerl, K. (2010). *SLRT-II: Lese- und Rechtschreibtest; Weiterentwicklung des salzburger lese- und rechtschreibtests (slrt)*. Huber.
- Muthén, L. K., Muthén, B. O. (1998–2017). *Mplus user's guide. Eighth edition*. Muthén & Muthén. https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf
- Nestler, S. (2022). An extension of the mixed-effects growth model that considers between-person differences in the within-subject variance and the autocorrelation. *Statistics in Medicine*, 41(3), 471–482. <https://doi.org/10.1002/sim.9280>
- Nestler, S., & Blozis, S. A. (2024). A latent variable mixed-effects location scale model that also considers between-person differences in the autocorrelation. *Statistics in Medicine*, 43(1), 89–101. <https://doi.org/10.1002/sim.9943>
- Potthoff, R. F., Tudor, G. E., Pieper, K. S., & Hasselblad, V. (2006). Can one assess whether missing data are missing at random in medical studies? *Statistical Methods in Medical Research*, 15(3), 213–234. <https://doi.org/10.1191/0962280206sm448oa>
- R Core Team. (2021). *R: A language and environment for statistical computing* [Computer software].
- Radach, R., & McConkie, G. W. (1998). Determinants of fixation positions in words during reading. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 77–100). Elsevier.
- Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparison in the analysis of variance*. Cambridge University Press.
- 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>
- Rovine, M. J., & Molenaar, P. C. M. (2000). A structural modeling approach to a multilevel random coefficients model. *Multivariate Behavioral Research*, 35(1), 51–88. https://doi.org/10.1207/S15327906MBR3501_3
- Rovine, M. J., & Molenaar, P. C. M. (2001). A structural equations modeling approach to the general linear mixed model. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 67–98). American Psychological Association. <https://doi.org/10.1037/10409-003>
- Sivo, S., Fan, X., & Witta, L. (2005). The biasing effects of unmodeled arma time series processes on latent growth curve model estimates. *Structural Equation Modeling: A Multidisciplinary Journal*, 12(2), 215–231. https://doi.org/10.1207/s15328007sem1202_2
- Smyth, J. M., & Stone, A. A. (2003). Ecological momentary assessment research in behavioral medicine. *Journal of Happiness Studies*, 4(1), 35–52. <https://doi.org/10.1023/A:1023657221954>
- Stan Development Team. (2023a). *RStan: The R interface to Stan*. [Software package].
- Stan Development Team. (2023b). *Stan modeling language users guide and reference manual*, version 2.25.
- Trull, T. J., & Ebner-Priemer, U. (2014). The role of ambulatory assessment in psychological science. *Current Directions in Psychological Science*, 23(6), 466–470. <https://doi.org/10.1177/0963721414550706>
- VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press.
- West, S. G., Aiken, L. S., & Krull, J. L. (1996). Experimental personality designs: Analyzing categorical by continuous variable interactions. *Journal of Personality*, 64(1), 1–48. <https://doi.org/10.1111/j.1467-6494.1996.tb00813.x>

- Williams, D. R., Zimprich, D. R., & Rast, P. (2019). A Bayesian nonlinear mixed-effects location scale model for learning. *Behavior Research Methods*, 51(5), 1968–1986. <https://doi.org/10.3758/s13428-019-01255-9>
- Yaremych, H. E., Preacher, K. J., & Hedeker, D. (2023). Centering categorical predictors in multilevel models: Best practices and interpretation. *Psychological Methods*, 28(3), 613–630. <https://doi.org/10.1037/met0000434>