STRUCTURAL EQUATION MODELING
2026, VOL. 33, NO. 3
ISSN: 1070-5511 | E-ISSN: 1532-8007

Analyzing the Dynamics of Auditory Distraction Using Multigroup Multilevel Structural Equation Modeling

Salome Li G. M. Keintzel[a], Nicole Wetzel[b,c,d], and Florian Scharf[a]

[a]University of Kassel; [b]Leibniz Institute for Neurobiology; [c]Center for Behavioral Brain Sciences Magdeburg; [d]University of Applied Sciences Magdeburg-Stendal

ABSTRACT

Advanced Structural Equation Models (SEMs) have gained popularity in recent years, but they are rarely applied in experimental psychology. We illustrate how analyzing longitudinal phenomena in repeated-measures experiments using SEM can enhance understanding of trial-level dynamics and fundamental cognitive processes. Previous work shows that children exposed to novel sounds during a visual task display increased response times in that trial and the following trial (post-novel effect), suggesting a carryover process. We reanalyzed data from adults and children to directly test this hypothesis, revealing age-related differences in attentional dynamics. We provide a comprehensive account of our analyses and parameter interpretations to help researchers who are interested in similar questions about the temporal dynamics of cognitive effects and complement our findings by a sample size simulation study assessing the model's data requirements.

KEYWORDS

Attention; autoregressive effects; children; experimental psychology; multilevel SEM

Experimental research often produces longitudinal data in the form of repeated measures across multiple trials. Whereas researchers frequently aggregate these data across trials to examine marginal effects, many psychological theories also make more specific predictions at the trial level. Testing such predictions requires longitudinal models that operate on trial-level data. Explicitly modeling these temporal dynamics can reveal new opportunities for theory development and empirical testing (e.g., Baayen et al., 2022; Fink et al., 2024; Von Gunten et al., 2018; Widmann et al., 2024). Experimental psychologists frequently use linear mixed-effects models (LMMs) to analyze longitudinal data. By contrast, (longitudinal) Structural Equation Models (SEMs) are less commonly applied, although they are functionally similar and offer greater flexibility—particularly for investigating lagged, dynamic effects, such as autoregressive effects, where current outcomes depend on their previous values (Curran, 2003; Curran et al., 2014). One key reason for the limited use of SEMs is that most standard SEM software does not support the large number of measurement occasions (e.g., $T > 50$) that are typically employed in behavioral or eye-tracking experiments (Asparouhov & Muthén, 2023). To address this gap, the present study demonstrates how longitudinal SEMs—originally developed for designs with relatively few measurement occasions—can be adapted for experimental settings with many repeated observations by segmenting data into shorter prototypical sequences. We reanalyzed data from a visual categorization task with a longitudinal multigroup multilevel SEM, investigating the effect of different experimental manipulations on autoregressive (carryover) effects in different participant groups. We begin by outlining the theoretical background of our substantive research question and explaining how it translates into statistical terms. We then provide a brief overview of key considerations in the analysis of longitudinal data. The Method section offers a detailed account of model specification and parameter interpretation. Our aim is to support researchers who have similar questions and wish to apply comparable modeling approaches.

1. Theoretical Background

The analysis of correlations between indicators of cognitive processes and their dynamics over time provides new insights into complex cognitive functions (e.g., attention) and their development. Selective attention involves the ability to focus on task-relevant information and ignore task-irrelevant information, an ability that is important for learning and academic performance. For example, when reading an article, the sound of one's mobile phone receiving a text message may involuntarily distract attention and can impair performance on the reading task. Neurophysiological and behavioral studies have suggested that task-irrelevant but unexpected events can involuntarily capture attention and

trigger evaluation processes regarding possible behavioral adjustments (Escera et al., 2000; Horváth et al., 2008). If no behavioral adaptation is required (e.g., no action on the phone), attention is voluntarily reallocated to the task at hand (for a review, see Escera et al., 2000). The mechanisms underlying the detection of unexpected changes in the environment (phone notification sound) and the attentional orienting toward the distractor have been studied intensively, especially in the auditory modality (Escera et al., 2000; Schröger & Wolff, 1998; Wetzel et al., 2006), but the mechanisms underlying the reorienting of attention back to the primary task are not yet well understood. As these attentional mechanisms require resources that are no longer available for the task at hand, performance may be impaired (distraction effect; for a review, see Escera et al., 2000). The distraction effect has been studied extensively through the use of oddball paradigms, which consist of frequently presented standard stimuli and infrequently presented deviant or novel (oddball) stimuli. Participants typically perform a task that is unrelated to the oddball event or the oddball features (e.g., categorizing pictures while a sequence of task-irrelevant standard and oddball sounds is presented). Responses to the target are typically slower following unexpected oddball sounds than following standard sounds. This behavioral distraction effect–typically quantified as the difference in response times (RTs) between standard and oddball trials–is small in adults but has been reliably observed across sensory modalities (Escera et al., 2000; Parmentier, 2014; Wetzel & Schröger, 2014). Consistent with the long-term development of selective attention into adolescence, a number of oddball studies have reported larger distraction effects in children than in adults (Hoyer et al., 2021; Wetzel et al., 2006, 2019). A recent study analyzed the temporal (block-wise) trends of distraction effects using pooled data from different oddball studies with children (aged 4–10 years) and adults (Volkmer et al., 2022). The results showed increased distraction effects in children compared with adults. Over the course of the experimental session, the initially strong distraction effects decreased to adult levels in the 6–10-year-olds, but not in the 4–5-year-olds. The results demonstrate that attention control is a time-varying process and suggest that schoolchildren, but not preschoolers, are capable of attentional control similar to that of adults when given more time to practice and adapt the processes underlying the control mechanisms. These new findings about the time course of distraction control can be used in the development of models of attention and the design of learning environments, and may provide additional information for the diagnosis and treatment of attentional disorders (Volkmer et al., 2022).

In addition, some oddball studies have reported prolonged RTs also in the first post-novel standard trial compared with the remaining, physically identical standard trials in children (Wetzel, 2015; Wetzel et al., 2019) and adults (Ahveninen et al., 2000; Berti, 2008; Kähkönen et al., 2002; Ljungberg & Parmentier, 2012; Parmentier & Andrés, 2010; Roeber et al., 2003). Increased RTs in post-novel standard trials have been accompanied by brain activity that also differs from the activity that occurs during pre-novel standard trials in adults (Roeber et al., 2003, 2005) and children (Wetzel, 2015). These behavioral and neurophysiological differences between pre- and post-novel standards have been discussed as reflecting a continuous reallocation of attention back to task-relevant stimuli after distraction (Roeber et al., 2003). If post-novel effects indeed reflect the continuous reallocation of attention after distraction, then larger distraction effects should be followed by larger post-novel effects, that is, longer RTs in novel trials should carry over to subsequent standard trials, and this carryover effect should specifically occur for novel to post-novel transitions. Furthermore, a decrease in the distraction effect over time as a function of age (Volkmer et al., 2022) should also lead to a similar age-sensitive decrease in post-novel effects. However, it has also been discussed that mechanisms related to task set switching, error-processing, or conflict monitoring contribute to the prolonged RTs in post-novel trials (Berti, 2008; Parmentier & Andrés, 2010; Roeber et al., 2003). The contributions of these processes to the post-novel effect do not necessarily depend on RTs in the preceding novel trial. Examining the trial-level relationship between RTs in novel and post-novel trials across the experiment and age groups may clarify the underlying mechanisms and help distinguish between reorienting and non-reorienting accounts of post-novel effects.

1.1. Present Study

In the present study, we investigated the dynamic relationship between RTs on consecutive trials in an audio-visual oddball paradigm. Specifically, we reanalyzed the data set from Volkmer et al. (2022) to test the hypothesis that post-novel effects result from a *carryover* of distraction-related processes to subsequent trials. That is, we hypothesized that an increased RT on a trial with a novel stimulus would be associated with an increased RT on the trial following this novel stimulus. Furthermore, we examined whether this relationship is specific to the novel–post-novel transition, as compared with other transitions (e.g., standard–novel), and whether these dynamic effects differ between age groups. The specification of a statistical model that was able to test these hypotheses posed two major challenges: The first challenge involved the central hypothesis that the post-novel effect reflects a *carryover* phenomenon. Evaluating this assumption required the specification of a lag-1 autoregressive (AR) effect of RT (moderated by experimental condition and age). In other words, we assumed that the RT on a given trial is not only shaped by the current stimulus but also by the RT on the *preceding* trial. Thus, we needed a statistical model that could operate at the level of individual trials *and* allow us to test lagged (autoregressive) effects. LMMs, though widely used in experimental psychology, are not ideal for this purpose, because there is a well-known bias that affects parameters of observed lagged dependent variables as predictors (Nickell bias; Nickell, 1981) and that cannot be resolved by simple data transformations such as centering (Falkenström, 2024; Hamaker & Grasman, 2014).

The Nickell bias is especially pronounced with fewer time points, but can be mitigated using latent instead of manifest lagged variables (Asparouhov et al., 2018). Thus, we used a longitudinal multilevel SEM framework which permitted the explicit modeling of lagged effects between observed variables at the trial level. The second challenge concerned some other longitudinal phenomena that needed to be taken into account (Von Gunten et al., 2018; Widmann et al., 2024)–such as between-person variability and (individual) trends. These had to be appropriately accounted for because they could otherwise confound the interpretation of autoregressive effects and lead to incorrect inferences about within-person processes. The following section briefly outlines the specific modeling decisions we made in response to these challenges.

1.1.1. Within-Person Versus Between-Person Effects

Most importantly, within-person processes had to be distinguished from between-person variability (Von Gunten et al., 2018). Trial-level autoregressive effects are inherently within-person in nature and failing to account for between-person differences—such as individual's average RT or susceptibility to distraction—can bias these estimates (e.g., Hamaker et al., 2015). This separation is typically implemented via random intercepts and random slopes, which account for individual baselines and heterogeneous responses to experimental conditions between individuals (Singmann & Kellen, 2019). This separation was achieved by including latent person-level intercept and slope factors and by allowing model parameters to vary across individuals where necessary.

1.1.2. Temporal Trends

Temporal trends, that is, systematic changes in an outcome over time, are commonly observed in experimental data, reflecting factors such as fatigue, learning, or changing task engagement (Baayen et al., 2017; Lorist et al., 2000; Widmann et al., 2024). Such trends were also reported for the data set we reanalyzed here (Volkmer et al., 2022). From a time-series perspective, trends represent violations of the so-called stationarity assumption, which underlies the interpretation of autoregressive parameters (for a comprehensive overview, see Ryan et al., 2025). If left unmodeled, trends can inflate or obscure autoregressive effects, leading to misinterpretations (Mund et al., 2021; Woodward et al., 2017). Thus, we included slopes for a time variable to account for trends.

1.1.3. Moderating Factors Between and Within Participants

Finally, we assumed that carryover effects and trends may vary depending on experimental condition (e.g., whether a trial follows a novel stimulus) and between age groups, reflecting differences in distraction susceptibility or cognitive control. These moderator effects were explicitly modeled via condition-specific trends and autoregressive effects in a mul-tigroup SEM, allowing us to assess interactions between longitudinal processes and age. Failing to model such interactions can again violate the stationarity assumption because the time-varying autocorrelation structure is not properly addressed. Similarly, hidden interactions with trends can lead to erroneous conclusions if ignored (Widmann et al., 2024).

Together, these considerations necessitated a modeling approach that is both flexible and sensitive to the structure of experimental longitudinal data. The SEM framework allowed us to jointly model autoregressive (carryover) effects, trends, and moderator effects within and between individuals in a unified structure, while maintaining a clear separation of within- and between-person processes.

2. Method

2.1. Original Experiments and Data

The data were pooled from previous auditory-visual oddball studies by Wetzel et al. (2019, 2016, 2021, available upon request from the authors) and were previously reanalyzed by Volkmer et al. (2022). In line with the detailed poolability analysis of Volkmer et al. (2022), we pooled data from the three experimental blocks in which all three studies had an almost identical experimental setup, that is, blocks featuring non-repeating environmental sounds as novel sounds (for details, see the additional online material or Volkmer et al., 2022). The number of trials per condition, sound duration and loudness, task, and visual targets were therefore identical across studies.

2.1.1. Participants

We reanalyzed data from children in three age groups (ages 4-5, 6-7 and 8-10) and one adult group for a total of 352 participants. Age groups were formed with the aim of having at least 50 participants each. All studies were performed in quiet rooms at participants' kindergartens, after-school care or in the laboratory. Adult participants provided written informed consent and confirmed that they did not suffer from hearing disorders, neurological diseases, or attention deficit hyperactivity disorder and that they had not taken any medication that would affect the central nervous system. Parents confirmed the same conditions for their children and provided parental consent. Children provided oral consent. The studies were approved by the local ethics committees. We applied the same exclusion criteria as in the original studies. Therefore, 14 children were excluded because they did not complete all experimental blocks, and an additional seven children had to be excluded due to disturbances, technical problems or misunderstanding the task. In addition, two adults and nine children were excluded because their average RTs were more than 2 *SDs* away from the mean of their respective age group. Additionally, as we investigated specific sequences of trials (see the Analysis section), we also excluded participants who did not have such complete sequences in their data (e.g., due to errors), resulting in another six participants aged 4-5 years who had to be excluded. The final sample consisted of data from $n = 48$ children aged 4-5, $n = 86$ children aged 6-7, $n = 128$ children aged 8-10 and $n = 51$ adults (total $N = 313$).

2.1.2. Visual Stimuli

One of three sets of background scenes and targets was presented in each experimental block. The first set featured a palace and a fortress, with princesses and knights as targets. The second set depicted a basket and a hen-roost in a village, with cats and hens as targets. The third set included a flowering shrub and a pond in a coastal landscape, with butterflies and fish as targets. Each target appeared in two variants, which differed in orientation (facing left or right), shape, and color. The variants were presented with equal probability (25% each).

2.1.3. Sounds

The sounds were environmental sounds (e.g., sneezing, a bird song, a bell) which were rated as identifiable (Wetzel et al., 2021). For frequently presented standard sounds, Wetzel et al. (2021) used different sounds across participants whereas the other studies used one sound for all participants (fragment of a bell). The sounds were presented at an average loudness of 55 dB SPL, and the sound duration was 500 ms including raised cosine fade-in and fade-out windows of 10 ms. In Wetzel et al. (2016) and Wetzel et al. (2021) sound intensity was equalized to 6.5 sone (Zwicker et al., 1991), whereas in Wetzel et al. (2019), it was RMS equalized.

2.1.4. Task and Procedure

All studies used the same oddball-paradigm. The task was introduced as a story in which participants were instructed to help the targets reach their destinations which were shown on the left or right side of the background by pressing the left or right button (e.g., clicking right to guide the fish to the ocean shown on the right side of the background). Participants were asked to respond to the visual target stimuli as quickly and correctly as possible. Figure 1 presents the prototypical trial structure. Participants responded with their respective index fingers, except for one child in Wetzel et al. (2019), who responded with their thumbs. Correct responses within the 2,000 ms RT window were followed by a 600 ms feedback animation of the target moving toward the destination. Data consisted of three consecutive experimental blocks with 36 trials each (total of 108 trials) per participant. The block order was counterbalanced across participants. Every block was preceded by a training block with no limit for RT, which was repeated if the participant did not respond correctly in at least 50% of the training trials. One experimental block took approximately 2 min to complete. Within each block, every target was preceded by a task-irrelevant sound, which was either a standard sound ($n = 28$ per block, $n = 84$ per condition) or an interspersed novel sound ($n = 8$ per block, $n = 24$ per condition). The order was pseudorandomized per participant, so that novel trials were always preceded by at least two standard sound trials and also followed by two standard trials unless participants reached the end of the block. All four target versions per block were presented with equal number of novel sounds (balanced). The task was performed on a Macbook (13.3") with external loudspeakers placed on the left and right sides of the laptop, using Matlab (The MathWorks, Natick, U.S.A.) and the Psychtoolbox (Kleiner et al., 2007). An RTBox (Lee et al., 2012) was used to measure the RTs. Two response buttons for children were placed on the left and right sides of the laptop.

2.2. Data Analysis

Data were preprocessed with R version 4.4.2 (R Core Team, 2025) and the data were analyzed with Mplus version 8.11 (Muthén & Muthén, 1998) via MplusAutomation version

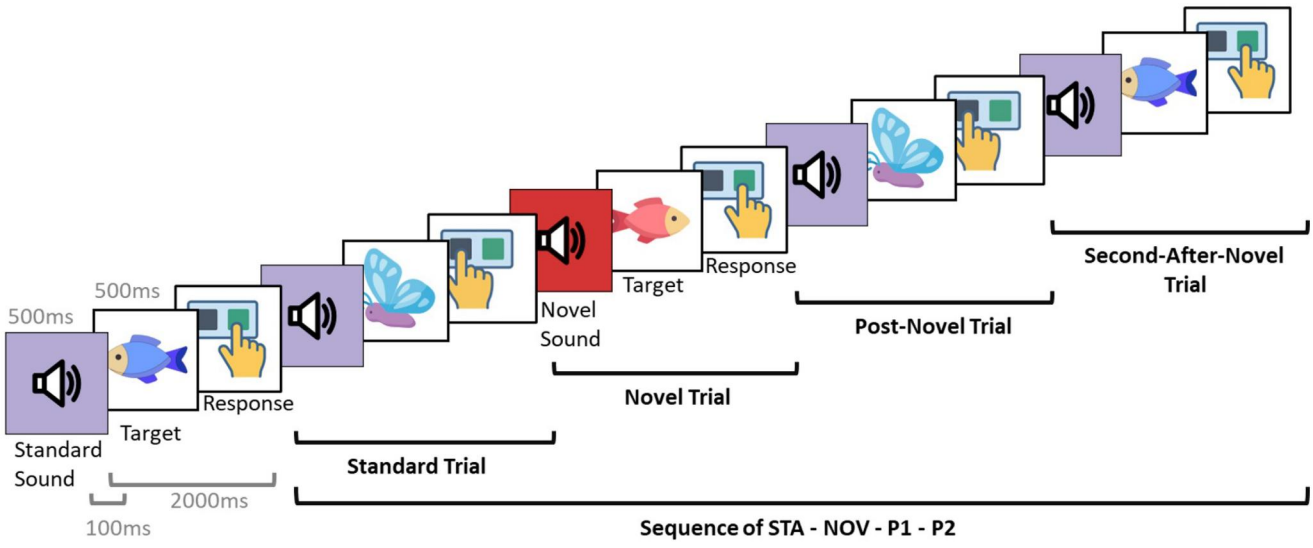

Figure 1. Trial and sequence structure.
Note: Participants were instructed to respond to targets (here: fish and butterfly) by pressing the left or right button within a 2000 ms response window. Every target was preceded by a standard (78% of trials) or novel (22% of trials) task-irrelevant environmental sound (oddball paradigm). The data for analysis only contained complete sequences of "standard - novel - post-novel - second-after-novel" as highlighted in the Figure. The target images are illustrative examples and not the actual stimuli used in the studies

1.1.1 (Hallquist & Wiley, 2018). All analysis scripts and annotated Mplus files are available in the additional online material.

2.2.1. Exclusion Criteria for Trials

The first two trials per block were excluded because they were needed to establish the rule for distinguishing between standard and novel sounds (Bendixen et al., 2007). Trials with no response within 2,000 ms after target onset or responses faster that 100 ms after target onset were considered invalid and thus excluded from the analysis. Furthermore, trials with an incorrect response (error) and trials following incorrect or invalid responses (post-error trials) were excluded from the analysis to avoid confounding by error-related processes (Dutilh et al., 2012). After these preprocessing steps, we extracted sequences of interest to investigate RT dynamics across distinct trial types. Specifically, we focused on four trial types: standard trials (STA), novel trials (NOV), the first standard trial following a novel trial (post-novel, P1), and the second standard trial after a novel trial (P2; see Figure 1). Extracting fixed STA–NOV–P1–P2 sequences allowed us to examine carryover effects for relevant transitions between trial types and to distinguish specific novel-related carryover from general carryover effects between any trials. The number of sequences extracted per participant was as follows: for children aged 4–5, $m = 5$, range $[1; 11]$; children aged 6–7, $m = 8$, range $[1; 15]$; children aged 8–10, $m = 9$, range $[2; 16]$; and adults, $m = 12$, range $[4; 17]$.

2.2.2. Model Building

In statistical modeling terms, the hypothesized carryover effects from novel RTs to post-novel RTs can be conceptualized as an autoregressive effect, which captures the influence of the previous observation on the current observation. In a simplified scenario, where the carryover effect operates in isolation, it could be estimated using a simple regression model with the novel RT as a predictor of the post-novel RT. However, in realistic longitudinal data, multiple processes must be considered; otherwise, the autoregressive parameter may be confounded and will then likely fail to represent the trial-level carryover effect of interest. In the following, we explain the full model step by step (as shown in Figure 2), illustrating how it can be understood as an extension of the simple regression framework by conceptually successively adding components that account for specific potential confounding processes. The model was parameterized to resemble conventional LMMs commonly used in experimental psychology, ensuring that the model coefficients provide a clear substantive interpretation of the relevant longitudinal phenomena. All corresponding model equations for the full model can be found in the additional online material.

As mentioned above, the central parameter of interest is a regression coefficient α that describes the linear relationship between post-novel RTs (**P1**) and the preceding novel RTs (**NOV**; Figure 2a):

$$\mathbf{P1} = \alpha \cdot \mathbf{NOV} + \mathbf{e}_{P1}$$

As a first step, it was important for our hypothesis to account for the possibility that the carryover effect was not

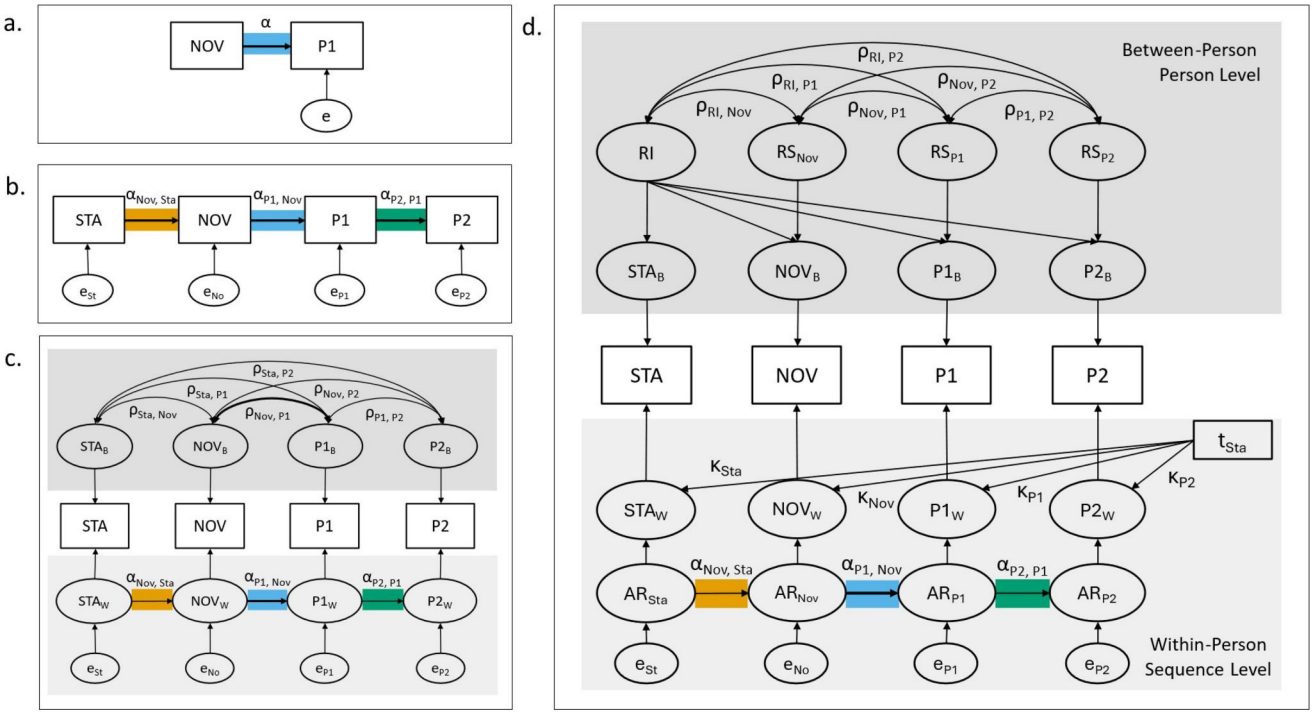

Figure 2. Building blocks of the analysis model.
Note: The panels illustrate the building blocks leading to the full model in Panel d. Panel a shows the autoregressive effect simplified to a simple regression. Panel b shows the autoregressive effect of all trial types in a sequence. Panel c shows the distinction between between-person (dark grey, Index *B*) and within-person (light grey, Index *W*) components of the model. Panel d shows the final multilevel structural equation model for one age group including trends. All paths without Greek letters are fixed to 1. STA = standard RT, NOV = novel RT, P1 = post-novel RT, P2 = second after novel RT, $\mathbf{t}_{Sta}$ = time variable coding trial number of STA, RI = random intercept, RS = random slope, e = residual variable

specific to novel–post-novel transitions. All RTs, regardless of trial type, might exhibit an autoregressive effect, potentially with different magnitudes. For the substantive hypothesis to be plausible, the novel–post-novel transition should differ substantially from other transitions (i.e., the autoregressive effect should be larger than in other transitions). Thus, we investigated longer STA-NOV-P1-P2 sequences and also included autoregressive paths for the surrounding standard trials (Figure 2b), allowing us to compare parameter magnitudes and test whether the carryover from novel to post-novel RTs ($\alpha_{P1,Nov}$) was larger than that of other trial types—that is, whether there was an interaction between trial type and autoregressive effect. The residuals, denoted by **e**, were specified as latent variables whose variances were allowed to differ between trial types to account for possible heteroscedasticity.

The model formula for **P1** corresponding to Figure 2b is as follows:

$$\begin{aligned}\mathbf{P1} &= \alpha_{P1,Nov} \cdot \mathbf{NOV} + \mathbf{e}_{P1}\\ \mathbf{P1} &= \alpha_{P1,Nov} \cdot \underbrace{(\alpha_{Nov,Sta} \cdot \underbrace{\mathbf{e}_{Sta}}_{\text{Substitute for STA}} + \mathbf{e}_{Nov})}_{\text{Substitute for NOV}} + \mathbf{e}_{P1}\end{aligned}$$

The second building block of our model accounted for the multilevel structure of sequences nested within participants. To distinguish person-level and trial-level processes, the multilevel SEM decomposes each observation into a between-participant (person-specific) and a within-participant (sequence-specific) component. Conceptually, the multilevel SEM can be understood as a combination of two models–one representing between-person differences in RTs and one representing within-person variation in RTs (see Figure 2c):

$$\mathbf{P1} = \underbrace{\mathbf{P1}_B}_{\text{between part}} + \underbrace{\mathbf{P1}_W}_{\text{within part}}$$

The between-person model captures differences between participants in their *average* RTs for each trial type. Thus, the between-person components $\mathbf{STA}_B$, $\mathbf{NOV}_B$, $\mathbf{P1}_B$, and $\mathbf{P2}_B$ take on a constant value for each participant, such that $\mathbf{NOV}_B$ reflects a participant's average RT in novel trials and $\mathbf{P1}_B$ reflects their average RT in post-novel trials. The means of these components correspond to the grand-average RTs for each trial type, their variances represent the variability in average RTs across participants, and the correlations $\rho_{..}$ quantify the association between participants' average RTs in the respective trial types.

The within-person model captures the deviation of the RT in a specific trial from the (latent) average[1] RTs of the participant in the respective trial type (latent centering within participants). That is, the within-person components $\mathbf{STA}_W$, $\mathbf{NOV}_W$, $\mathbf{P1}_W$, and $\mathbf{P2}_W$ take on different values for each trial. For instance, $\mathbf{NOV}_W$ quantifies how much faster or slower a participant responded in a specific novel trial compared with their average RT in novel trials. The autoregressive effects α quantify the extent to which the within-person part of a specific trial is carried over to the subsequent trial:

$$\mathbf{P1}_W = \alpha_{P1,Nov} \cdot \underbrace{(\alpha_{Nov,Sta} \cdot \underbrace{\mathbf{e}_{Sta}}_{\text{Substitute for } \mathbf{STA}_W} + \mathbf{e}_{Nov})}_{\text{Substitute for } \mathbf{NOV}_W} + \mathbf{e}_{P1}$$

For example, $\alpha_{P1,Nov}$ estimates to what extent a participant responds *particularly* slowly (relative to their own average) in a specific post-novel trial, if they also responded *particularly* slowly in the preceding novel trial. This effect is fundamentally different from the correlation of the respective between components, as this between-person correlation quantifies whether participants who respond more slowly to novel trials on average also respond more slowly to post-novel trials on average. The between-level correlation thus refers to stable differences between participants, whereas the within-level carryover captures a trial-specific process. To accurately model the carryover effect of interest, it is essential to separate the between- and within-person components (Hamaker et al., 2015).

In our full model, we used an alternative parameterization at the between level to ensure that its interpretation more closely resembled the random effects typically specified in an LMM (see Figure 2d, upper part). Specifically, we introduced a common random intercept (RI) factor for all trial types, representing each participant's average standard trial RT, which is added to the components of every trial type. This factor captures the idea that between-person differences may result in generally faster or slower responses across all trial types. The first standard trial in each sequence serves as the reference. The RI factor mean corresponds to the fixed intercept in an LMM, and the RI factor variance corresponds to the random intercept variance. The random slope (RS) factors quantify the *differences* between the average standard trial RT (reflected by the RI factor) and the average RT in the respective trial type for each participant (i.e., their individual person-level experimental effects of novel, post-novel, or second-after-novel trials on RT). Thus, $\mathbf{RS}_{Nov}$ serves as a marker of participants' distraction susceptibility (higher values indicate greater susceptibility), and $\mathbf{RS}_{P1}$ serves as a marker of various stable person-level characteristics contributing to the difference between standard trial and post-novel RT (individual person-level post-novel effect). Many processes may come into play here, such as susceptibility to distraction or the ability to refocus on the task. The correlation $\rho_{Nov,P1}$ quantifies the association between participants' distraction susceptibility and their individual person-level post-novel effect (participants' characteristics). Expressed as a formula, this parameterization would be:

$$\begin{aligned}\mathbf{P1} &= \underbrace{\mathbf{P1}_B}_{\text{between part}} + \underbrace{\mathbf{P1}_W}_{\text{within part}}\\ \mathbf{P1}_B &= \mathbf{RI} + \mathbf{RS}_{P1}\end{aligned}$$

The RS factor means quantify the average experimental effects of the respective trial types across individuals, corresponding to the fixed effects of the trial types in an LMM.

[1]For brevity, we use the therm average to denote latent means in the remainder of the paper.

The RS factor variances correspond to the random effects variances.

We also further extended the within-person part of our full model to reduce confounding of the autoregressive effects $\alpha_{..}$ with temporal trends (i.e., systematic changes in RT over time; see, e.g., Woodward et al., 2017, Chapter 5). Previous analyses of this data set, as well as data from similar experimental designs, have identified such trends in standard and novel trial RTs (Volkmer et al., 2022). Specifically, we added linear trends for all trial types by regressing the within-person components on a time variable $\mathbf{t}_{Sta}$ (see Figure 2d, lower part). To ensure an intuitive interpretation of the trend parameters κ, we included an additional layer of autoregressive components (labeled **AR**) in the within-person part:

$$\mathbf{P1}_W = \underbrace{\kappa_{P1} \cdot \mathbf{t}_{Sta}}_{\text{trend part}} + \underbrace{\mathbf{AR}_{P1}}_{\text{AR part}}$$

$$\begin{aligned}
\mathbf{AR}_{P1} &= \underbrace{\alpha_{P1,No} \cdot \mathbf{AR}_{Nov}}_{\text{carryover}} + \underbrace{\mathbf{e}_{P1}}_{\text{residual}} \\
&= \alpha_{P1,No} \cdot \underbrace{(\alpha_{No,St} \cdot \mathbf{AR}_{Sta} + \mathbf{e}_{No})}_{\text{substitute for } \mathbf{AR}_{Nov}} + \mathbf{e}_{P1} \\
&= \alpha_{P1,No} \cdot (\alpha_{No,St} \underbrace{\cdot \mathbf{e}_{St}}_{\text{substitute}} + \mathbf{e}_{No}) + \mathbf{e}_{P1} \\
&= \underbrace{\alpha_{P1,No} \cdot \alpha_{No,St} \cdot \mathbf{e}_{St}}_{\text{carryover from STA residual}} + \underbrace{\alpha_{P1,No} \cdot \mathbf{e}_{No}}_{\text{carryover from NOV residual}} + \mathbf{e}_{P1}
\end{aligned}$$

This modeling choice separates temporal trends from trial-to-trial carryover effects, so that κ directly represents the estimated change in RT for each trial type with a 1-unit increase in the time variable, independent of the autoregressive effect (Jongerling & Hamaker, 2011; Xiao & Liu, 2025). The **AR** components quantify RTs after controlling for both the between-person components and the temporal trend. More specifically, they capture the deviation of the observed RT from the expected RT, given each participant's average RT and the time on task (i.e., these are the residuals after accounting for both the between-person model and the trend analysis). The interpretation of the autoregressive parameters $\alpha_{..}$ in the full model is as follows: In trials where participants are more distracted than expected on the basis of their average RT, distraction susceptibility, and time elapsed during the experiment, do they also show greater distraction in the post-novel trial than would be expected on the basis of those same factors (i.e., average RT, individual post-novel effect, and time on task)? This type of carryover effect is sometimes referred to as a *dynamic residual* effect (Asparouhov & Muthén, 2023). This principle is illustrated in Figure 3a. The red dots depict RTs from an exemplary sequence from a specific participant. The purple squares represent the RTs that would be predicted on the basis of the between-person components, that is, the average RT per trial type for that participant. The residuals are illustrated by black arrows; the carryover effect α of the dynamic residuals quantifies the extent to which the length and direction of an arrow in one trial are associated with the length and direction of the arrow in the preceding trial.

The introduction of the time variable also affects the interpretation of the between-person components, as all respective effects become *conditioned* on the time point zero. As a result, the between-person factor means (i.e., the fixed intercepts and fixed slopes for the trial types in an LMM) refer to this specific reference point. To ensure meaningful interpretation of these conditional effects, we coded the time variable $\mathbf{t}_{Sta}$ such that the value 0 corresponds to the midpoint of the experiment (Trial 54) as a reference point. Note that the temporal trends were modeled using a single time variable, $\mathbf{t}_{Sta}$, which codes the trial number of the first standard trial in each sequence (as indicated by "Sta" in the variable name). This means that $\mathbf{t}_{Sta}$ is constant within each sequence and does not increase per trial within a sequence. This coding was chosen to ease the interpretation of the model parameters. With $\mathbf{t}_{Sta}$ defined in this way, the mean of the RI factor can be interpreted as the conditional standard trial RT across participants when $\mathbf{t}_{Sta}$ is equal to zero, that is, in the middle of the experiment (Trial 54; analogous to the fixed intercept in an LMM). Similarly,

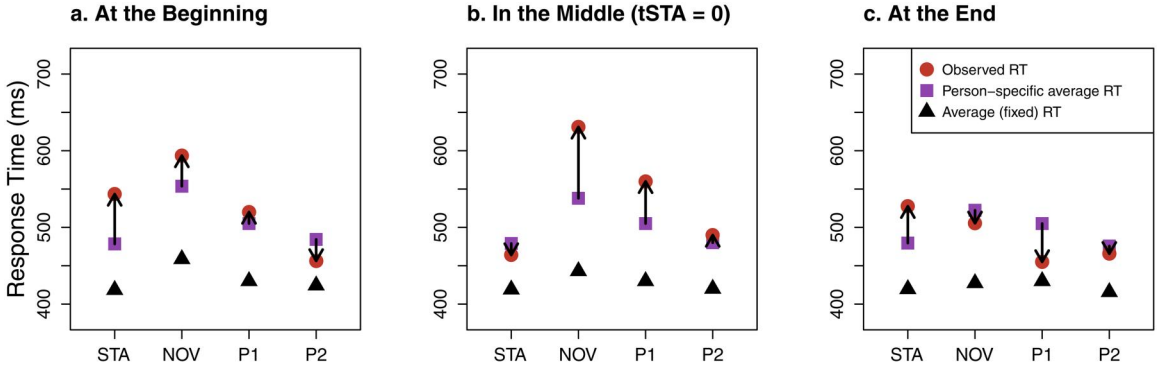

Figure 3. Visualization of fixed effects, random effects and dynamic residuals with data from exemplary sequences from a hypothetical participant.
Note: Black triangles depict the estimated fixed RT across participants for a specific time point at the beginning, in the middle (when $\mathbf{t}_{Sta}$ is equal to 0) or at end of the experiment. Depending on trial type and time point, this value consists of the fixed intercept (for standard trials STA in the middle of the experiment), plus the fixed effect of the respective trial type (for novel trials NOV, post-novel trials P1, second-after-novel trials P2, in the middle of the experiment), plus the fixed trend for the trial type (at all other time points). Person-specific random effects (random intercept plus random slope) are deviations between black triangles and purple squares. The depicted participant is generally slower than average and shows a larger distraction effect (larger difference between STA and NOV for purple squares compared with black triangles). Deviations between red dots (observed RT in a specific trial) and purple squares (person-specific average RT per trial type) are depicted by arrows (residuals). The carryover effects α in the model quantify how much the residual (arrow) is predicted by the previous trial's residual

the RS factor means represent the conditional difference between the standard trial RT and the RT of the respective trial type across participants when $\mathbf{t}_{Sta}$ equals zero (i.e., in the sequence beginning with trial 54). Thus, the RS factor means reflect the *conditional* effect of the respective trial type in the middle of the experiment, analogous to the fixed slope of the trial type in an LMM with an accordingly centered time variable. To illustrate the conditional effects, we show example sequences at different time points in Figure 3. The values of the black triangles vary between panels due to the trend. The *conditional* effects of our full model correspond to differences between the STA black triangle and the triangles from the other types of trials in Figure 3b.

The full model shown in Figure 2d was estimated as a multigroup model. In this approach, the same SEM is fit simultaneously to multiple (age) groups within a single model framework. This approach allows for the testing or constraining of parameters to equality across groups. Whereas multigroup SEM is commonly used to assess measurement invariance in psychological instruments (Meredith, 1993), the framework is applicable to any SEM context. In our case, we used the multigroup SEM to obtain parameter estimates for all parameters in Figure 2d for each age group within one model. This approach enabled us to conveniently calculate and test for differences in model parameters between age groups (e.g., to test the equality of $\alpha_{P1,Nov}$ across groups).

Readers familiar with recent advancements in intensive longitudinal data analysis might question why we did not employ a residual dynamic SEM (RDSEM, Asparouhov & Muthén, 2020). While the RDSEM framework theoretically aligns with our data structure, Mplus lacked a convenient option for modeling all the latent interactions at the within level at the time of our data analysis. Furthermore, we were specifically interested in these STA-DEV-P1-P2 trial sequences and were reluctant to treat all remaining standard trials outside these sequences as a single trial type, as we were uncertain whether the autoregressive effect might change when multiple standards are presented consecutively due to the repetition of auditory stimuli (Bendixen et al., 2007).

3. Results

The model was fit in Mplus 8.11 (Muthén & Muthén, 1998) using maximum likelihood estimation with heteroscedasticity-robust standard errors (MLR).

3.1. Convergence

Due to the complexity of our model, we encountered convergence issues, that is, the numerical algorithm failed to (reliably) locate the maximum likelihood, resulting in no output or a warning about unreliable results. Such issues are not unusual when implementing complex SEMs and need to be addressed when implementing such models (see e.g., Lüdtke et al., 2011). It is beyond the scope of this paper to discuss the technical details and reasons for convergence issues, but we disclose the steps we followed to attempt to solve the convergence issues in the additional online material, as they are needed to replicate our results and may help researchers handle convergence issues in their models. Eventually, the full model converged, but the algorithm reached a saddle point as indicated by a warning message. However, as standard errors were returned and withstood our evaluation (see the additional online material), we interpreted the model as advised by the authors of the Mplus package (Muthén, 2020). To increase confidence in the robustness of the results with respect to the estimation method, we also used a Bayesian estimator with default default priors coming with Mplus 8.11, that is, normal distribution for intercepts ($N(0, \infty)$) and regression slopes and inverse Wishart distributions for covariance matrices with $\sigma^2 \sim \text{IW}(0, -p-1)$, where p is the size of the covariance matrix.[2] Overall, the Bayesian estimation yielded the same result pattern for the parameter estimates (see the additional online material).

3.2. Model Results

Table 1 presents all estimated coefficients, along with their standard errors (SEs) and concise interpretation guidelines. All post-tests comparing coefficients within or between age groups were corrected for a family-wise false discovery rate (FDR) in accordance with Benjamini & Hochberg (1995).

3.2.1. Conditional Effects of Trial Types

The fixed effects of the trial types represent the conditional differences between standard trial RTs and the respective trial type RTs across participants in the middle of the experiment (Trials 54 to 57), as modeling the temporal trends requires a reference point. For the sake of readability, we refer to these simply as *conditional effects*. All age groups showed a significant conditional distraction effect in novel trials (m_{NOV}; RT difference between standard and novel trials in trials 54 and 55), that is, participants' responses to novel trials were slower than their responses to previous standard trials in the middle of the experiment. The conditional distraction effects declined with age: Children aged 4-5 showed significantly larger conditional distraction effects than all other age groups ($p_{4-5\ vs\ 6-7} = .047$, $p_{4-5\ vs\ 8-10} < .001$, $p_{4-5\ vs\ adults} < .001$), followed by children aged 6-7 who showed significantly larger conditional distraction effects than adults ($p = .001$) and children aged 8-10 ($p = .008$). Children aged 8-10 showed no significant difference in distraction effects from adults ($p = .086$). All age groups except for children aged 8-10 years also showed a significant conditional post-novel effect (m_{P1}; RT difference between standard and post-novel trials in Trials 54 and 56), that is, participants' responses to post-novel trials were slower than their responses to previous standard trials in the middle of the experiment. Although post-novel effects declined with age on a descriptive level, no pairwise comparisons between age groups were significant.

[2]The multigroup model is not (yet) available with Bayesian estimation in Mplus 8.11 (Muthén & Muthén, 1998).

Table 1. Results for parameter estimates.

	4-5			6-7			8-10			Adults			
Parameter	Est.	*SE*	*p*	Est.	*SE*	*p*	Est.	*SE*	*p*	Est.	*SE*	*p*	Brief Interpretation
$\alpha_{NOV,STA}$	−0.01	0.06	0.912	**0.13**	0.05	0.007	**0.13**	0.06	0.017	**0.17**	0.05	0.001	Carryover from standard to novel trials
$\alpha_{P1,NOV}$	−0.02	0.07	0.736	**0.16**	0.05	0.001	**0.11**	0.03	0.001	**0.18**	0.05	< 0.001	Carryover from novel to post-novel trials
$\alpha_{P2,P1}$	**0.20**	0.07	0.004	**0.19**	0.06	0.002	**0.11**	0.04	0.003	**0.13**	0.05	0.014	Carryover from post-novel to second-after-novel trials
κ_{STA}	0.62	0.57	0.277	**0.44**	0.17	0.009	**0.21**	0.10	0.033	−0.01	0.13	0.944	Trend in standard trial RT
κ_{NOV}	**−1.08**	0.41	0.009	**−0.90**	0.24	< 0.001	**−0.56**	0.13	< 0.001	**−0.29**	0.11	0.009	Trend in novel trial RT
κ_{P1}	−0.56	0.56	0.319	**0.66**	0.21	0.002	0.21	0.12	0.085	−0.00	0.12	0.982	Trend in post-novel trial RT
κ_{P2}	−0.80	0.54	0.140	**0.48**	0.19	0.010	**0.37**	0.11	0.001	−0.08	0.10	0.401	Trend in second-after-novel trial RT
m_{RI}	**734.28**	19.82	< 0.001	**593.52**	9.92	< 0.001	**536.68**	6.45	< 0.001	**419.69**	6.55	< 0.001	Conditional standard trial RT in middle of experiment
m_{NOV}	**122.87**	22.02	< 0.001	**71.99**	11.04	< 0.001	**37.05**	5.95	< 0.001	**24.25**	4.50	< 0.001	Conditional distraction effect in middle of experiment
m_{P1}	**49.98**	22.55	0.027	**14.35**	6.86	0.037	10.02	5.32	0.060	**11.59**	3.90	0.003	Conditional post-novel effect in middle of experiment
m_{P2}	−3.78	20.77	0.856	5.17	7.72	0.503	−8.04	4.18	0.054	1.78	4.47	0.691	Conditional second-after-novel effect in middle of experiment
sd_{RI}	**97.01**	21.38	0.023	**84.49**	12.74	0.001	**64.08**	7.52	< 0.001	**42.00**	9.30	0.024	*SD* of standard trial (baseline) RT between individuals
sd_{NOV}	44.38	69.32	0.749	68.50	22.15	0.122	**44.08**	7.60	0.004	16.41	8.83	0.353	*SD* of distraction effect (distraction susceptibility) between individuals
sd_{P1}	41.88	55.87	0.708	11.30	30.42	0.853	25.95	10.89	0.233	4.93	29.67	0.934	*SD* of post-novel effect between individuals
sd_{P2}	36.21	28.67	0.528	26.09	14.01	0.352	15.50	13.38	0.562	3.12	45.77	0.973	*SD* of second-after-novel effect between individuals
$\rho_{RI,NOV}$	0.57	1.83	0.754	−0.10	0.14	0.512	−0.36	0.19	0.051	0.21	0.51	0.679	Correlation between individual's standard trial RT and distraction susceptibility
$\rho_{RI,P1}$	0.31	1.26	0.807	−0.03	0.85	0.974	0.07	0.34	0.844	0.88	5.81	0.880	Correlation between individual's standard trial RT and post-novel effect
$\rho_{RI,P2}$	0.85	1.18	0.472	0.16	0.42	0.705	−0.40	0.41	0.333	0.70	12.49	0.955	Correlation between individual's standard trial RT and second-after-novel effect
$\rho_{NOV,P1}$	0.47	1.70	0.784	0.92	2.40	0.701	**0.71**	0.34	0.037	0.53	1.99	0.789	Correlation between individual's distraction susceptibility and post-novel effect
$\rho_{NOV,P2}$	0.71	2.30	0.757	0.19	0.55	0.726	0.99	0.60	0.099	0.39	3.87	0.919	Correlation between individual's distraction susceptibility and second-after-novel effect
$\rho_{P1,P2}$	0.74	3.82	0.846	0.03	1.80	0.986	0.66	0.47	0.166	0.72	7.40	0.922	Correlation between individual's post-novel effect and second-after-novel effect
σ_{STA}	236.54	19.29	< 0.001	118.97	4.88	< 0.001	114.02	6.89	< 0.001	72.48	4.84	< 0.001	Residual *SD* for standard trials
σ_{NOV}	220.46	21.85	< 0.001	159.93	13.52	< 0.001	123.18	8.25	< 0.001	70.32	4.63	< 0.001	Residual *SD* for novel trials
σ_{P1}	224.62	19.76	< 0.001	136.34	8.43	< 0.001	132.15	9.85	< 0.001	80.48	6.96	< 0.001	Residual *SD* for post-novel trials
σ_{P2}	192.37	23.44	< 0.001	146.87	9.33	< 0.001	108.41	4.53	< 0.001	90.19	16.51	0.006	Residual *SD* for second-after-novel trials

Notes: Estimates are reported separately for each age group (adults: $n = 51$; children aged 8–10: $n = 128$; children aged 6–7: $n = 86$; children aged 4–5: $n = 48$). SE = standard error; *SD* = standard deviation. The dependent variable, response time (RT), was measured in milliseconds (ms). The middle of the experiment corresponds to Trials 54 to 57. Values significantly different from zero are printed in bold ($\alpha = .05$), except for residual variances, all of which were significantly different from zero. Standard errors for standard deviations were calculated using the delta method (see the additional online material).

We did not find any conditional post-post-novel effects (m_{P2}; RT difference between standard and second-after-novel trials in Trials 54 and 57).

3.2.2. Carryover Effects α

A visualization of results for autoregressive effects α, reflecting the carryover in RT from one trial to the following trial, is provided in Figure 4. Children aged 6-10 and adults all showed significant carryover effects (α) for all trial types. Carryover effects did not differ significantly between trial types or between age groups of children aged 6-7 years, children aged 8-10 years, and adults. However, children aged 4-5 showed a significant carryover effect only from post-novel to second-after-novel RT $\alpha_{P2,\,P1}$, but not from standard to novel $\alpha_{Nov,\,Sta}$ or novel to post-novel $\alpha_{P1,\,Nov}$. The carryover from post-novel to second-after-novel RTs $\alpha_{P2,\,P1}$ was significantly larger than the previous carryover effects ($p = .045$ and .045, respectively; Figure 4). Between age groups, the post-tests revealed no significant differences after the FDR correction ($p \geq .093$).

3.2.3. Trend Parameters

All age groups showed a significant decline in novel trial RTs over time (κ_{NOV} trend parameter). Children aged 6 - 10 also showed significant increases in RT in other trial types (STA, P1, P2) over time. Nevertheless, we only accounted for fixed, linear trends per trial type, which is likely an oversimplification as the underlying trend in novel trial RTs has been suggested to follow a non-linear form (Volkmer et al., 2022). Additionally, trends might differ between individuals, so that some participants may show a steeper decline or increase in RTs than others (random trends). As mentioned before, when the trends are incorrectly or insufficiently controlled (incorrect form or neglect of random trends) the bias estimates of autoregressive (carryover) effects may be biased (Ryan et al., 2025; Woodward et al., 2017). To address this issue, we estimated a refined model in which we incorporated a second-order polynomial trend for novel trials as well as random trend parameters for all trial types to examine the robustness of the carryover effect results under different trend specifications. We used a Bayesian single-group model because polynomial random trends are not supported by the use of maximum likelihood estimation in Mplus 8.11. In the same vein we also specified random effects for the residual variances, relaxing the assumption that all participants within an age group share the same residual variances. This approach allowed residual variances to vary not only between the trial types and age groups but also between participants within those age groups, accounting for another form of heteroscedasticity. This analysis produced the same result pattern for the autoregressive effects, with overall slightly smaller estimates for the carryover effects compared with the robust maximum likelihood estimation results reported above (see the additional online material), thereby strengthening confidence in our findings.[3]

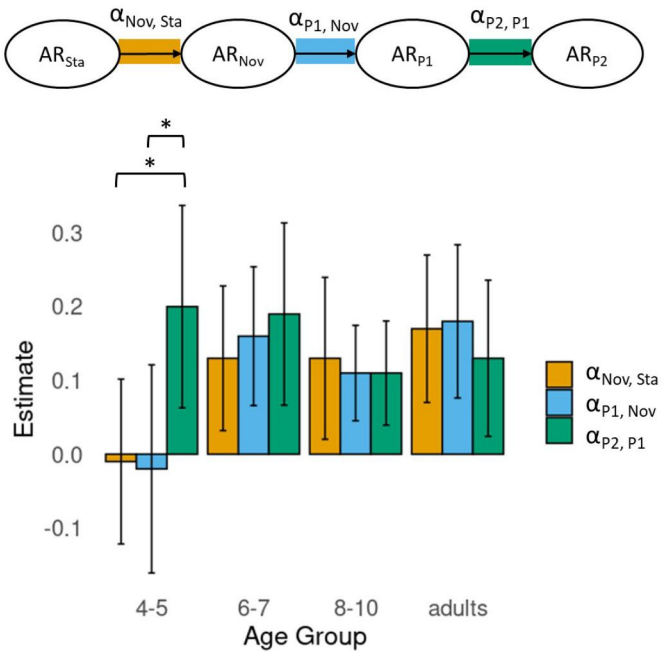

Figure 4. Visualization of carryover effect results.
Note: Results for the autoregressive part of the SEM. The upper panel displays the autoregressive component (AR) from Figure 2d (carryover, autoregressive part) to facilitate interpretation. The lower panel presents a bar plot of carryover effect estimates for the autoregressive transitions between trial types: standard (STA) to novel (NOV), novel to post-novel (P1), and post-novel to second-after-novel (P2), reported separately for each age group (4–5 years: $n = 48$; 6–7 years: $n = 86$; 8–10 years: $n = 128$; adults: $n = 51$). Error bars indicate 95% confidence intervals. Asterisks denote significant differences in parameter values ($p < .05$, false-discovery-rate-corrected)

4. Discussion

We reanalyzed sequences of RT data from a visual categorization task from adults and children of different age groups using a multigroup multilevel SEM to investigate carryover (autoregressive) effects. We implemented an advanced SEM for common sample sizes per group in experimental psychology. Even though the model is computationally demanding, the model finally yielded interpretable results, demonstrating that a well-specified SEM can be used to investigate longitudinal phenomena even in noisy experimental measures such as RTs with relatively small samples. We briefly discuss the substantive empirical findings. Then we discuss the model properties and conclude with results from a sample-size simulation to investigate statistical power.

4.1. Conditional Effects and Trends

The estimated SEM provided rich and detailed insights into the behavioral patterns observed in the data set. Importantly, the model subsumed several simpler analyses that had previously been conducted on the data set. A comparison with these prior findings underscores the plausibility

[3]A model comparison using the DIC favored the simpler Bayesian models without the quadratic trend, random trends and random residual variances (differences in DIC: 1424.3, 3446.03, 10635 and 763.78, for age groups 4-5, 6-7, 8-10 year-old children and adults, respectively).

of the present results: The conditional distraction effects replicated previous research by revealing conditional distraction effects for all age groups. This finding is consistent with earlier analyses of this data set as well as with findings from comparable experiments (Volkmer et al., 2022). We also found (conditional) post-novel effects in most age groups, replicating and extending results for children of a wide age range (Wetzel, 2015; Wetzel et al., 2019) and adults (Wetzel, 2015). We observed a post-novel effect of comparable size in children aged 8–10 years, although it was not significantly different from zero. However, we did not replicate the significant differences in post-novel effects between 4-year-old children and other age groups reported by Wetzel et al. (2019) on a subset of the current data set. One possible explanation is that we combined 4- and 5-year-olds into a single age group to achieve acceptable sample sizes per group, whereas only 4-year-olds showed significant differences in the original study.

Again replicating previous research, we found a decline in RTs for novel trials over time (temporal trend), a finding that has been interpreted as an increase in attentional control over time (Volkmer et al., 2022; Wetzel et al., 2016, 2021). Given that these effects have already been studied in this data set, their replication is not surprising. Nevertheless, their replication illustrates that effects typically identified using LMMs can also be equivalently modeled within in multilevel SEM framework, provided that the random effect factors are specified to match the random effects structure of an LMM. Extending the previous findings, we found that children aged 6-10 years showed an increase in RTs over time for standard-sound trials (STA, P2 and P1 for children aged 6-7), possibly indicating fatigue effects (Hoyer et al., 2021). A similar pattern was present in a previous analysis by Volkmer et al. (2022), but a significant increase in standard RTs across blocks was not reported. A reason for this difference may be that the trend analysis in Volkmer et al. (2022) was based on aggregation per block, which was likely less sensitive to small changes, as it was more coarse than our analysis, which involved a linear trend across time. In addition to conditional experimental effects and temporal trends, another set of parameters modeled in previous research comprised the random effect (co)variances. The interindividual differences reflected by random effects have rarely been examined explicitly in experimental psychology despite their potential relevance for theory development. However, our results indicate that, apart from the random intercepts, these parameters are associated with substantial standard errors. A systematic investigation of between-person correlations would require considerably larger sample sizes (Hox & Maas, 2001). Overall, these results underscore the plausibility of the model and highlight the necessity of accounting for temporal trends and random effects to obtain interpretable estimates of autoregressive effects. Given these results, we conclude that the model is appropriate, and thus, we proceed to interpreting the analyses of carryover effects.

4.2. Carryover Effects

We observed significant carryover effects for all trial types among children aged 6-10 years and adults. That is, participants responded particularly slowly in a certain trial if they responded particularly slowly in the previous trial, regardless of trial type. In contrast to our hypothesis, these carryover effects were comparable in size for transitions from novel to post-novel trials compared with transitions from other trial types. An examination of the parameter estimates revealed no clear pattern in the magnitudes of the carryover effects, neither within nor between age groups among children aged 6-10 years and adults. All estimates of autoregressive effects ranged from 0.11 to 0.19, meaning that for a 10 ms slower response on one trial (than expected on the basis of participant characteristics and time), an increase of 1.1 to 1.9 ms in RT on the subsequent trial was predicted–a rather small effect. On the basis of the present results, we suggest that the carryover effects represent a general phenomenon observable in this experimental context—regardless of trial type—rather than an indicator of a specific novel-related reorienting process, at least among children aged 6–10 years and adults. Similar general carryover effects have been reported across various experimental paradigms and could either reflect fluctuation in general processes, such as cognitive control and attention (Voormann & Miller, 2024). This finding is in line with previous research, where models including an AR(1) residual correlation were found to be appropriate for this data set (Volkmer et al., 2022) and for RT data in general.

Interestingly, we found a different pattern for children aged 4-5 years. These young children only yielded carryover effects of 0.20 from post-novel to second-after-novel trials, that is, in transitions when the sound type remained the same standard sound. Carryover effects between trials with varying sound types, that is, standard to novel trials or novel to post-novel (first standard after novel) trials, were close to zero for children aged 4-5 years. This finding may indicate a different underlying process in young children compared to older children either when the task-irrelevant acoustic stimuli change, or when distraction occurs, compared with when the stimuli types remain the same. It could also be the case that the general dynamic fluctuation of attention (reflected by a positive autoregressive effect) is interrupted by distraction. The following predictable repetition of a task-irrelevant sound (each novel sound was followed by at least two standard sounds) may reduce the preceding behavioral variability, resulting in significant carryover effects in the youngest group. This process appears to qualitatively change around age 6, whereas the other effects (conditional effects, trends) show the same basic patterns for all age groups and merely vary in effect size. Future research should explicitly examine sequences of repeated versus changing stimuli across age groups to validate this pattern of results. Nevertheless, the relatively large standard errors indicate uncertainty around these parameter estimates. As is the case with any effect, it is possible that very small differences between parameters within or between age groups were present, but the model lacked the necessary precision to detect it. Similarly, we did not find significant differences in carryover effects between age groups after FDR correction, although children aged 4–

5 years exhibited a different pattern compared with the other age groups. Therefore, we conducted a sample size simulation to estimate the differences in carryover effects that can be reliably detected using this model across varying numbers of observations.

4.3. Sample Size Simulation

It is a common concern that advanced SEMs may be "too complex" for the noisy measurements and limited sample sizes typical of fields such as cognitive psychology (Rosseel, 2020). These concerns often center on potential convergence issues or insufficient statistical power. Our data analysis demonstrated that this type of SEM can be successfully implemented for RT data with samples of $N = 48$ per group, provided that convergence issues are appropriately addressed. In this section, we address the remaining concern: Did we have "sufficient" statistical power to detect carryover (autoregressive) effects? And, building on this question, how many trials from how many participants should be collected when planning a similar study investigating autoregressive effects in trial sequences? To this end, we conducted a simulation to estimate power and 95% confidence interval width for autoregressive effects for three scenarios: (a) testing autoregressive effects against zero, (b) comparing autoregressive effects within a group (i.e., $\alpha_{Nov,\,Sta}$ vs. $\alpha_{P1,\,Nov}$), and (c) comparing autoregressive effects between groups (i.e., $\alpha_{P1,\,Nov}$ for adults vs. children aged 4-5). Sample size planning for models with numerous parameters is inherently challenging because power analysis requires prior information about the parameters, which is typically unavailable before data collection. Making informed assumptions about realistic values for residual variances or covariances among random effects is particularly difficult and often involves a high degree of uncertainty. Therefore, we used the parameter estimates from our empirical analysis as input for a power simulation. We acknowledge that this approach simplifies reality, as the estimates themselves are subject to uncertainty and may vary substantially across experimental designs. Accordingly, the following results should be understood as a rough approximation of what might or might not be realistically achievable in similar designs with similar measures, rather than as precise sample size recommendations. We simulated power for three sample sizes: $N = 30$, $N = 50$ (a realistic sample size in the field of developmental cognitive psychology), and $N = 100$ (a relatively large sample size in the field). We varied the number of sequences per participant between 5 and 40, assuming for simplicity that each participant contributed the same number of sequences. Given that estimating the full model required approximately 1.4 hr per run, we used a reduced model for the simulation. This model excluded the second-after-novel (P2) trial and focused on comparing children aged 4–5 years (who showed the largest residual variances) with adults (who showed the smallest residual variances) to achieve a feasible runtime for 10,000 replications per condition. Further details and simulation code is available in the additional online material. These reduced models converged in 99.95% of runs without any further measures taken for helping with convergence.

We simulated effects of varying sizes that were either even smaller than or equal in magnitude to the effect observed in our data. The results of these simulations are presented in Figure 5. It is important to note that there are no widely accepted interpretations of effect sizes for autoregressive effects, and their interpretation depends on the specific research context. For tests of autoregressive parameters against a fixed value of zero, as well as for comparisons within age groups, we found only minor differences in power between the simulated samples of children and adults. This similarity is reflected in the close alignment of the respective dashed and solid lines in Figure 5. Overall, for relatively small effects of 0.1 and above, a power of 80% can be achieved with $N > 50$ and a reasonable number of sequences, even for between-group comparisons. Among the three types of tests we examined, the highest power was observed for tests of a single autoregressive parameter against zero, followed by the within-age-group comparison of two carryover parameters. The lowest power overall was found for the between-age-group comparison. For both within- and between-group comparisons, statistical power for detecting very small effects of 0.05 remained at or only slightly above the chance level of a coin flip–even with a sample size of $N = 100$ and 40 sequences per participant. In a follow-up simulation, we found that achieving 80% power to detect a 0.05 difference in carryover effects between age groups would require approximately 70 sequences per participant with $N = 100$ (see the additional online material).

To summarize, our simulation shows that autoregressive effects of 0.1 and larger can be detected with reasonably large sample sizes of $N > 50$ per group and a reasonable number of sequences (around 20 per participant) for similar data. To put the effect size into perspective, an autoregressive effect of 0.1 in our case means that a 10 ms increase in RT in a novel trial predicts a 1 ms increase in the subsequent post-novel trial. We consider this a rather small effect, given that, for instance, the average distraction effect for adults was 24 ms, the average post-novel effect was 12 ms, and the within participants residual standard deviation was 70 - 80 ms. For comparisons of autoregressive effects within or between groups, larger sample sizes and numbers of sequences are needed. Given our empirical data situation with about $N = 50$ for children aged 4-5 years and adults, and a mean of 5 and 10 sequences per participant, respectively, we had enough power to reliably detect autoregressive effects of 0.15, but we lacked sufficient power to reliably detect differences within or between groups of 0.15 or smaller.[4]

Again, we emphasize that this simulation was specifically tailored to our experimental design and data, and we do not claim that the results are generalizable to different designs.[5]

[4]Another followup simulation with more effect sizes can be found in the additional online material.

[5]In a series of small simulations we found that the magnitude of trend parameters and between factor means do not influence the power as long as the variance components remain constant.

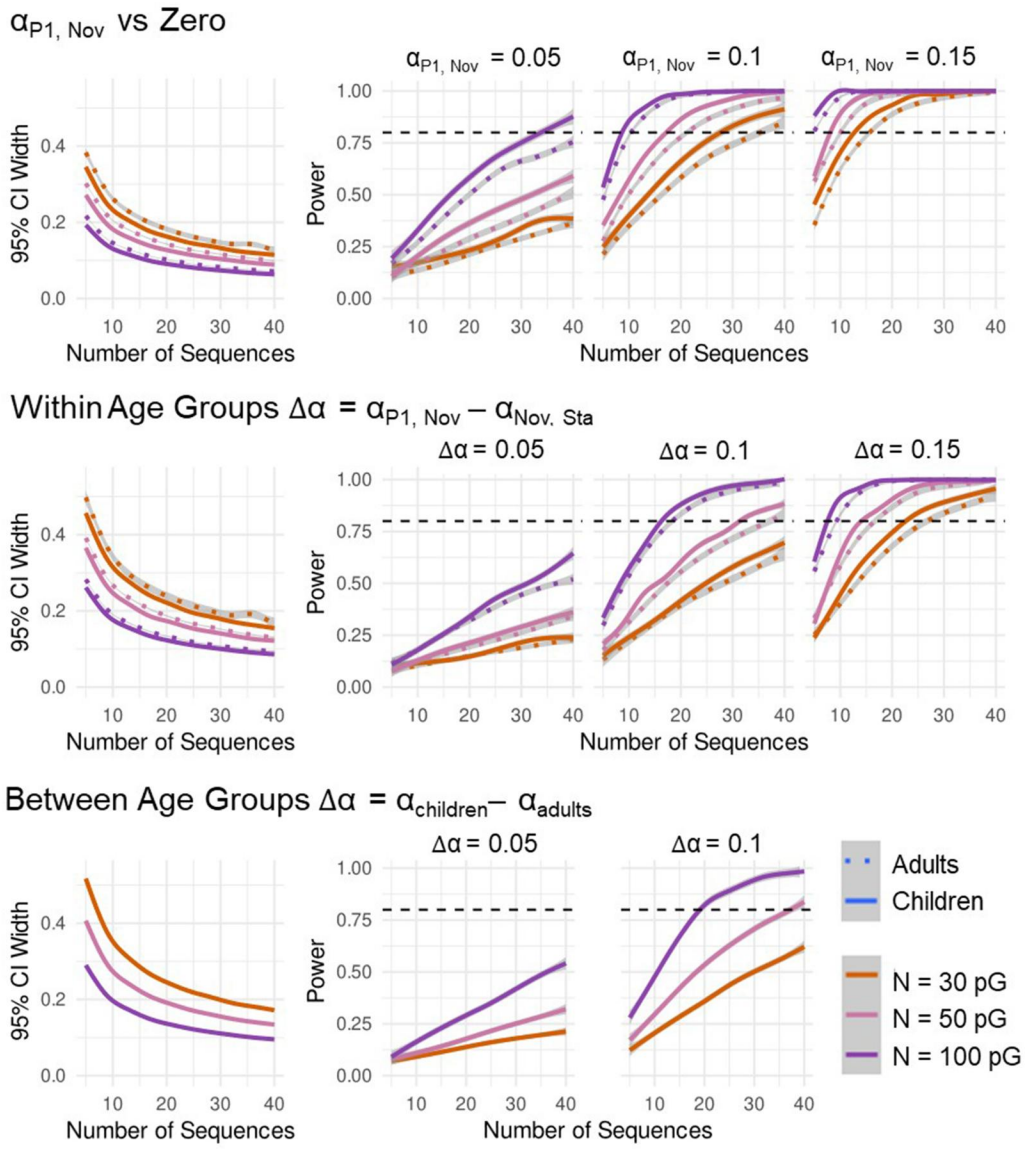

Figure 5. Sample size simulation results.
Note: We simulated power for three sample sizes per age group (pG): $N = 30$, $N = 50$ and $N = 100$. We varied the number of sequences per participant between 5 and 40, assuming for simplicity that each participant contributed the same number of sequences. Each sequence consisted of three trials. Results are depicted for 10,000 replications of a power simulation per panel for different N and different effect sizes for the test of $\alpha_{P1,Nov}$ against 0 (row 1), against $\alpha_{Nov,Sta}$ within age groups (i.e., between experimental conditions manipulated within participants, row 2) and against $\alpha_{P1,Nov}$ between age groups (row 3) for adults (dotted) and 4-5 year-old children (straight line) for 200 trials with different numbers of sequences drawn. Lines are smoothed by General Additive Models (GAM), which model the width of the 95% confidence interval (CI) or the proportion of significant tests (power). Grey areas show 95% confidence intervals of GAM estimation. Power of .80 is shown as a dashed line.

Still, this prospect is promising given the typical measures and sample sizes used in the field, even for data from preschool children. We encourage researchers interested in investigating carryover effects of different trial types to conduct their own power simulations based on their specific study designs and to use reanalyses of existing data as a starting point for defining simulation parameters.

4.4. Evaluation the Potential of Multilevel SEMs for Experimental Psychology

To our knowledge, we are the first to apply a multigroup multilevel SEM approach to experimental longitudinal data. We demonstrated how research questions concerning trial-by-trial processes, such as the dynamics of auditory distraction, can be conceptualized and modeled in terms of longitudinal phenomena, such as trends and autoregressive effects. SEMs offer a flexible framework for modeling these phenomena and their interactions with experimental conditions and between-person characteristics, while also mitigating estimation biases such as the Nickell bias (Asparouhov et al., 2018; Nickell, 1981). Additionally, accounting for different sources of variance, such as random effects and trends, can increase statistical power (Widmann et al., 2024). Our empirical analysis and sample size simulation show that applying these SEMs to typical RT data is feasible, and that reasonably small effects can be detected with sample sizes between 50 and 100 participants per group and 10 to 40 sequences per participant, depending on the parameter of interest. Even smaller samples may be sufficient for less noisy

measures, such as pupil dilation. However, reaching these sample sizes may be difficult in certain experimental contexts (e.g., when testing children or using costly measures). A promising strategy in such cases is to pool data from experiments with identical or highly similar setups and paradigms, as we did in our analysis. Such reanalyses can provide new insights into the temporal dynamics of established phenomena, such as the distraction effect. These insights, in turn, can guide future research and inform the design of new studies aimed at testing emerging hypotheses with adequate statistical power.

5. Conclusion

Our analysis highlights the value of using SEMs to investigate autoregressive effects in repeated-measures experimental data, offering new insights into underlying cognitive and psychological processes. SEMs provide a flexible framework for modeling various longitudinal dynamics and their interactions with experimental conditions and between-person factors, while mitigating estimation biases such as the Nickell bias (Asparouhov et al., 2018; Nickell, 1981). With regard to our substantive research question, we found no evidence for a specific carryover effect that links the post-novel effect to the preceding distraction. Instead, we observed a general attentional dynamic across trials in participants older than six years and a qualitatively different pattern in children aged 4–5 years, suggesting a distinct attentional process in this younger age group. To our knowledge, this is the first study to directly test the reorienting hypothesis for post-novel effects in children and adults, providing initial evidence against this account. More broadly, our findings underscore the potential of longitudinal SEM approaches for advancing cognitive theory. We encourage researchers to conceptualize repeated-measures data explicitly in terms of longitudinal structure, and to make use of the growing set of SEM tools available for such analyses (e.g., Bollen & Zimmer, 2010; Hamaker et al., 2023; Muthén & Muthén, 1998; Rosseel, 2012).

Acknowledgments

We want to thank Louisa Reinhardt for her assistance with data preprocessing and Kristian Genz for his feedback on early versions of the manuscript. We used ChatGPT-4.0 for LaTeX debugging as well as initial language support, as we are not native English speakers. We also thank Jane Zagorski for her language support.

Author Note

R and Mplus scripts for all the analyses reported in this paper can be found in the additional online material: https://osf.io/ryq97/overview?view_only=59551ad5637d48119c2328d529d40f45.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

The data acquisition was funded by the Center for Behavioral Brain Sciences (CBBS) Magdeburg financed by the European Regional Development Fund (ZS/2016/04/78120), the Leibniz Association (P58/2017), and the Deutsche Forschungsgemeinschaft (DFG) (WE5026/1-2).

Data Availability Statement

The pooled dataset is available upon request from the authors to comply with the informed consent filled out by the participants at the time of data collection. The sample size simulation results are available in the additional online material.

References

Ahveninen, J., Jääskeläinen, I. P., Pekkonen, E., Hallberg, A., Hietanen, M., Näätänen, R., Schr¨oger, E., & Sillanaukee, P. (2000). Increased distractibility by task-irrelevant sound changes in abstinent alcoholics. *Alcoholism, Clinical and Experimental Research*, *24*, 1850–1854.

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 359–388. https://doi.org/10.1080/10705511.2017.1406803

Asparouhov, T., & Muthén, B. (2020). Comparison of models for the analysis of intensive longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 275–297. https://doi.org/10.1080/10705511.2019.1626733

Asparouhov, T., & Muthén, B. (2023). Residual structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*, 1–31. https://doi.org/10.1080/10705511.2022.2074422

Baayen, H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234. https://doi.org/10.1016/j.jml.2016.11.006

Baayen, R. H., Fasiolo, M., Wood, S., & Chuang, Y.-Y. (2022). A note on the modeling of the effects of experimental time in psycholinguistic experiments. *The Mental Lexicon*, *17*, 178–212. https://doi.org/10.1075/ml.21012.baa

Bendixen, A., Roeber, U., & Schr¨oger, E. (2007). Regularity extraction & application in dynamic auditory stimulus sequences. *Journal of Cognitive Neuroscience*, *19*, 1664–1677. https://doi.org/10.1162/jocn.2007.19.10.1664

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical & powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berti, S. (2008). Cognitive control after distraction: Event-related brain potentials (ERPs) dissociate between different processes of attentional allocation. *Psychophysiology*, *45*, 608–620. https://doi.org/10.1111/j.1469-8986.2008.00660.x

Bollen, K. A., & Zimmer, C. (2010). An overview of the Autoregressive Latent Trajectory (ALT) model. In K. Van Montfort, J. H. L. Oud, & A. Satorra (Eds.), *Longitudinal research with latent variables* (pp. 153–176). Springer Berlin Heidelberg.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, *38*, 529–569. https://doi.org/10.1207/s15327906mbr3804_5

Curran, P. J., Howard, A. L., Bainter, S. A., Lane, S. T., & McGinley, J. S. (2014). The separation of between-person & within-person components of individual change over time: A latent curve model

with structured residuals. *Journal of Consulting and Clinical Psychology*, *82*, 879–894. https://doi.org/10.1037/a0035297
Dutilh, G., Vandekerckhove, J., Forstmann, B. U., Keuleers, E., Brysbaert, M., & Wagenmakers, E.-J. (2012). Testing theories of post-error slowing. *Attention, Perception & Psychophysics*, *74*, 454–465. https://doi.org/10.3758/s13414-011-0243-2
Escera, C., Alho, K., Schröger, E., & Winkler, I. W. (2000). Involuntary attention & distractibility as evaluated with event-related brain potentials. *Audiology & Neuro-Otology*, *5*, 151–166. https://doi.org/10.1159/000013877
Falkenström, F. (2024). Time-lagged panel models in psychotherapy process & mechanisms of change research: Methodological challenges & advances. *Clinical Psychology Review*, *110*, 102435. https://doi.org/10.1016/j.cpr.2024.102435
Fink, L., Simola, J., Tavano, A., Lange, E., Wallot, S., & Laeng, B. (2024). From pre-processing to advanced dynamic modeling of pupil data. *Behavior Research Methods*, *56*, 1376–1412. https://doi.org/10.3758/s13428-023-02098-1
Hallquist, M. N., & Wiley, J. F. (2018). *MplusAutomation:* An R package for facilitating large-scale latent variable analyses in M *plus*. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 621–638. https://doi.org/10.1080/10705511.2017.1402334
Hamaker, E., Asparouhov, T., & Muthén, B. (2023). Dynamic structural equation modeling as a combination of time series modeling, multilevel modeling, & structural equation modeling. In R. H. Hoyle (Ed.), *The handbook of structural equation modeling* (2nd ed., Chapter 31, pp. 576–596). Guilford Press.
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., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*, 102–116. https://doi.org/10.1037/a0038889
Horváth, J., Winkler, I., & Bendixen, A. (2008). Do N1/MMN, P3a, & RON form a strongly coupled chain reflecting the three stages of auditory distraction? *Biological Psychology*, *79*, 139–147. https://doi.org/10.1016/j.biopsycho.2008.04.001
Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups & small samples. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 157–174. https://doi.org/10.1207/S15328007SEM0802_1
Hoyer, R. S., Elshafei, H., Hemmerlin, J., Bouet, R., & Bidet-Caulet, A. (2021). Why are children so distractible? Development of attention & motor control from childhood to adulthood. *Child Development*, *92*, e716–e737. https://doi.org/10.1111/cdev.13561
Jongerling, J., & Hamaker, E. L. (2011). On the trajectories of the predetermined ALT model: What are we really modeling? *Structural Equation Modeling: A Multidisciplinary Journal*, *18*, 370–382. https://doi.org/10.1080/10705511.2011.582004
Kähkönen, S., Ahveninen, J., Pekkonen, E., Kaakkola, S., Huttunen, J., Ilmoniemi, R. J., & Jääskeläinen, I. P. (2002). Dopamine modulates involuntary attention shifting & reorienting: An electromagnetic study. *Clinical Neurophysiology*, *113*, 1894–1902. https://doi.org/10.1016/s1388-2457(02)00305-x
Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in psychtoolbox-3. *Perception*, *36*, 1–16.
Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2×2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full & partial error correction models. *Psychological Methods*, *16*, 444–467. https://doi.org/10.1037/a0024376
Lee, C.-Y., Yen, H-l., Yeh, P-w., Lin, W.-H., Cheng, Y.-Y., Tzeng, Y.-L., & Wu, H.-C. (2012). Mismatch responses to lexical tone, initial consonant, & vowel in Mandarin-speaking preschoolers. *Neuropsychologia*, *50*, 3228–3239. https://doi.org/10.1016/j.neuropsychologia.2012.08.025
Ljungberg, J. K., & Parmentier, F. B. R. (2012). Cross-modal distraction by deviance: Functional similarities between the auditory & tactile modalities. *Experimental Psychology*, *59*, 355–363. https://doi.org/10.1027/1618-3169/a000164
Lorist, M. M., Klein, M., Nieuwenhuis, S., De Jong, R., Mulder, G., & Meijman, T. F. (2000). Mental fatigue & task control: Planning & preparation. *Psychophysiology*, *37*, 614–625. https://doi.org/10.1111/1469-8986.3750614
Meredith, W. (1993). Measurement invariance, factor analysis & factorial invariance. *Psychometrika*, *58*, 525–543. https://doi.org/10.1007/BF02294825
Mund, M., Johnson, M. D., & Nestler, S. (2021). Changes in size & interpretation of parameter estimates in within-person models in the presence of time-invariant & time-varying covariates. *Frontiers in Psychology*, *12*, 666928. https://doi.org/10.3389/fpsyg.2021.666928
Muthén, B. (2020). *Answer to How to interpret warning saddle point*. http://www.statmodel.com/discussion/messages/12/23080.html?1589925410
Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide* (8th ed.). Muthén & Muthén.
Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, *49*, 1417. https://doi.org/10.2307/1911408
Parmentier, F. B. R. (2014). The cognitive determinants of behavioral distraction by deviant auditory stimuli: A review. *Psychological Research*, *78*, 321–338. https://doi.org/10.1007/s00426-013-0534-4
Parmentier, F. B. R., & Andrés, P. (2010). The involuntary capture of attention by sound: Novelty & postnovelty distraction in young & older adults. *Experimental Psychology*, *57*, 68–76. https://doi.org/10.1027/1618-3169/a000009
R Core Team. (2025). *R: A language & environment for statistical computing*. R Foundation for Statistical Computing.
Roeber, U., Berti, S., Widmann, A., & Schröger, E. (2005). Response repetition vs. response change modulates behavioral & electrophysiological effects of distraction. *Brain Research. Cognitive Brain Research*, *22*, 451–456. https://doi.org/10.1016/j.cogbrainres.2004.10.001
Roeber, U., Widmann, A., & Schröger, E. (2003). Auditory distraction by duration & location deviants: A behavioral & event-related potential study. *Brain Research. Cognitive Brain Research*, *17*, 347–357. https://doi.org/10.1016/s0926-6410(03)00136-8
Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02
Rosseel, Y. (2020). Small sample solutions for structural equation modeling. In *Small sample size solutions* (1st ed., pp. 226–238). Routledge.
Ryan, O., Haslbeck, J. M. B., & Waldorp, L. J. (2025). Non-stationarity in time-series analysis: Modeling stochastic & deterministic trends. *Multivariate Behavioral Research*, *60*, 556–588. https://doi.org/10.1080/00273171.2024.2436413
Schröger, E., & Wolff, C. (1998). Behavioral & electrophysiological effects of task-irrelevant sound change: A new distraction paradigm. *Brain Research. Cognitive Brain Research*, *7*, 71–87. https://doi.org/10.1016/s0926-6410(98)00013-5
Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In *New methods in cognitive psychology* (pp. 4–31). Taylor & Francis.
Volkmer, S., Wetzel, N., Widmann, A., & Scharf, F. (2022). Attentional control in middle childhood is highly dynamic—Strong initial distraction is followed by advanced attention control. *Developmental Science*, *25*, e13275. https://doi.org/10.1111/desc.13275
Von Gunten, C. D., Volpert-Esmond, H. I., & Bartholow, B. D. (2018). Temporal dynamics of reactive cognitive control as revealed by event-related brain potentials. *Psychophysiology*, *55*, e13007. https://doi.org/10.1111/psyp.13007
Voormann, A., & Miller, J. (2024). Sequential effects on reaction time distributions: Commonalities & differences across paradigms. *Journal of Cognition*, *7*, 68. https://doi.org/10.5334/joc.395
Wetzel, N. (2015). Effects of the short-term learned significance of task-irrelevant sounds on involuntary attention in children & adults. *International Journal of Psychophysiology*, *98*, 17–26. https://doi.org/10.1016/j.ijpsycho.2015.06.003
Wetzel, N., Scharf, F., & Widmann, A. (2019). Can't Ignore—Distraction by task-irrelevant sounds in early & middle childhood. *Child Development*, *90*, e819–e830. https://doi.org/10.1111/cdev.13109

Wetzel, N., & Schröger, E. (2014). On the development of auditory distraction: A review. *PsyCh Journal*, *3*, 72–91. https://doi.org/10.1002/pchj.49

Wetzel, N., Schröger, E., & Widmann, A. (2016). Distraction by novel & pitch-deviant sounds in children. *Frontiers in Psychology*, *7*, 1949. https://doi.org/10.3389/fpsyg.2016.01949

Wetzel, N., Widmann, A., Berti, S., & Schröger, E. (2006). The development of involuntary & voluntary attention from childhood to adulthood: A combined behavioral & event-related potential study. *Clinical Neurophysiology*, *117*, 2191–2203. https://doi.org/10.1016/j.clinph.2006.06.717

Wetzel, N., Widmann, A., & Scharf, F. (2021). Distraction of attention by novel sounds in children declines fast. *Scientific Reports*, *11*, 5308. https://doi.org/10.1038/s41598-021-83528-y

Widmann, A., Herrmann, B., & Scharf, F. (2024). Pupillometry is sensitive to speech masking during story listening: The critical role of modeling temporal trends.

Woodward, W., Gray, H. L., & Elliott, A. C. (2017). *Applied time series analysis with R* (2nd ed.). CRC Press.

Xiao, Y., & Liu, H. (2025). Detrending for intensive longitudinal dyadic data analysis using DSEM. *Structural Equation Modeling: A Multidisciplinary Journal*, *32*, 450–459. https://doi.org/10.1080/10705511.2024.2442980

Zwicker, E., Fastl, H., Widmann, U., Kurakata, K., Kuwano, S., & Namba, S. (1991). Program for calculating loudness according to DIN 45631 (ISO 532B). *Journal of the Acoustical Society of Japan (E)*, *12*, 39–42. https://doi.org/10.1250/ast.12.39