Journal of Neuroscience Methods 413 (2025) 110299

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Pupillometry is sensitive to speech masking during story listening: A commentary on the critical role of modeling temporal trends

Andreas Widmann [a,*,1], Björn Herrmann [b,c,2], Florian Scharf [d,3]

[a] Wilhelm Wundt Institute for Psychology, Leipzig University, Germany
[b] Rotman Research Institute, Baycrest Academy for Research and Education, North York, Ontario, Canada
[c] Department of Psychology, University of Toronto, Ontario, Canada
[d] Department of Psychology, University of Kassel, Germany

A B S T R A C T

An increase in pupil size is an important index of listening effort, for example, when listening to speech masked by noise. Specifically, the pupil dilates as the signal-to-noise ratio decreases. A growing body of work aims to assess listening effort under naturalistic conditions using continuous speech, such as spoken stories. However, a recent study found that pupil size was sensitive to speech masking only when listening to sentences but not under naturalistic conditions when listening to stories. The pupil typically constricts with increasing time on task during an experimental block or session, and it may be necessary to account for this temporal trend in experimental design and data analysis in paradigms using longer, continuous stimuli. In the current work, we re-analyze the previously published pupil data, taking into account a problematic constraint of randomization and time-on-task, and use the data to outline methodological solutions for accounting for temporal trends in physiological data using linear mixed models. The results show that, in contrast to the previous work, pupil size is indeed sensitive to speech masking even during continuous story listening. Furthermore, accounting for the temporal trend allowed modeling the dynamic changes in the speech masking effect on pupil size over time as the continuous story unfolded. After demonstrating the importance of accounting for temporal trends in the analysis of empirical data, we provide simulations, methodological considerations, and user recommendations for the analysis of temporal trends in experimental data using linear mixed models.

1. Introduction

Speech masked by background noise reduces intelligibility and increases listening effort. Typically, when listening to masked speech, larger pupil sizes are observed as the signal-to-noise ratio (SNR) decreases, that is, as listening effort increases (up to a point when listening becomes impossible; Wendt et al., 2018; Zekveld and Kramer, 2014). This sensitivity of the pupil size to listening effort is a well-established finding (for review see, for example, the 2018 collection of articles in Hearing Science, including Naylor et al., 2018; Winn et al., 2018; Zekveld et al., 2018). In a recent publication, Cui and Herrmann (2023) demonstrated that fixation duration and spatial gaze dispersion eye movement measures are also sensitive to speech masking during sentence listening (experiments 1 and 2) and story listening (experiment 3). Pupil size was found to be sensitive to speech masking during sentence listening (cf., Cui and Herrmann, 2023), but surprisingly not during story listening. The absence of a signal-to-noise ratio (SNR) effect on pupil size during story listening was attributed to the absence of baseline normalization (i.e., centering around the mean of a baseline period) because for continuous stories no neutral, speech-devoid time period was available (Cui and Herrmann, 2023). The authors concluded that the lack of sensitivity of the pupil size to speech masking may highlight "challenges with pupillometric measures for the assessments of listening effort during naturalistic speech listening." (Cui and Herrmann, 2023). Here, we would like to review two potential shortcomings in the design and analysis of the reported experiment 3 and present a re-analysis of the pupillometry data that addresses these shortcomings. We will show that by accounting for temporal trends in the data analysis, the data favor a conclusion that speech masking does indeed increase pupil size and that thus pupillometry may be used to assess listening effort during story listening. In the theoretical part, we will summarize some recommendations for analyzing temporal trends in experimental data with linear mixed models.

* Correspondence to: Wilhelm Wundt Institute for Psychology, Leipzig University, Neumarkt 9-19, Leipzig D-04109, Germany.
E-mail address: widmann@uni-leipzig.de (A. Widmann).
[1] https://orcid.org/0000-0003-3664-8581
[2] https://orcid.org/0000-0001-6362-3043
[3] https://orcid.org/0000-0003-1659-4774

https://doi.org/10.1016/j.jneumeth.2024.110299
Received 28 March 2024; Received in revised form 25 September 2024; Accepted 9 October 2024
Available online 19 October 2024

2. Empirical part: data re-analysis

2.1. Pupil size and time-on-task temporal trend

In experiment 3 of the study by Cui and Herrmann (2023), individuals listened to two continuous podcast stories of ~10 min duration. Stories were masked by background babble noise at 5 different SNRs: −4, +1, +6, +11, and +16 dB SNR. The SNR level changed seamlessly every 28 s to one of the five levels, leading to overall 22 segments of 28 s duration (in the following referred to as "trials"). The order of SNR levels was randomized with the constraint that the experiment started and ended with the +16 dB SNR level to enable the listener to understand the beginning and end of the story. Each SNR level was presented four times, with the exception of the +16 dB SNR level, which was presented six times (four plus the beginning and end). The two stories were presented in two conditions. In the "intact" condition, the story was presented in the original temporal order. In the "scrambled" condition, the story was cut into short phrases and sentences and shuffled prior to adding background babble. Pupil size was averaged across time points per SNR level and story type ("intact", "scrambled"; effectively removing any information about temporal order or trends). A linear regression function predicting pupil size from SNR level was fitted per participant and story type and tested against zero using one-sample *t* tests per story type and dependent sample *t* tests between story types.

Critically, pupil size can change with the time passing during an experimental session (for a summary see, Fink et al., 2023; Martin et al., 2022; McLaughlin et al., 2023; Unsworth et al., 2019; van den Brink et al., 2016; but see Murphy et al., 2011) and this may have not been sufficiently considered in Cui and Herrmann (2023). Typically, a reduction in the pupil size is most prominent within the first minutes of an experiment. If the pupil size changes over time while a participant listens to a story, the constraint that the story always begins with a specific condition, such as the highest SNR level (+16 dB SNR), could introduce a bias into the results. Specifically, in experiment 3 of Cui and Herrmann (2023), the highest SNR level was systematically presented at a time when pupil size may have been generally large. Although the authors included the highest SNR level also at the end of a story, this may not be sufficient to balance the highest SNR level at the beginning, because the largest pupil-size reduction with time passing typically occurs within the first few minutes. As a result, the average pupil size for the highest SNR level could have been systematically increased, which would counteract the expectation that this SNR level requires less listening effort and should therefore be associated with a *smaller* pupil size. That is, the temporal trend in the data could explain the unexpected result of increased pupil size in the highest SNR condition in the original publication. An analysis accounting for temporal trends in the data may help reveal SNR effects on pupil size.

We have re-analyzed the data replicating the original pre-processing to the best of our knowledge to analyze the effects of time-on-task during story listening (with the exception of the correction of pupil foreshortening error by regression as several of the per participant and story type regression models gave implausible results). The change in pupil size per story type ("intact", "scrambled") over trials is displayed in Fig. 1. On average the mean pupil size decreased by about 25 % from trial 1 to trial 22. A decline of the pupil size during listening to the story was observed in all participants and for both story types except for three participants which showed a small positive change in the "scrambled" story (Fig. 1 Panel B). That is, the constraint to present +16 dB SNR level in trial 1 could have introduced an upward bias in the estimated pupil size for this SNR level.

2.2. Growth curve modeling the time-on-task temporal trend

The results presented above show a trend for the pupil size, such that it decreased with increasing time in the experimental block. This raises a question about the extent to which the change in pupil size over time

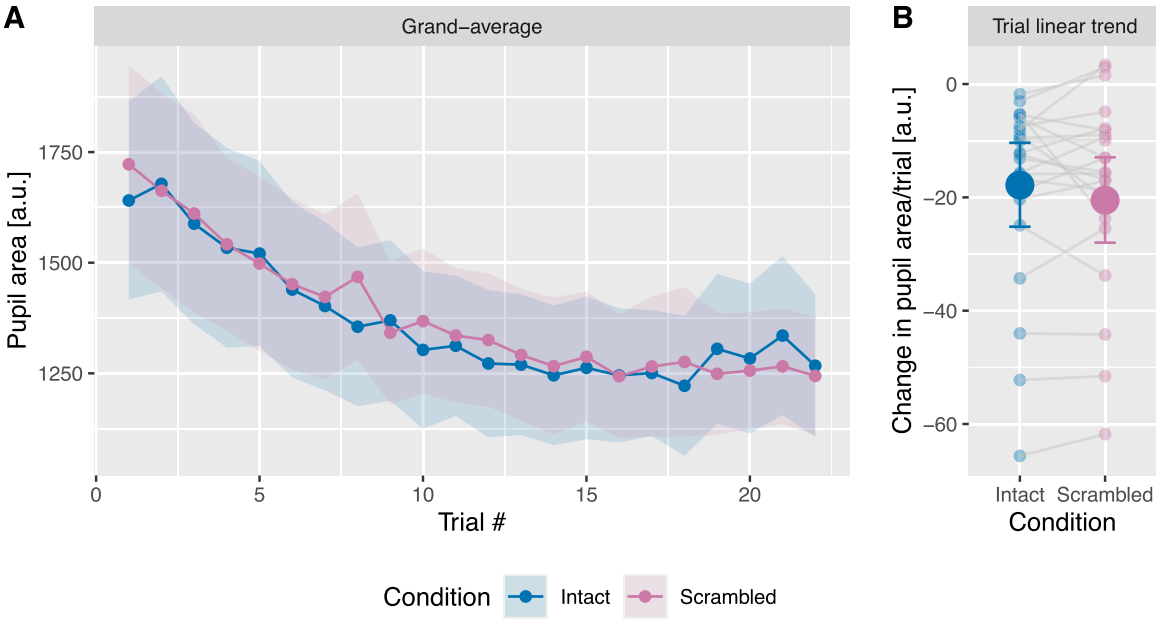

Fig. 1. The observed pupil size significantly decreased during the block in both story types. The grand-average mean pupil size per trial and story type is displayed in Panel A. Panel B illustrates the linear trend for trial estimated from a mixed effects model in which pupil size was allowed to vary quadratically over trials. The linear term can be interpreted as the change of pupil size per trial separately for each story type (large, solid dots), including individual estimates (random effects) of the linear trend for trial (transparent dots) and the trial × story type interaction effect (gray lines).[1] Shaded areas and error bars display 95 % CIs. The trial × story type interaction effect was neither significant for the linear nor for the quadratic term.

[1] Technically, this linear term corresponds to the expected conditional change of pupil size from trial to trial in the middle of the experiment. Here, due to the balanced design, the linear term may also be interpreted as the average (i.e., marginal) expected change from trial to trial across all trials and participants.

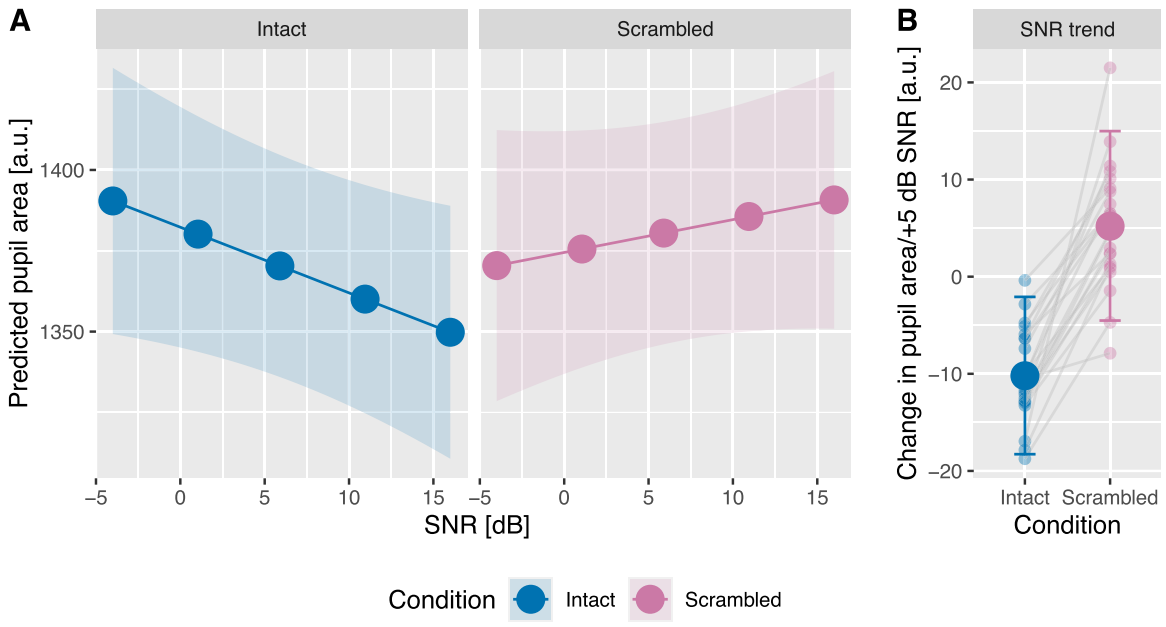

Fig. 2. In the "intact" story type, a significant decrease of −10.3 pupil size units per +5 dB change in SNR was estimated by the model. In the "scrambled" story type, a non-significant increase of 5 pupil size units per +5 dB change in SNR was estimated. Panel A illustrates the predicted pupil size for each type of story and SNR level. Please note that we display marginal values (i.e., "averaged" over trials) which are adjusted for the influence of the variable trial number. Panel B illustrates the estimated SNR effect, that is, the change in pupil size per +5 dB change in SNR for each story type (large, solid dots) including individual estimates of the SNR effect (transparent dots) and the individual differences in the SNR effect between story types (i.e., the interaction effect; gray lines). Shaded areas in Panel A reflect the standard error estimates *within* the average person (for better comparability with Fig. 9 in Cui and Herrmann, 2023, who removed between participants variance for display). They were obtained by first computing the estimates and standard errors in Panel A separately for each participant, and then averaging these values across participants. Error bars in Panel B display 95 % CIs.

may obscure the speech-masking effects of interest. Systematic effects of variables such as time-on-task can be accounted or corrected for by either removing temporal trends for each experimental block and participant prior to further analysis (but missing potential interaction effects) or by including them as covariates in regression models (see for example, Alday, 2019, for a derivation in the framework of the general linear model).

We aim for a more general treatment of the approach, including recommendations in the second part of the current work. We have re-analyzed the data from experiment 3 of Cui and Herrmann's study with a linear mixed model (LMM) predicting pupil size (mean pupil area per 28 s trial) including story type as dummy coded fixed effect (0 = "intact"; 1 = "scrambled"), and SNR level (−4, +1, +6, +11, +16 dB; centered around the mean SNR) and trial number (1–22; centered around the midpoint of the experiment; including linear and quadratic terms) as continuous covariates (pupil_area ~ snr_centered * story_type + trialnr_centered * story_type + I(trialnr_centered^2) * story_type + (1 + story_type + trialnr_centered + I(trialnr_centered^2) | subj); R-code used for model estimation and figure generation can be found online at https://github.com/widmann/cui2023exp3_reanalysis). We opted to fit a raw polynomial model instead of an orthogonal polynomial model because we were explicitly interested in the conditional effect of the experimental condition. It is important to note that both the raw polynomial and orthogonal polynomial models fit equivalent temporal trends. Therefore, this choice primarily affects the interpretation of the model parameters (Mirman, 2014). Critically, any bias introduced by the fixed SNR level of the first and the last trial is also implicitly accounted for by including trial as a covariate in the model.

Accounting for the linear and quadratic effects of trial including their interaction with story type, the growth curve model estimated a significant change of −10.3 units/SNR level (95 % CI [−17.6 −3.0], p =.006) for the "intact" story and a non-significant change of +5 units/SNR level (95 % CI [–2.3 12.3], p =.177) for the "scrambled" story. The SNR level × story type interaction was significant (+15.3 units/SNR level; 95 % CI [5.0 25.6], p =.004). Fig. 2 (Panel A) illustrates the predicted pupil size for each type of story and SNR level.[2] Please note that we display *marginal* values (i.e., "averaged" over trials; see below) which are adjusted for the influence of the variable trial number.

2.3. Modeling the change in the SNR-effect with the time-on-task trend

van den Brink and colleagues (2016) reported that periods during which pupils were constricting were characterized by better performance (e.g., lower reaction times and less false alarms) compared to periods during which pupils were stable or dilating. That is, the temporal derivative of the pupil size had a linear relationship with behavioral performance. In the framework of the adaptive gain theory, pupil constriction might indicate a transition from a tonic to a phasic mode in Locus coeruleus (Gilzenrat et al., 2010). Here, the observed reduction in pupil size during the first half of a story-listening block compared to the relative stability of pupil size during the second half of the block might also imply a change in effort. Such a change in effort should be reflected

[2] Originally, we estimated the LMM in a frequentist framework using (restricted) Maximum-Likelihood estimation. Due to convergence issues of the estimation process, it was not possible to include a random effect for the SNR level × story type interaction effect (i.e., to allow for the interaction effect to vary across participants). In order to illustrate the distributions of the estimated SNR effects per story type and the interaction corresponding to Figure 9 in the original publication (Cui and Herrmann, 2023), we have re-fitted the model as a Bayesian LMM (brms R-package, Bürkner, 2017). This allowed us to circumvent the estimation issues. We used the default minimally informative priors. None of the estimated effects and CIs substantially differed from the frequentist models. We set the number of iterations to a total of 100000 (from 4 chains). Convergence of the Markov Chain Monte Carlo-sampler was established using potential scale reduction, effective sample size estimates as well as visual inspection of trace plots. All measures agreed that the models converged successfully.

in a change in the effect of SNR on pupil size with increasing trial number. We have therefore fitted another Bayesian LMM, also including the three-way interaction SNR level (−4, +1, +6, +11, +16 dB) × story type ("intact", "scrambled") × trial (1−22). The data strongly supported the model including the three-way interaction compared to the model including only the two-way SNR level × story type and trial × story type interaction effects (BF = 406452.8). The SNR level × story type × trial interaction was significant (–1.7 units/trial; 95 % CI [–3.2 –0.1]). The model estimated a significant change of the SNR effect over trials of +1.3 units/trial (95 % CI [0.2 2.4]) for the "intact" story and a non-significant change of the SNR effect over trials of –0.4 units/trial (95 % CI [–1.4 0.7]) for the "scrambled" story. In other words, the observed (conditional) effect of SNR on pupil size for the "intact" story was present and strong at the beginning of the story and declined towards zero at the end of the story. This relation is visualized in Fig. 3.

The observed change in the effect of SNR on the pupil size over time-on-task for the "intact" story may reflect either a true change in the effort invested during story listening or a change in sensitivity of the pupil size over time. In a corresponding model fitted to the fixation duration data, which Cui and Herrmann (2023) also showed to be sensitive to listening effort, we also observed a significant decrease in the SNR effect on fixation duration with time-on-task. Therefore, we suggest that the decrease in the SNR effect with time-on-task for pupil size reflects a true change in effort (e.g., due to practice, motivation, etc.) rather than a change in the sensitivity of pupillometry.

The ability to study listening effort under naturalistic conditions using pupillometry is a promising advance for the field. Story-like speech materials resemble more closely listening conditions in everyday life than traditional sentence paradigms (Bohanek et al., 2009; Mullen and Yi, 1995), where speech follows a coherent narrative thread, and the listener is typically intrinsically motivated to comprehend. Moving towards investigations of listening under naturalistic conditions is further important, because not all speech-perception effects typically investigated with isolated sentences actually generalize to naturalistic speech materials (Irsik et al., 2022). Moreover, assessing listening effort during story listening rather than for disconnected sentences can increase participant compliance, because story materials tend to be highly absorbing and enjoyable (Herrmann and Johnsrude, 2020), enabling examinations of listening effort including its dynamics depending on various variables, such as motivation and interest, speech materials, age, and many others.

In summary, we conclude that pupillometry is sensitive to speech masking during both sentence listening tasks (Kadem et al., 2020; Wendt et al., 2016; Zekveld et al., 2019) and during story listening (current analyses). As shown here, the temporal dynamics of listening effort, which change over time in naturalistic settings, can be modeled using pupillometry. The current analyses show that pupillometric measures may index listening effort during naturalistic speech listening.

3. Theoretical part: consideration of temporal trends in data analysis

3.1. Temporal trends in experimental data

Essentially, experimental data sets can be considered longitudinal data because there are typically multiple measurements (e.g., trials) from each participant and these measurements have a natural temporal order (e.g., trial 1, trial 2, etc.). In the past, these measurements have often been aggregated across trials to cope with the limitations of data-analytic approaches. Recognizing this longitudinal nature of experimental data offers the opportunity to analyze them as a time series, which enables asking new, interesting questions regarding psychological processes. Linear Mixed-effect models (LMMs) allow for analyzing experimental data while avoiding aggregation and have become a standard tool for the statistical analysis of data sets from experimental linguistics to psychology (DeBruine and Barr, 2021). They are well-suited to model longitudinal relationships (see e.g., Raudenbush and Bryk, 2002, chapter 5). Typically, participants serve as the higher-order clustering variable ("second level") whereas trials represent the units of observation for which the outcome such as response time varies ("first level"). In the following, we will describe in more detail why it is important to consider psychological experiments from a time series perspective. We will provide guidance how this can be achieved in the framework of mixed-effect models. Note, however, that LMMs are not the only tool to account for trends. For instance, *detrending* before further analysis is common in many research areas and may also be effective as long as the trend is equal across conditions as will be explained below (but see, Raffalovich, 1994).

In multi-level time series data, various effects may emerge (see for instance, McNeish and Hamaker, 2020, for an overview). These effects can be categorized as follows (Box et al., 2015): (1) *temporal trends*, that is, changes in the mean of the outcome over time (e.g., pupil size decreases over time), (2) *seasonal effects*, that is, mean changes that occur with a regular rhythm and a certain time interval (e.g., RT follows a sinusoidal pattern due to fluctuations of the participants' arousal), (3) *covariate effects* on the outcome either on a participant-level ("between" or time-invariant covariates; e.g., participants with a larger average pupil size tend to have faster RTs) or on the trial-level ("within" or time-varying covariates; e.g., participants respond faster in trials with a concurrent large pupil size), (4) *dynamic effects*, that is, lagged effects of covariates or the outcome of itself from previous time points (e.g., the larger pupil size in trial i, the faster RT in trial $i + 1$). Dynamic effects may be (4a) *autoregressive* effects acting on the level of the original values of the variables (e.g., the slower the raw RT value of trial i, the slower the raw RT in trial $i + 1$), (4b) *residual autoregressive* effects acting on the residuals (i.e., the difference from the model-predicted values, e.g., the slower the residual RT in trial i the faster residual RT in trial $i + 1$), or (4c) *moving average* effects where the outcome value depends on previous residuals (e.g., the slower the residual RT in trial i the faster raw RT in trial $i + 1$). All these effects are potentially interesting depending on the experimental setup and have been investigated occasionally (Bonmassar et al., 2023; Kristjansson et al., 2007; LoTemplio et al., 2021; Tremblay and Newman, 2015; Volkmer et al., 2022). Different effects may occur concurrently depending on the phenomenon under investigation and neglecting any of them can lead to erroneous conclusions. For instance, it may be important to control for autoregressive residuals in addition to temporal trends (van Rij et al., 2019). Such complex scenarios may be investigated in the framework of Dynamic Structural Equation Models (Thorson et al., 2024), but a detailed treatment of these models is beyond the scope of this comment paper. For the remainder of this article, we will focus on temporal trends because the original analysis of Cui and Herrmann (2023) would have benefited from incorporating temporal trend.

Why should rigorous experimentalists bother with temporal trends in the first place? After all, hasn't careful randomization addressed any potential confounders (refer to Thul et al., 2021, for a similar argument)? To understand the impact of unaccounted temporal trends in experimental settings, it is crucial to differentiate whether these temporal trends are consistent across experimental conditions. Fig. 4 shows simulations for an outcome variable (e.g., pupil size) for 100 trials and two experimental conditions A and B to illustrate two scenarios: In the first scenario (illustrated in Fig. 4, Panel A), a temporal trend is present in both condition A and condition B (depicted by colored curves). In other words, the outcome variable decreases over time for both conditions. In the second scenario (illustrated in Fig. 4, Panel C), temporal trends differ between condition A and condition B, with a strong temporal trend in condition A (blue curve) and a weak temporal trend in condition B (green curve). In the first scenario, the experimental effect (i.e., the difference between the blue and the green curve) is constant over time (Fig. 4, Panel B), whereas, in the second scenario, it decreases over time (Fig. 4, Panel D).

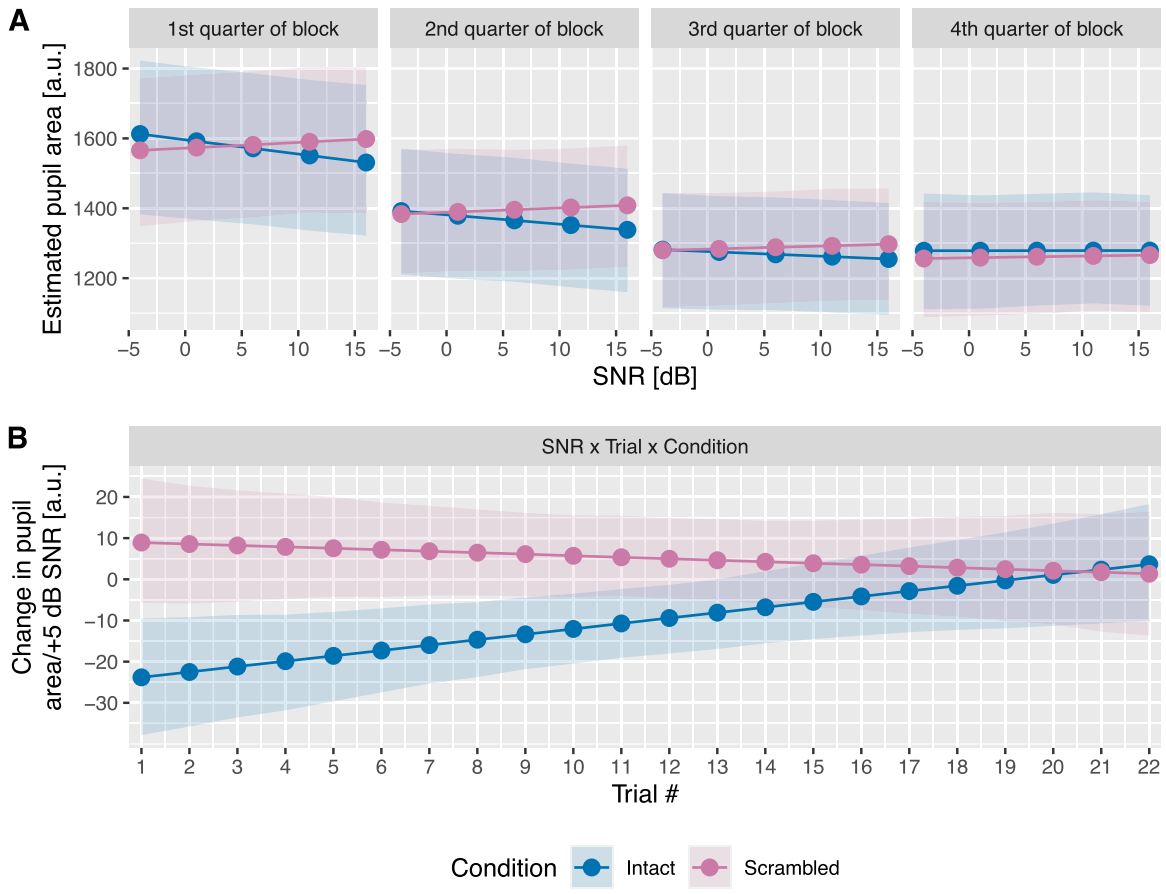

Fig. 3. A significant decrease of pupil size with increasing SNR was observed in the first and second quarter of the block for the "intact" story (cf., 95 % CI illustrated in panel B). Panel A illustrates the predicted pupil size for each type of story and SNR level separately for each quarter of the block. Please note that we display marginal values (i.e., "averaged" over trials) which are adjusted for the influence of the variable trial number. Panel B illustrates the model implied change in pupil size per SNR level per trial and story type. The effect of SNR level on pupil size significantly decreases with time-on-task in the "intact" story condition.

3.2. Consequences of ignoring temporal trends

There are two potential perspectives on the condition effect that can differ when viewing an experimental block or session as longitudinal: Our re-analysis above showed that the SNR effect on the pupil size was larger for early compared to late trials (in fact, absent for the last quarter of the "intact" story). The average across all trials was a mixture of SNR effects for earlier and later trials. The right-hand column compares these two perspectives more formally by computing the condition effect in different ways: The solid line indicates the *conditional effect*, which is determined by the difference in average outcomes between conditions within the same trial. The dashed line represents the *marginal effect*, defined as the difference in average outcomes across all previous trials up to the current trial, for the specific condition. In other words, the marginal effect is what one would analyze when comparing conditions means across an experiment of that length (without considering the temporal trend). That is, the marginal effect represents the average difference between the conditions that one would obtain from an experiment with the respective trial number. Note that the conditional and the marginal effect are equivalent when the temporal trend is identical across conditions (top row), whereas they differ and change over time when temporal trends differ between conditions (lower row; see Fitzmaurice et al., 2011, for a similar discussion in a more general context of longitudinal data modeling). That is, one would find smaller condition effects in longer experiments in the second scenario.

How much temporal trends affect the conclusions that can be drawn from analyses depends on whether the behavior of the outcome variable in a specific experiment aligns more with the first or the second scenario. In the first scenario, an analysis which does not account for the temporal trend can still estimate the experimental effect correctly as long as the presentation of the conditions if fully randomized across time. If trials are randomly assigned to any of the conditions, the marginal mean difference between the conditions reflects the true difference between the curves. However, depending on the experimental context, it can be impossible to fully randomize the presentation of stimuli and conditions across time. For example, Cui and Herrmann (2023) fixed the SNR of the first and the last presented story segments to the highest level in order to enable their participants to understand the beginning and end of the story. Similarly, full randomization across time can be incompatible with the experimental task whenever the order of stimulus presentation is crucial. Imagine that the first 20 trials originate from condition B due to an experimental constraint, for example, familiarization trials. In such a case of partial randomization, time and condition are confounded (i.e., the predictors correlate) and the mean difference between the conditions would be underestimated. Hence, when full randomization is not feasible, time and condition predictors might be correlated and including time in the statistical model is necessary for accurate estimation of the condition effect (Steyer and Schmitt, 1994).

Even if randomization or counterbalancing has addressed potential temporal trends, there are additional reasons to consider time as a predictor. First, even if the estimate of the condition effect is not biased in a specific study, other parameters of the model may still be biased. For instance, it has been shown that the extent to which participants differ in their individual condition effects can be overestimated in the presence of unaccounted trial-level effects such as temporal trends (Barr, 2013; Barr et al., 2013; Bauer and Cai, 2009; Thul et al., 2021). Second, controlling

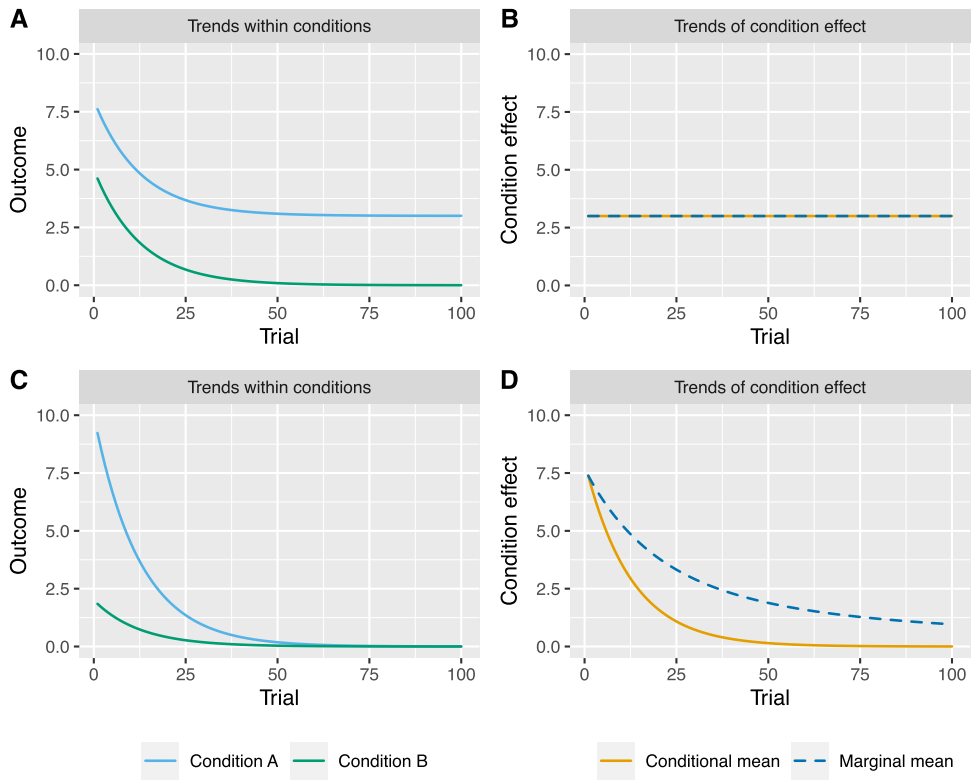

Fig. 4. Illustrating the potential impact of time for two edge cases. Top row: The temporal trend is identical across conditions (Panel A, blue and green solid lines). Consequently, neither conditional (difference between the lines in the left plot) nor marginal condition effect (difference in cumulative means between conditions up to the respective trial) vary over time (Panel B). Bottom row: The temporal trend differs between conditions (Panel C). Conditional and marginal effects are not equivalent anymore and depend on the time (Panel D). Put more simply, the marginal effect represents the average difference between the conditions that one would obtain from an experiment with the respective trial number. That is, one would find smaller condition effects in longer experiments.

for temporal trends by incorporating them into the statistical model can increase statistical power when testing for condition effects (Thul et al., 2021). We illustrated this in a simplistic simulated data set in Fig. 5 for a more intuitive understanding. The two panels contrast the results of mean comparison before and after accounting for a temporal trend. Importantly, although both analyses arrive at the same estimates for the condition averages, the confidence intervals are much narrower after accounting for the trend. This phenomenon is well-known in regression analysis. Including meaningful predictors in the model consistently reduces the residual variance of the outcome variable and, hence, reduces the standard error for other effects of interest (see, e.g., Fahrmeir et al., 2013). Therefore, controlling for known temporal trends enhances the

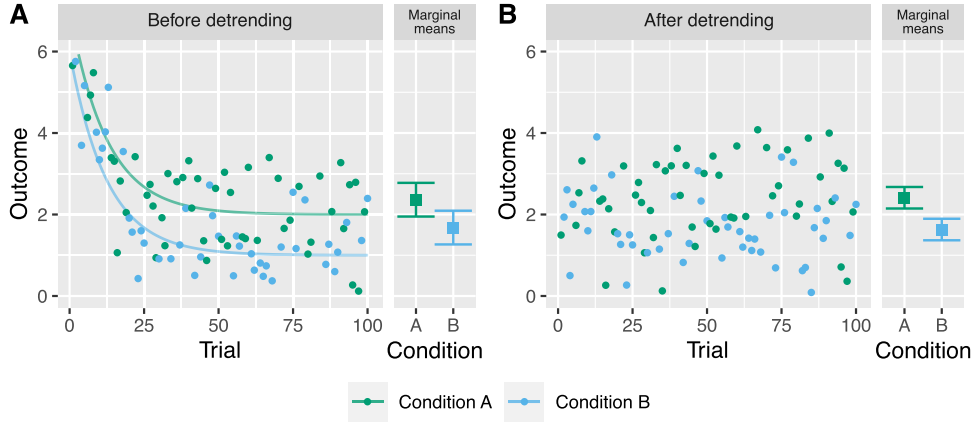

Fig. 5. Illustrating the potential power increase by accounting for temporal trends. We simulated data according to the first scenario with an identical temporal trend across conditions (Panel A, green and blue solid lines). Each dot represents the outcome value for a specific trial *within* a single participant. (For the sake of simplicity, we refrained from simulating data from multiple participants). The colored square dots represent the marginal means for both conditions and their corresponding 95 % CIs. In Panel A, the marginal means were computed from the original data without accounting for the temporal trend. In Panel B, the trend was first approximated by a fourth-order polynomial. Then, the residuals of this model were used (for better visual comparability, we added the grand-mean of the outcome to the residuals) and the same analysis was repeated. Note that the same marginal means were estimated in both panels, but the CIs are much narrower after detrending (Panel B), reflecting an increase in statistical power for tests of mean differences. Adding time as a predictor to the model achieves the same but without the need for a two-step procedure (and allows detecting changes of condition effects over time, that is, interactions of condition effects with temporal trends).

statistical power for the effects that researchers are genuinely interested in, even in the first scenario.

Beyond these arguments, the impact of temporal trends on the results of data analyses is even greater in the second scenario where temporal trends differ between conditions (Fig. 4, Panel C). As outlined above, conditional means and marginal means differ systematically in that case if the experimental effect varies over time. This is potentially problematic for substantive conclusions because the marginal effects can become uninterpretable or misleading. For instance, in our example in Fig. 4, researchers would systematically find smaller effects in longer experiments with more trials–even if they tried to account for temporal trends in their experimental design with full randomization. In the case of differential temporal trends between conditions, experimental effects should be interpreted *conditional* for specific times points. To model the (conditional) experimental effect at arbitrary time points within the experiment, it is necessary to model an interaction effect between the condition effect and time in the statistical model (as in our re-analysis of Cui and Herrmann, 2023). For a more intuitive grasp, this situation closely resembles a phenomenon in analyses of variance: Imagine, condition and time as binary predictors (e.g., first vs. second half of the experiment). If there is no interaction between the predictors, one may safely interpret the marginal mean difference ("main effect") between the cells of the single predictors (e.g., the marginal mean difference of the conditions averaged over experiment halves). However, if there is an interaction effect, caution is advised and researchers need to investigate the conditional (or "simple") effects of the predictors *within* the cells of the other predictor, that is, the conditional effects of the predictors.

Considering the high prevalence of temporal trends within experimental sessions (Dignath et al., 2019; Gouret and Pfeuffer, 2021; Langner et al., 2010; Volkmer et al., 2022; Wetzel et al., 2021), we strongly recommend routinely exploring temporal trends as part of the statistical analysis. Otherwise, the length of the experiment may become a hidden moderator, and findings may be inconsistent across studies (Volkmer et al., 2022; Wetzel et al., 2021). Apart from a technical perspective, the investigation of temporal trends in experimental effects offers new interesting insights (Baayen et al., 2022). For instance, it was found that participants from various age groups mainly differ in the temporal trends of their distraction effects (Volkmer et al., 2022; Wetzel et al., 2021). An interaction between time and condition can represent various psychological processes. Depending on the experimental context, it may provide insights into learning processes (e.g., "Are younger participants slower to learn to shield against behavioral distraction than older participants?"), arousal fluctuations (e.g., "Does quicker habituation of changes in pupil size across time reveal meaningful neurophysiological differences between participants?"), or assist in characterizing the processes related to different outcome measures (e.g., "Do changes in pupil size and response time exhibit similar or distinct temporal trends?"). In conclusion, exploring experimental effects in relation to time and other inter-individual variables may provide genuinely new insights into the underlying processes and prevent misinterpretations of marginal effects (Baayen et al., 2022).

3.3. Accounting for temporal trends with mixed effect models

After emphasizing the importance of considering temporal trends in experimental psychology, we would like to conclude by offering guidance on how to approach this. Although we recognize a need for an accessible introduction to advanced statistical modeling options, a detailed tutorial is beyond the scope of this article. As mentioned, mixed effect models are not the only data-analytic approach to account for trends. Nevertheless, we would like to briefly summarize some options within this framework for readers who want to gain a deeper understanding of the methods used in our re-analysis and point these readers to available implementations in the open statistical software R (R-Core-Team, 2022). Three prominent options exist for modeling temporal trends with mixed effect models, which also allow effects to be non-linear and heterogeneous across participants (see, Beller and Baier, 2013): (1) exact parametric mixed-effect models, (2) approximate linear or polynomial mixed-effect models, (3) semi-parametric mixed-effect splines (also known as generalized additive linear mixed models, GAMMs).

Exact parametric models require precise knowledge regarding the functional form of the temporal trend because they try to estimate a specific functional form of the temporal trend. For instance, one may model an exponential decay or a sinusoidal fluctuation of the outcome if the underlying theory predicts it (Lindstrom and Bates, 1990). This can be achieved in the R packages nlme and brms (Bürkner, 2017; Pinheiro et al., 2023). However, these models are notoriously hard to estimate, and precise knowledge of the functional form is rarely available. *Approximate models* address this by fitting a linear or polynomial relationship of the outcome variable with time. That is, using an approximate model, researchers admit that the true functional form is unknown and try to fit a "good enough" alternative function to the data. In theory, with increasing order of the polynomial, arbitrarily complex temporal trends could be approximated given sufficient data. These models are also referred to as growth curve models in the literature (Kristjansson et al., 2007) and can be estimated, for instance, in lme4, nlme or brms (Bates et al., 2015; Bürkner, 2017; Pinheiro et al., 2023). Linear or polynomial growth curve models are easy to implement and comprehensible as long as the degree of the polynomial is rather low (say third or fourth order). It is easily possible to derive characteristic points of the trend from them including corresponding inferential statistics and individual estimates for participants (McCormick, 2023). Even more complex non-linear relationships are more easily modeled using splines. Splines are a class of semi-parametric models that are able to approximate arbitrarily complex smooth relationships including arbitrarily complex interactions (Baayen et al., 2022). They are typically interpreted using graphical illustrations of the modeled relationships such as contour plots or by deriving characteristic points on the curves. They offer a highly flexible data-driven modeling approach, but they are harder to interpret, especially if many continuous predictors are involved. Mixed-effect splines are implemented in the packages gamm4 and brms (Bürkner, 2017; Wood and Scheipl, 2020).

In sum, all mentioned modeling approaches can account for temporal trends and the choice of the modeling approach ultimately depends on the analysis goals. In our re-analysis of the data from Cui and Herrmann (2023), we did not have specific expectations regarding the form of the trend (except that we expected a non-linear decrease of pupil size over time). Therefore, we used the second approach and fit a polynomial growth curve model to the data. Model comparisons revealed that a second-order polynomial was sufficient to capture the decrease in pupil size and higher-order polynomials did not improve the model fit any further. Furthermore, the temporal trend was rather consistent across participants. Based on these results, we would not expect much gain from the increased complexities of a spline-based analysis.

4. Conclusion

In many experiments, the outcome variable of interest may be affected by time, order, repetition, or other longitudinal confounding factors. Using the data of Cui and Herrmann (2023), we demonstrated that this can significantly affect the conclusions that are drawn from the data. Our re-analysis shows that the pupil size increases as speech masking increases, even during continuous story listening. Moreover, we show through simulated examples that explicitly modelling temporal trends during data analysis can avoid biases, increase power, and gain insight into the temporal dynamics of experimental effects. We hope to raise awareness of the presence of temporal trends in experimental data, their impact on outcome variables (e.g., such as the pupil size), and provide recommendations for modeling approaches.

CRediT authorship contribution statement

Andreas Widmann: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Björn Herrmann:** Writing – review & editing, Investigation, Data curation. **Florian Scharf:** Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We would like to thank Salomé Li Keintzel for her assistance with literature research and for her helpful comment on our list of possible effects in a time series. BH is supported by the Canada Research Chair program (CRC-2019–00156), the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant: RGPIN-2021–02602), and the Canadian Institutes of Health Research (CIHR: 186236).

Data Availability

We have re-analyzed an already publicly available dataset.

References

Alday, P.M., 2019. How much baseline correction do we need in ERP research? Extended GLM model can replace baseline correction while lifting its limits. Psychophysiology 56 (12), e13451. https://doi.org/10.1111/psyp.13451.

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. Ment. Lex. 17 (2), 178–212. https://doi.org/10.1075/ml.21012.baa.

Barr, D.J., 2013. Random effects structure for testing interactions in linear mixed-effects models. Front. Psychol. 4, 328. https://doi.org/10.3389/fpsyg.2013.00328.

Barr, D.J., Levy, R., Scheepers, C., Tily, H.J., 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. J. Mem. Lang. 68 (3). https://doi.org/10.1016/j.jml.2012.11.001.

Bates, D., Mächler, M., Bolker, B.M., Walker, S.C., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67 (1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Bauer, D.J., Cai, L., 2009. Consequences of unmodeled nonlinear effects in multilevel models. J. Educ. Behav. Stat. 34 (1), 97–114. https://doi.org/10.3102/1076998607310504.

Beller, J., Baier, D., 2013. Differential effects: are the effects studied by psychologists really linear and homogeneous? Eur. 'S. J. Psychol. 9 (2), 378–384. https://doi.org/10.5964/ejop.v9i2.528.

Bohanek, J.G., Fivush, R., Zaman, W., Lepore, C.E., Merchant, S., Duke, M.P., 2009. Narrative interaction in family dinnertime conversations. Merrill-Palmer Q. 55 (4), 488–515. https://doi.org/10.1353/mpq.0.0031.

Bonmassar, C., Scharf, F., Widmann, A., Wetzel, N., 2023. On the relationship of arousal and attentional distraction by emotional novel sounds. Cognition 237, 105470. https://doi.org/10.1016/j.cognition.2023.105470.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015. Time Series Analysis: Forecasting and Control, 5th ed. Wiley.

Bürkner, P.C., 2017. brms: an R package for bayesian multilevel models using stan. J. Stat. Softw. 80 (1), 1–28. https://doi.org/10.18637/jss.v080.i01.

Cui, M.E., Herrmann, B., 2023. Eye movements decrease during effortful speech listening. J. Neurosci. 43 (32), 5856–5869. https://doi.org/10.1523/JNEUROSCI.0240-23.2023.

DeBruine, L.M., Barr, D.J., 2021. Understanding mixed-effects models through data simulation. Adv. Methods Pract. Psychol. Sci. 4 (1). https://doi.org/10.1177/2515245920965119.

van den Brink, R.L., Murphy, P.R., Nieuwenhuis, S., 2016. Pupil diameter tracks lapses of attention. PLoS One 11 (10), e0165274. https://doi.org/10.1371/journal.pone.0165274.

Dignath, D., Berger, A., Spruit, I.M., van Steenbergen, H., 2019. Temporal dynamics of error-related corrugator supercilii and zygomaticus major activity: evidence for implicit emotion regulation following errors. Int. J. Psychophysiol. 146, 208–216. https://doi.org/10.1016/j.ijpsycho.2019.10.003.

Fahrmeir, L., Kneib, T., Lang, S., Marx, B., 2013. Regression: models, methods and applications. Springer. https://doi.org/10.1007/978-3-642-34333-9.

Fink, L., Simola, J., Tavano, A., Lange, E., Wallot, S., Laeng, B., 2023. From pre-processing to advanced dynamic modeling of pupil data. Behav. Res. Methods. https://doi.org/10.3758/s13428-023-02098-1.

Fitzmaurice, G.M., Laird, N.M., Ware, J.H., 2011. Contrasting marginal and mixed effects models. In. In: Fitzmaurice, G.M., Laird, N.M., Ware, J.H. (Eds.), Applied Longitudinal Analysis. Wiley, pp. 473–486. https://doi.org/10.1002/9781119513469.ch16.

Gilzenrat, M.S., Nieuwenhuis, S., Jepma, M., Cohen, J.D., 2010. Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. Cogn., Affect., Behav. Neurosci. 10 (2), 252–269. https://doi.org/10.3758/CABN.10.2.252.

Gouret, F., Pfeuffer, C.U., 2021. Learning to expect and monitor the future: how fast do anticipatory saccades toward future action consequences emerge? J. Exp. Psychol.: Hum. Percept. Perform. 47 (7), 992–1008. https://doi.org/10.1037/xhp0000924.

Herrmann, B., Johnsrude, I.S., 2020. Absorption and enjoyment during listening to acoustically masked stories, 2331216520967850 Trends Hear. 24. https://doi.org/10.1177/2331216520967850.

Irsik, V.C., Johnsrude, I.S., Herrmann, B., 2022. Age-related deficits in dip-listening evident for isolated sentences but not for spoken stories. Sci. Rep. 12 (1), 5898. https://doi.org/10.1038/s41598-022-09805-6.

Kadem, M., Herrmann, B., Rodd, J.M., Johnsrude, I.S., 2020. Pupil dilation is sensitive to semantic ambiguity and acoustic degradation, 2331216520964068 Trends Hear. 24. https://doi.org/10.1177/2331216520964068.

Kristjansson, S.D., Kircher, J.C., Webb, A.K., 2007. Multilevel models for repeated measures research designs in psychophysiology: an introduction to growth curve modeling. Psychophysiology 44 (5), 728–736. https://doi.org/10.1111/j.1469-8986.2007.00544.x.

Langner, R., Steinborn, M.B., Chatterjee, A., Sturm, W., Willmes, K., 2010. Mental fatigue and temporal preparation in simple reaction-time performance. Acta Psychol. 133 (1), 64–72. https://doi.org/10.1016/j.actpsy.2009.10.001.

Lindstrom, M.J., Bates, D.M., 1990. Nonlinear mixed effects models for repeated measures data. Biometrics 46 (3), 673–687. https://doi.org/10.2307/2532087.

LoTemplio, S., Silcox, J., Federmeier, K.D., Payne, B.R., 2021. Inter- and intra-individual coupling between pupillary, electrophysiological, and behavioral responses in a visual oddball task. Psychophysiology 58 (4), e13758. https://doi.org/10.1111/psyp.13758.

Martin, J.T., Whittaker, A.H., Johnston, S.J., 2022. Pupillometry and the vigilance decrement: task-evoked but not baseline pupil measures reflect declining performance in visual vigilance tasks. Eur. J. Neurosci. 55 (3), 778–799. https://doi.org/10.1111/ejn.15585.

McCormick, E.M., 2023. Deriving models of change with interpretable parameters: linear estimation with nonlinear inference. PsyArxiv. https://doi.org/10.31234/osf.io/r4vxb.

McLaughlin, D.J., Zink, M.E., Gaunt, L., Reilly, J., Sommers, M.S., Van Engen, K.J., Peelle, J.E., 2023. Give me a break! Unavoidable fatigue effects in cognitive pupillometry. Psychophysiology 60 (7), e14256. https://doi.org/10.1111/psyp.14256.

McNeish, D., Hamaker, E.L., 2020. A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. Psychol. Methods 25 (5), 610–635. https://doi.org/10.1037/met0000250.

Mirman, D., 2014. Growth curve analysis and visualization using R. Chapman and Hall/CRC. https://doi.org/10.1201/9781315373218.

Mullen, M.K., Yi, S.H., 1995. The cultural-context of talk about the past - implications for the development of autobiographical memory. Cogn. Dev. 10 (3), 407–419.. https://doi.org/10.1016/0885-2014(95)90004-7.

Murphy, P.R., Robertson, I.H., Balsters, J.H., O'Connell, R., G., 2011. Pupillometry and P3 index the locus coeruleus-noradrenergic arousal function in humans. Psychophysiology 48 (11), 1532–1543. https://doi.org/10.1111/j.1469-8986.2011.01226.x.

Naylor, G., Koelewijn, T., Zekveld, A.A., Kramer, S.E., 2018. The application of pupillometry in hearing science to assess listening effort. Trends Hear. 22, 2331216518799437. https://doi.org/10.1177/2331216518799437.

Pinheiro, J., Bates, D., & R-Core-Team. (2023). nlme: Linear and Nonlinear Mixed Effects Models (R package version 3.1-164). https://cran.r-project.org/package=nlme.

Raffalovich, L.E., 1994. Detrending time-series - a cautionary note. Sociol. Methods Res. 22 (4), 492–519.. https://doi.org/10.1177/0049124194022004003.

Raudenbush, S.W., Bryk, A.S., 2002. Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd ed. Sage Publications.

R-Core-Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. ⟨https://www.R-project.org/⟩.

van Rij, J., Hendriks, P., van Rijn, H., Baayen, R.H., Wood, S.N., 2019. Analyzing the time course of pupillometric data, 2331216519832483 Trends Hear. 23. https://doi.org/10.1177/2331216519832483.

Steyer, R., Schmitt, T., 1994. The theory of confounding and its application in causal modeling with latent variables. In: von Eye, A., Clogg, C.C. (Eds.), Latent variables analysis: Applications for developmental research. Sage Publications, pp. 36–67.

Thorson, J.T., Andrews III, A.G., Essington, T.E., Large, S.I., 2024. Dynamic structural equation models synthesize ecosystem dynamics constrained by ecological mechanisms. Methods Ecol. Evol. 15 (4), 744–755. https://doi.org/10.1111/2041-210x.14289.

Thul, R., Conklin, K., Barr, D.J., 2021. Using GAMMs to model trial-by-trial fluctuations in experimental data: More risks but hardly any benefit. J. Mem. Lang. 120. https://doi.org/10.1016/j.jml.2021.104247.

Tremblay, A., Newman, A.J., 2015. Modeling nonlinear relationships in ERP data using mixed-effects regression with R examples. Psychophysiology 52 (1), 124–139. https://doi.org/10.1111/psyp.12299.

Unsworth, N., Robison, M.K., Miller, A.L., 2019. Individual differences in baseline oculometrics: Examining variation in baseline pupil diameter, spontaneous eye blink rate, and fixation stability. Cogn., Affect., Behav. Neurosci. 19 (4), 1074–1093. https://doi.org/10.3758/s13415-019-00709-z.

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. Dev. Sci. 25 (6), e13275. https://doi.org/10.1111/desc.13275.

Wendt, D., Dau, T., Hjortkjaer, J., 2016. Impact of background noise and sentence complexity on processing demands during sentence comprehension. Front. Psychol. 7, 345. https://doi.org/10.3389/fpsyg.2016.00345.

Wendt, D., Koelewijn, T., Ksiazek, P., Kramer, S.E., Lunner, T., 2018. Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test. Hear. Res. 369, 67–78. https://doi.org/10.1016/j.heares.2018.05.006.

Wetzel, N., Widmann, A., Scharf, F., 2021. Distraction of attention by novel sounds in children declines fast. Sci. Rep. 11 (1), 5308. https://doi.org/10.1038/s41598-021-83528-y.

Winn, M.B., Wendt, D., Koelewijn, T., Kuchinsky, S.E., 2018. Best practices and advice for using pupillometry to measure listening effort: an introduction for those who want to get started, 2331216518800869 Trends Hear. 22. https://doi.org/10.1177/2331216518800869.

Wood, S., & Scheipl, F. (2020). gamm4: Generalized Additive Mixed Models using mgcv and lme4 (R package version 0.2-6). https://cran.r-project.org/package=gamm4.

Zekveld, A.A., Koelewijn, T., Kramer, S.E., 2018. The pupil dilation response to auditory stimuli: current state of knowledge, 2331216518777174 Trends Hear. 22. https://doi.org/10.1177/2331216518777174.

Zekveld, A.A., Kramer, S.E., 2014. Cognitive processing load across a wide range of listening conditions: insights from pupillometry. Psychophysiology 51 (3), 277–284. https://doi.org/10.1111/psyp.12151.

Zekveld, A.A., van Scheepen, J.A.M., Versfeld, N.J., Veerman, E.C.I., Kramer, S.E., 2019. Please try harder! The influence of hearing status and evaluative feedback during listening on the pupil dilation response, saliva-cortisol and saliva alpha-amylase levels. Hear. Res. 381, 107768. https://doi.org/10.1016/j.heares.2019.07.005.