Limited cognitive resources reduce the language predictability benefit across the adult lifespan

Merle Schuckart[1,2]*, Sandra Martin[3]*, Sarah Tune[1,2], Lea-Maria Schmitt[4], Gesa Hartwigsen[3,5]⋆ & Jonas Obleser[1,2]⋆

[1] Department of Psychology, University of Lübeck, Germany
[2] Center of Brain, Behavior and Metabolism, University of Lübeck, Germany
[3] Research Group Cognition and Plasticity, Max Planck Institute for Human Cognitive and Brain Sciences, Germany
[4] Donders Institute for Brain, Cognition and Behaviour, Radboud University, The Netherlands
[5] Wilhelm Wundt Institute for Psychology, Leipzig University, Germany
* Shared first authorship
⋆ Shared last authorship

ABSTRACT

In everyday communication, humans predict upcoming language seemingly effortlessly. However, it remains unclear to what extent the formation of such predictions taxes executive resources. Our study set out to investigate how a limitation of executive resources impacts natural language prediction on multiple timescales in a novel dual-task paradigm, and how this impact is modulated by age. Participants (N = 175; 18–85 years) read short newspaper articles, presented word by word in varying font colours. This self-paced reading task was either performed in isolation or paired with a competing n-back task (1-back or 2-back) on the words' font colour. We measured word-level reading time and block-level reading comprehension as well as n-back performance. To quantify word predictability, we estimated word surprisal on four distinct timescales (i.e., context lengths ranging from words to paragraphs) using a large language model. Under high cognitive load, adults aged 60 and over benefited most from high word predictability. Our results show that independent of timescale, higher cognitive load diminishes the benefits of high word predictability on reading time, suggesting language predictions draw on executive resources.

INTRODUCTION

In everyday life, we constantly rely on our ability to accurately and swiftly process linguistic input. When reading a book, watching television, or manoeuvring a car through busy traffic while heeding navigation instructions, language prediction can act as a catalyst to facilitate linguistic processing (Pickering & Garrod, 2007; Onnis et al., 2022). It remains to be tested whether maintaining such an elaborate language process really incurs no cognitive cost itself.

Language processing is posited to heavily lean on predictive coding mechanisms: Predictions from higher levels in the processing hierarchy are passed on top-down to lower levels, where they are compared to the actual incoming signals. If a prediction aligns with the actual stimulus, there is no need to transmit information about the stimulus bottom-up. Only in case of a mismatch, the unexpected difference of actual and predicted signal – i.e., the prediction error – is transmitted bottom-up in the processing hierarchy, where it can be considered for future predictions, allowing the system to continuously adapt to a changing environment. The pre-activation of upcoming linguistic input reduces processing demands compared to a scenario where all incoming signals have to be processed along the entire language processing hierarchy, which facilitates real-time language processing (Clark, 2013; Rao & Ballard, 1999; Ryskin & Nieuwland, 2023; Schmitt et al., 2021).

The hierarchical organisation of language processing in the cortex indicates that the representation of previous context is not necessarily restricted to a singular process happening word by word: Instead, language is processed on multiple levels of abstraction, referred to as *timescales*, ranging from individual linguistic features to progressively more complex representations of previous context (Nieuwland et al., 2018; Ryskin & Nieuwland, 2023). Indeed, several recent neuroimaging studies report evidence for speech prediction on different timescales, from phonemes (Ettinger et al., 2014; Donhauser & Baillet, 2020; Ylinen et al., 2016) to individual words, phrases, sentences, and larger narrative structures (Heilbron et al., 2022; Lerner et al., 2011; Schmitt et al., 2021).

Typically, when investigating mechanisms underlying language prediction, researchers focus on smaller linguistic units such as single words or sentences, while trying to minimise any distraction during the experiment that could divert executive resources from the language task at hand. However, while suitable for many enquiries, this approach cannot fully capture the complex dynamics of language comprehension in everyday life, where individuals draw information from longer, natural text segments to prepare responses or perform secondary tasks. It has been hypothesised that in such demanding situations, language-specific processes are supported by

executive resources (Diachek et al., 2020; MacGregor et al., 2022; Ryskin & Nieuwland, 2023). However, to date, it is unclear to which extent the formation and utilisation of language predictions taxes executive resources (MacGregor et al., 2022; Shain et al., 2022; Xie et al., 2023; Ryskin & Nieuwland, 2023), and how this dynamic interplay between the two systems is modulated by age.

Advancing age has been shown to be associated with sensory and executive decline (Idowu & Szameitat, 2023; Salthouse et al., 2003; Verhaeghen et al., 2003). However, previous research presents conflicting results on how cognitive ageing affects the use of linguistic predictions. Behavioural studies show a decline in using context to make semantic predictions with age (e.g., Häuser et al., 2018), yet EEG studies reveal a varied response, indicating either heightened sensitivity to unexpected information in older adults (Cheimariou et al., 2019) or no significant age-related differences in neural reactions (Dave et al., 2018; Payne & Federmeier, 2018).

In the current study, we thus investigated the effect of limited executive resources on language prediction during natural reading across the adult lifespan. Specifically, we asked how increased cognitive load affects language prediction on four different timescales, ranging from single words to entire paragraphs. Moreover, we were interested in how the generation of language predictions change across the adult lifespan.

To this end, we employed a novel dual-task paradigm, incorporating a self-paced reading task alongside an n-back task. We tested the following hypotheses: First, both increased cognitive load and reduced word predictability, that is, higher word surprisal, should be reflected by longer reading times (Fig. 1a and b), alongside a decrease in comprehension question and n-back task performance. Second, we expected that the formation of language predictions should be contingent upon the availability of executive resources. Accordingly, a gradual limitation of these resources should result in an impairment of the formation of language predictions (Fig. 1c). We additionally investigated whether increased cognitive load differentially affected the sensitivity of predictions across different context lengths (i.e. timescales). Third, we explored how these previously described effects would be modulated by age: Expanding upon our hypothesis that language predictions should be impaired if executive resources are limited, we hypothesised that with increasing age and thus decreasing executive resources (Idowu & Szameitat, 2023; Salthouse, Atkinson & Berish, 2003), individuals should rely less on language predictions. This should be reflected by diminished surprisal effects (Fig. 1d, bottom). However, given the presumed stability of language comprehension across the lifespan (Shafto & Tyler, 2014), we also considered the opposite case, namely, that older adults might compensate for their sensory and

executive decline (Idowu & Szameitat, 2023; Salthouse et al., 2003; Verhaeghen et al., 2003) by relying more heavily on language predictions. In this case, we should see strong surprisal independent of age (Fig. 1d, top).

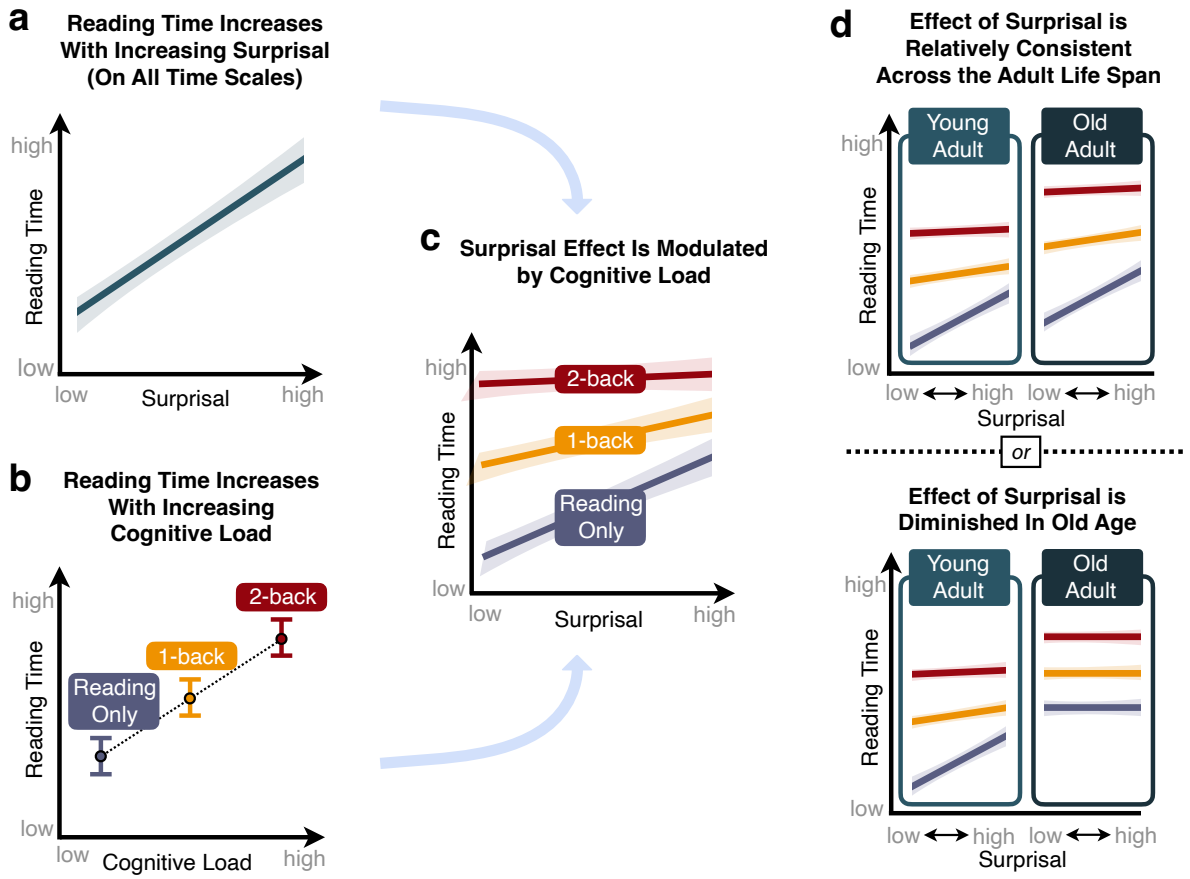

Figure 1. Visualisation of hypotheses. We expected main effects on reading time of (a) surprisal and (b) cognitive load, as well as (c) an interaction of surprisal and cognitive load). Additionally, (d) we explored how these effects are modulated by age. Here, two alternative scenarios seemed equally plausible a priori our study: The effect of surprisal could either be consistent across age groups (d, top panel; no moderation by age), or language prediction deteriorates with declining executive resources in old age, resulting in the absence of a surprisal effect in older age groups (**d**, bottom panel).

MATERIALS & METHODS

Participants

We recorded data from 178 participants, who were either tested online (*N* = 83) or in a controlled lab environment (*N* = 95). We excluded data from three participants from the online sample from further analysis, either due to technical issues (*N* = 1) or because they reported having been distracted during the task (*N* = 2). The resulting final sample comprised 175 participants aged 18 to 85 years (M = 26.943 ± 16.430 years) with a balanced gender distribution of 50.9% female, 47.4% male, and 1.7% non-binary identifying participants. All participants were native German speakers with normal or corrected-to-normal vision, and intact colour vision (assessed by a screening test). Exclusion criteria were dyslexia, illiteracy, a history of psychiatric or neurological disorders or drug abuse. Individuals who had consumed drugs or alcohol immediately prior to the study were not eligible for participation. All participants from the online sample enlisted through the recruitment platform Prolific, whereas lab participants were recruited via an existing database of the Max-Planck-Institute. Lab participants above the age of 40 performed the Mini-Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 1975) to screen for cognitive impairments. Middle-aged participants (40-59 years) had a mean score of 29.57 ± 0.84, whereas older participants (60-85 years) scored 28.23 ± 1.62 points.

Participants provided their informed consent prior to participation, and received financial compensation (12€/h). The study was conducted in accordance with the Declaration of Helsinki and was approved by the local ethics committees of the University of Lübeck and Leipzig University, respectively.

Study Design

During the experiment, participants were asked to read short newspaper articles which were adapted from articles from the news archive of a well-known German magazine ("*Der Spiegel*"). The nine selected articles were edited to be easy to understand, neutral in tone and non-emotional in content to avoid any influences of text content on reading time. Additionally, all texts were limited to a length of 300 words (trials). Participants were instructed to read the text carefully, as content-related multiple-choice questions were asked after each one. The comprehension questions served both as a measure of reading comprehension as well as a motivator to pay close attention to the reading material.

During this self-paced reading task (*Reading Only* task), each word was presented individually on screen and participants proceeded to the next word by pressing the space bar on their keyboard (see Fig. 2). To ensure that each word was displayed at least briefly, the response window started after the word had been shown for a fixed period of 50 ms. Each word was presented centred on screen in one of four font colours (Hex codes: #D292F3 [lilac], #F989A2 [muted pink], #2AB7EF [cerulean blue] and #88BA3F [leaf green]) against a white background.

In four of the six main blocks, the self-paced reading task was complemented by a secondary n-back task (see Fig. 2), in which participants were instructed to press a target button on the keyboard ("M" for right-handed, "C" for left-handed participants) whenever the font colour of the current word matched that of the previous (*1-back Dual Task*) or the penultimate word (*2-back Dual Task*). Participants were still required to press the spacebar to advance to the next word after pressing the target button. Reaction times were recorded for both kinds of responses. Neither the *Reading Only* task nor the *1-back* and *2-back Dual Task* blocks were speeded, allowing participants to complete the experiment at their own pace.

Before being introduced to this combination of reading and n-back task (see Fig. 2), participants could familiarise themselves with the n-back paradigm in short, non-linguistic, single-task blocks comprising coloured rectangles as stimuli. These blocks served as an introduction to the nature of an n-back task, and to quantify participants' working memory abilities. For all main blocks with an n-back task, a target ratio of 16.667% was used. The low target-ratio was chosen to prevent an excessive number of n-back reactions during the dual-task blocks.

Taken together, the experiment comprised three cognitive load conditions: A baseline condition, comprising a reading task without additional n-back task (*Reading Only*), a reading task with an additional 1-back task (*1-back Dual Task*), and a reading task with an additional 2-back task (*2-back Dual Task*). Each condition was presented in two blocks, each comprising 300 trials. For each block, one of nine texts was randomly selected, with no text occurring more than once.

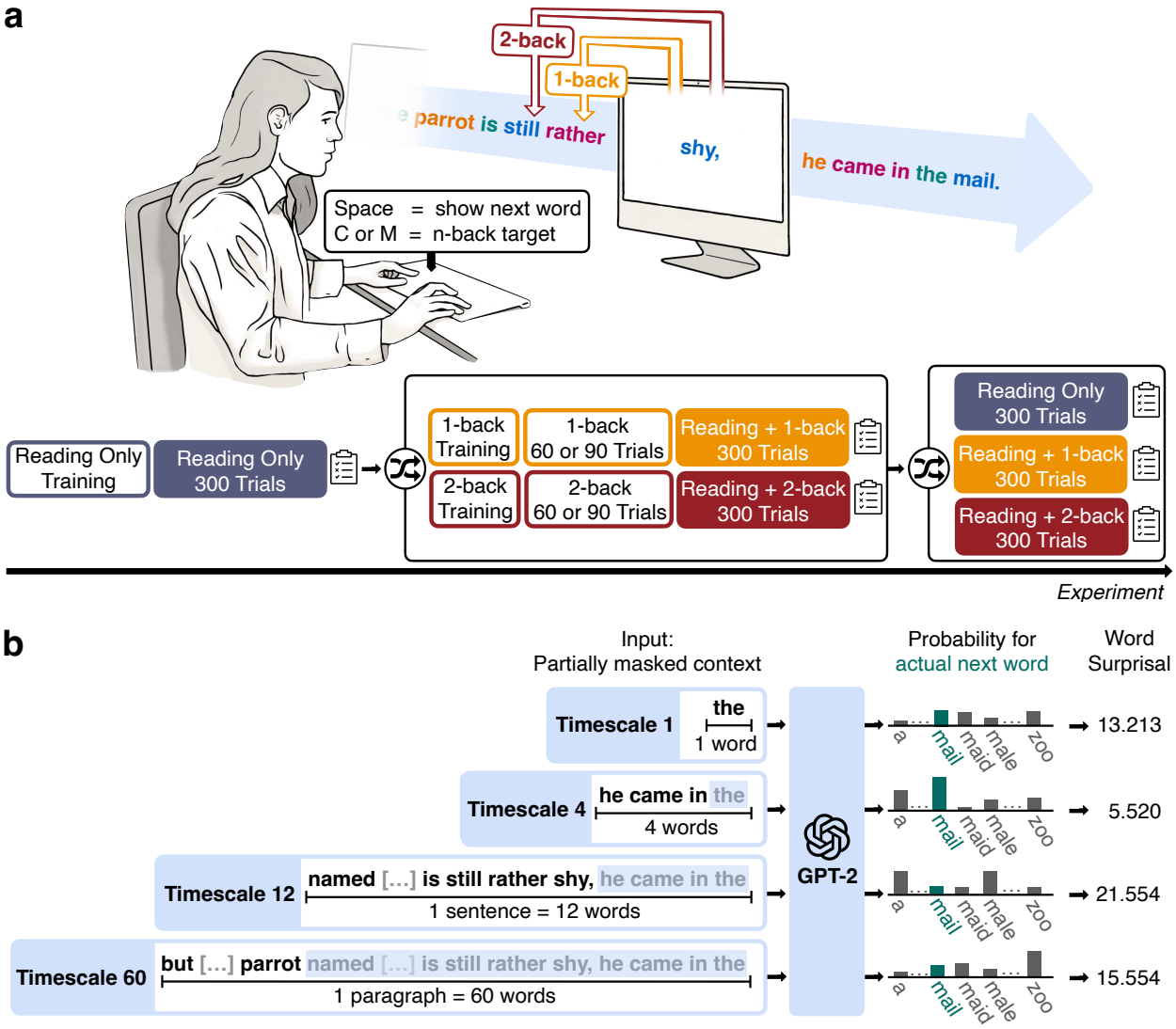

Figure 2. Experimental design and quantification of predictability as word surprisal using GPT-2. a. Participants were asked to perform a self-paced reading task (Reading Only) which was complemented in some blocks by a secondary n-back task on the font colour of the words (Reading + 1-back, and Reading + 2-back). The order of the blocks was pseudo-randomized, with Reading Only always being the first condition to be presented, followed by the two dual-task conditions, and another main block for each of the three conditions. Both dual-task paradigms (Reading + 1-back and Reading + 2-back) were first introduced in short single-task training sets. b. We generated four surprisal scores for each word in the reading material by using text chunks of different lengths (with n previous words set to 1, 4, 12, or 60 words) as prompts for next-word-predictions in GPT-2. To reduce redundancies between timescales, we masked all parts of the input chunks which were already covered by previous timescales. The resulting probabilities for the actual next word in the text were then transformed into surprisal scores, which reflected how predictable the respective word was on each of the four timescales. As each surprisal is based on a different context chunk, it is possible for one word to have four different surprisal scores. Please note that the example sentence used here has been translated to English for better comprehensibility, while the original text materials were in German.

Generation of Surprisal Scores

The predictability of each word was operationalised via word surprisal on different timescales. Word surprisal reflects the predictability of the current word given its preceding context (Hale, 2001; Levy, 2008; Smith and Levy, 2013) and is defined as the negative logarithm of the word's probability given its context (Hale, 2001):

$$surprisal_{w_n} = -\log(P(w_n | w_1, \dots, w_{n-1}))$$

If a word w_n has a high surprisal score, its occurrence given its preceding context w_1, w_2, …, w_{n-1} has a low probability, rendering it highly unpredictable (i.e. surprising). A word can be predicted on different timescales, defined by the length of the considered context (e.g. single-word, phrase, sentence, paragraph or entire-narration level). In the current study, we computed four surprisal scores for each word in the experimental texts, representing a distinct timescale each (1, 4, 12, and 60 words context, see Fig. 3). The selected timescales represent contextual information at the single-word, phrase, sentence, and paragraph level, as the average sentence length in our experimental texts was 12 words, and the shortest sentence contained 4 words. For paragraph-level analysis, we considered 60 words, which corresponds to the average paragraph length of 5 sentences observed in our texts. As the four timescales overlapped (see Fig. 2b), we masked all words in the context chunk that are shared with lower timescales to ensure their independence. The mean pairwise correlation of surprisal scores on the four different timescales was 0.658 ± 0.044 (see further information in the supplementary material), while surprisal scores increased slightly with increasing timescale (mean surprisal on timescale 1: 15.042 ± 6.045, prediction accuracy = 8.449 %; timescale 4: 15.673 ± 5.769, prediction accuracy = 9.728 %; timescale 12: 16.018 ± 6.042; prediction accuracy = 8.657 %; timescale 60: 16.884 ± 6.337, prediction accuracy = 9.663 %).

To generate surprisal scores, we used a 12-layered GPT-2 model (Radford et al., 2019), which was pre-trained on German texts by the MDZ Digital Library team (dbmdz) at the Bavarian State Library, and the corresponding tokenizer, both available from the Hugging Face model hub (Schweter, 2021). Surprisal scores were calculated using Python version 3.10.12 (Van Rossum & Drake, 2009).

Analysis

Preprocessing

To gauge participants' response accuracy in the n-back task, we computed the detection-prime (d') index. This measure quantifies the ability to distinguish between target and non-target stimuli, in our case trials (i.e. words in dual- and rectangles in single-task blocks) with colour repetitions and trials where the current colour doesn't match the colour from the *n*th previous trial. A d' value of 0 signifies an inability to discriminate between signal and noise stimuli, suggesting that participants indicated they saw a target in either no or all trials. Thus, we excluded all dual-task blocks with *d*-primes of 0 from further analyses, which affected only five participants. In total, we excluded three main blocks in the online sample and two main blocks in the lab sample.

After each text block, participants were asked to answer three multiple-choice questions as a measure of reading comprehension. For each question, we showed four response options, with only one of them being correct. To ensure participants performed the tasks as intended and read the words on screen, we excluded all blocks where none of the questions were answered correctly. In total, we excluded one *1-back* and seven *2-back Dual Task* blocks from the online sample, as well as four *1-back* and eleven *2-back Dual Task* blocks from the lab sample, from datasets of six and fourteen participants, respectively.

Lastly, we preprocessed the reading time data: First, any trial exhibiting a raw reading time exceeding 5000 ms was considered an extended break and subsequently excluded from further analyses. To further remove outliers, we followed the procedure recommended by Berger and Kiefer (2021) to ensure comparable exclusion criteria for long and short outliers in typically skewed reading times: First, reading times were transformed on block-level using the *Proportion of Maximum Scaling* method (POMS; Little, 2013). The square root of each value was then taken to ensure a symmetric distribution. We transformed the data on block level to account for potential differences in reading time distributions between blocks. Following this, we *z*-transformed the data and excluded all trials from further analysis where *z*-scores fell outside a range of -2 to 2 (Berger & Kiefer, 2021; Cousineau & Chartier, 2010). Taken together, we excluded 12968 trials (4.117% of all trials from the main blocks) with an average of 74.103 ± 12.654 excluded trials per participant.

To facilitate interpretability of units in the results, we subsequently continued working with the raw reading times, which had been cleaned of outliers at this stage, and log-transformed them for statistical analysis.

Statistical Analysis of n-Back Responses and Comprehension Questions

To ensure the validity of our cognitive load manipulation in the dual-task blocks, we examined whether increased cognitive load induced a decline in performance in both the n-back task – as indicated by reduced *d*-primes – and accuracy in the comprehension question task, as reflected by a lower number of correct answers. We employed a linear mixed-effects model (LMM) for d-primes and a logistic linear mixed-effects model (GLMM, logit link function) for comprehension question accuracy.

In both models, we included recording location (online or lab), cognitive load and continuously measured age (centred) in both models as well as the interaction of age and cognitive load as fixed effects. In the model for the *d*-primes, we additionally included measures of comprehension question accuracy (on participant and block level) as well as the block number as fixed effects to control for different response strategies and tiredness effects, respectively. Moreover, we included the mean d-primes from the *1-back* and *2-back Single Tasks* as a working memory measure.

We assigned simple coding schemes to the factors recording location and cognitive load. While the model for d-primes included by-participant random slopes for cognitive load, the model for comprehension question accuracy comprised random intercepts for participants. Both models included random intercepts for participants and texts.

d-prime ~ mean d-prime from single tasks +
mean comprehension question accuracy +
block-level deviation from mean comprehension question accuracy +
recording location + block number +
*age * cognitive load +*
(1 + cognitive load | ID) +
(1 | text number)

Note. Structure of the model for *d*-primes in the n-back task in dual-task blocks. The variable age was centred and the variable cognitive load encompassed only two levels (1-back and 2-back) as there was no n-back task in the Reading Only condition.

comprehension question accuracy ~ recording location +
*age * cognitive load +*
(1 | ID) + (1 | text number)

Note. Structure of the model for accuracy in the comprehension question task. The variable age was centred. We used a binomial family distribution with a logit link function for modelling the comprehension question accuracies.

Statistical Analysis of Reading Times

We explored the effects of cognitive load, age and surprisal on four timescales (timescales 1, 4, 12 and 60) as well as their 2-way and 3-way interactions on log-transformed reading times. All models included an interaction of age, surprisal, and cognitive load motivated by our hypotheses as well as additional fixed effects to control for nuisance effects. The final selection of fixed and random effects structure was based on highest R^2 values.

To model the influence of surprisal on reading times while addressing potential multicollinearity issues stemming from the high correlation between surprisal scores on the different timescales, we used an individual model for each of the four timescales. The four models deviated in structure only in the way we included the effects of surprisal (see example for model formula below), while all other fixed effects were kept constant across models. In all four models, we included the reading time of the previous word as a fixed effect to control for potential nuisance effects such as post-error-slowing following a missed n-back target in the previous trial, or sequential modulation effects if the previous trial was ended prematurely, leading to an extended reading time carried over to the current trial. As we only modelled reading times from trials where we had surprisal scores on all four timescales the first 60 trials of each block were not included in the statistical analyses.

As response strategies may differ between individuals, but also within an individual from block to block, we included two different regressors representing these distinct between- vs. within-participant effects on reading time. Between-participant effects were modelled by the individual mean comprehension performance whereas within-participant effects were modelled by the block-level deviation from this mean (cf. Bell et al., 2018; Tune et al., 2021). We further included block-wise d-primes and participant-wise mean single-task d-primes as a proxy of each participant's working memory capacity. By incorporating block- and participant level performance measures, which are designed to be sensitive to task difficulty, we accounted for the potential variation in perceived task load between age groups or samples. For instance, the 2-back task might present a greater challenge for an older individual compared to a younger counterpart, therefore rendering the tasks not entirely comparable between age groups if not appropriately controlled for.

The remaining fixed effects entailed the recording location (online or lab), word frequency (as estimated using Python's *wordfreq* package; Speer, et al., 2018), word length, the occurrence of n-back reactions, as well as the position of block and trial in the course of the experiment as main effects.

Finally, we included the three-way interaction of age (as a continuous predictor), cognitive load (on 3 levels: *Reading Only*, *1-back Dual Task* and *2-back Dual Task*; contrasted using a simple coding scheme with the *Reading Only* condition as the reference level), and surprisal score (continuous predictor). This entails the implicit inclusion of all two-way interactions of age, cognitive load and surprisal, as well as the main effects of those variables. To ensure independence of timescales, we preserved the order of surprisal scores on one timescale while randomising the surprisal scores on all others, resulting in four different models.
Random effects included random intercepts for participants, the effect of text, the current word, the colour of the current word, and by-participant random slopes for cognitive load. All continuous predictors were centred.

log(RT) ~ RT of previous word +
block-level d-prime + mean d-prime from single tasks +
mean comprehension question performance +
block-level deviation from mean comprehension question performance +
recording location +
word frequency + word length (without punctuation) +
n-back reaction + block number + trial number +
surprisal TS 4 shuffled + surprisal TS 12 shuffled + surprisal TS 60 shuffled +
surprisal TS 1 * age * cognitive load +
(1 + cognitive load | ID) + (1 | text number) + (1 | word) + (1 | colour)

Note. Example for the structure of the model for timescale 1. Equivalent models for timescales 4, 12 and 60 were constructed by replacing surprisal on timescale 1 (in this model) with surprisal on the respective timescale (4, 12 or 16). Additionally, we included surprisal on the remaining timescales (i.e. timescales 1, 4 and 60 in the model for timescale 12) in a shuffled order as main effects.
RT = Reading Time, ID = participant.

To gain a more nuanced understanding of the three-way interaction of age, cognitive load and surprisal, we performed a subsequent simple slopes analysis for each of the four models. This analysis allows exploring the interaction of two continuous predictors, in our case quantifying the slope of the surprisal effect on each timescale in each of the three cognitive load conditions as a function of age. This way, we determined for which age range and cognitive load condition we observed a significant effect of surprisal on reading time.

For all models, *p*-values were obtained using ANOVAs with type III sums of squares. Degrees of freedom for *p*-values and standard errors were estimated using Satterthwaite's

approximation for all LMMs, and Wald's approximation for the GLMM (Luke, 2017; Satterthwaite, 1946).

All analyses were carried out in *R* version 4.2.2 (R Core Team, 2023) using the packages *interactions, lmerTest, lme4 and sjPlot* (Bates et al., 2015; Kuznetsova et al., 2017; Long, 2019; Lüdecke, 2023).

RESULTS

In the following section, all effects reported are significant on an alpha level of 0.05 after FDR-correction for multiple comparisons (Benjamini & Hochberg, 1995). For full details of model outputs see Table 1 and Table S1 in the Supplementary Material.

Increased Cognitive Load and Older Age Reduce Task Performance

As expected, both increased cognitive load and advancing age were associated with a significant decrease in comprehension question accuracy (cognitive load: $OR_{1\text{-}back} = 0.253$, $t = -8.928$, $p_{1\text{-}back} = < 0.001$, $OR_{2\text{-}back} = 0.156$, $t = -12.403$, $p_{2\text{-}back} = < 0.001$; age: $OR = 0.956$, $t = -2.676$, $p = 0.013$; see Fig. 3a and Table 1), suggesting that participants were less likely to answer comprehension questions correctly if they were older or they had to engage in a secondary task while reading the associated text.

For task performance in the n-back task, we observed a significant interaction of age and cognitive load ($\beta = -0.014$, $t = -3.961$, $p = < 0.001$): While participants showed a decrease in performance with advancing age in the *2-back Dual Task*, this was not the case in the easier *1-back Dual Task* condition (see Fig. 3b and Table 1), suggesting that older and younger participants only showed differences in performance in the n-back task if task demands were high.

Table 1

Results from models for task performance measures

	LMM for d-primes					GLMM for comprehension question accuracy				
	Estimate	**Std. Error**	***t***	***p***		**OR**	**Std. Error**	***t***	***p***	
mean d-prime single tasks	0.475	0.0561	8.459	**2.227×10^{-14}**	*					
mean comprehension question performance	0.012	0.0037	3.203	**1.829×10^{-3}**	*					
de-meaned comprehension question performance	-0.001	0.0001	-15.441	**2.848×10^{-53}**	*					
block number	-0.005	0.0004	-12.992	**3.214×10^{-38}**	*					
recording location [online]	-0.52	0.0928	-5.651	**1.008×10^{-7}**	*	0.980	0.1876	-0.106	9.155×10^{-1}	
age	-0.005	0.0027	-1.922	5.630×10^{-2}		0.986	0.0053	-2.676	**1.304×10^{-2}**	*
cognitive load [1-back vs. Reading Only]						0.253	0.0390	-8.928	**1.011×10^{-18}**	*
cognitive load [2-back vs. Reading Only]						0.156	0.0234	-12.403	**1.753×10^{-34}**	*
cognitive load [2-back vs. 1-back]	-1.64	0.0643	-25.462	**2.764×10^{-59}**	*					
age * cognitive load [1-back vs. Reading Only]						0.990	0.0081	-1.183	3.313×10^{-1}	
age * cognitive load [2-back vs. Reading Only]						1.003	0.0080	0.395	8.081×10^{-1}	
age * cognitive load [2-back vs. 1-back]	-0.014	0.0036	-3.961	**1.409×10^{-4}**	*					
	Conditional / Marginal R^2			**ICC**		**Conditional / Marginal R^2**			**ICC**	
model fit	0.922 / 0.624			0.792		0.304 / 0.146			0.185	

Note. All continuous predictors were centred. Degrees of freedom for *p*-values, standard errors and confidence intervals (CI) were computed using Satterthwaite's approximation (LMM for d-primes) and Wald's approximation (GLMM for comprehension question accuracy). All *p*-values reported here are FDR-corrected and were computed using ANOVAs with type III sums of squares. Results that are significant on an alpha-level of 0.05 are marked with a star. OR = Odds Ratio.

Reading Time Increases with Increased Age, Surprisal, and Cognitive Load

As hypothesised, in all models we observed significantly longer reading times with advancing age, independent of cognitive load condition (model for timescale 1: $\beta = 0.009$, $t = 9.033$, $p = < 0.001$; model for timescale 4: $\beta = 0.014$, $t = 11.511$, $p = < 0.001$; model for timescale 12: $\beta = 0.014$, $t = 11.511$, $p = < 0.001$; model for timescale 60: $\beta = 0.009$, $t = 9.032$, $p = < 0.001$). For instance, in the model for timescale 1, an increase of one year in age was associated with an increase in reading time by 1.009%, which corresponds to 7.547 ms (see Fig. 3c, first panel).

Likewise, increased cognitive load was associated with significantly longer reading times (model for timescale 1: $\beta_{1\text{-back}} = 0.490$, $t_{1\text{-back}} = 33.0505$, $p_{1\text{-back}} = < 0.001$, $\beta_{2\text{-back}} = 0.812$, $t_{2\text{-back}} = 29.765$, $p_{2\text{-back}} = < 0.001$; model for timescale 4: $\beta_{1\text{-back}} = 0.493$, $t_{1\text{-back}} = 33.432$, $p_{1\text{-back}} = < 0.001$, $\beta_{2\text{-back}} = 0.812$, $t_{2\text{-back}} = 29.494$, $p_{2\text{-back}} = < 0.001$; model for timescale 12: $\beta_{1\text{-back}} = 0.493$, $t_{1\text{-back}} = 33.428$, $p_{1\text{-back}} = < 0.001$,

$\beta_{2\text{-}back}$ = 0.812, $t_{2\text{-}back}$ = 29.491, $p_{2\text{-}back}$ = < *0.001*; model for timescale 60: $\beta_{1\text{-}back}$ = 0.490, $t_{1\text{-}back}$ = 33.501, $p_{1\text{-}back}$ = < *0.001,* $\beta_{2\text{-}back}$ = 0.812, $t_{2\text{-}back}$ = 29.768, $p_{2\text{-}back}$ = < *0.001*; see Fig. 3c, second panel), indicating participants read more slowly when faced with a more challenging task.

Moreover, in line with our hypothesis, we observed a consistent increase in reading times with higher surprisal in all models (model for timescale 1: β = 0.0006, t = 5.385, p = < 0.001; model for timescale 4: β = 0.0006, t = 4.785, p = < 0.001; model for timescale 12: β = 0.0005, t = 3.887, p = < 0.001; model for timescale 60: β = 0.0004, t = 3.263, p = 0.001; see Fig. 3c, third panel). This suggests that highly predictable words were read more quickly than unpredictable words, regardless whether word surprisal was based solely on the very last word(s) (surprisal on timescales 1 and 4) or on a broader context (surprisal on timescales 12 and 60).

We observed these effects of age, cognitive load and surprisal after accounting for the effects of word length, word frequency, task performance and position of the respective trial in the course of the experiment.

Cognitive Load Modulates the Effect of Surprisal

In line with our hypotheses, cognitive load significantly modulated the effect of surprisal across timescales (model for timescale 1: $\beta_{1\text{-}back}$ = -0.0009, $t_{1\text{-}back}$ = -3.969, $p_{1\text{-back}}$ = < 0.001, $\beta_{2\text{-}back}$ = -0.0001, $t_{2\text{-}back}$ = -5.443, $p_{2\text{-back}}$ = < 0.001; model for timescale 4: $\beta_{1\text{-}back}$ = -0.001, $t_{1\text{-}back}$ = -4,616, $p_{1\text{-}back}$ = < 0.001, $\beta_{2\text{-}back}$ = -0.001, $t_{2\text{-back}}$ = -5.501, $p_{2\text{-}back}$ = < 0.001; model for timescale 12: $\beta_{1\text{-}back}$ = -0.001, $t_{1\text{-}back}$ = -4.505, $p_{1\text{-}back}$ = < 0.001, $\beta_{2\text{-}back}$ = -0.001, $t_{2\text{-}back}$ = -5.619, $p_{2\text{-}back}$ = < 0.001; model for timescale 60: $\beta_{1\text{-}back}$ = -0.0004, $t_{1\text{-}back}$ = -2.122, $p_{1\text{-}back}$ = 0.034, $\beta_{2\text{-}back}$ = -0.0007, $t_{2\text{-}back}$ = -3.591, $p_{2\text{-}back}$ = < 0.001). While we observed a clear increase in reading time in the Reading Only condition when surprisal was high, this effect was mitigated in the 1-back dual-task condition, and essentially absent in the 2-back dual-task condition, where cognitive load was highest (see Fig. 3d).

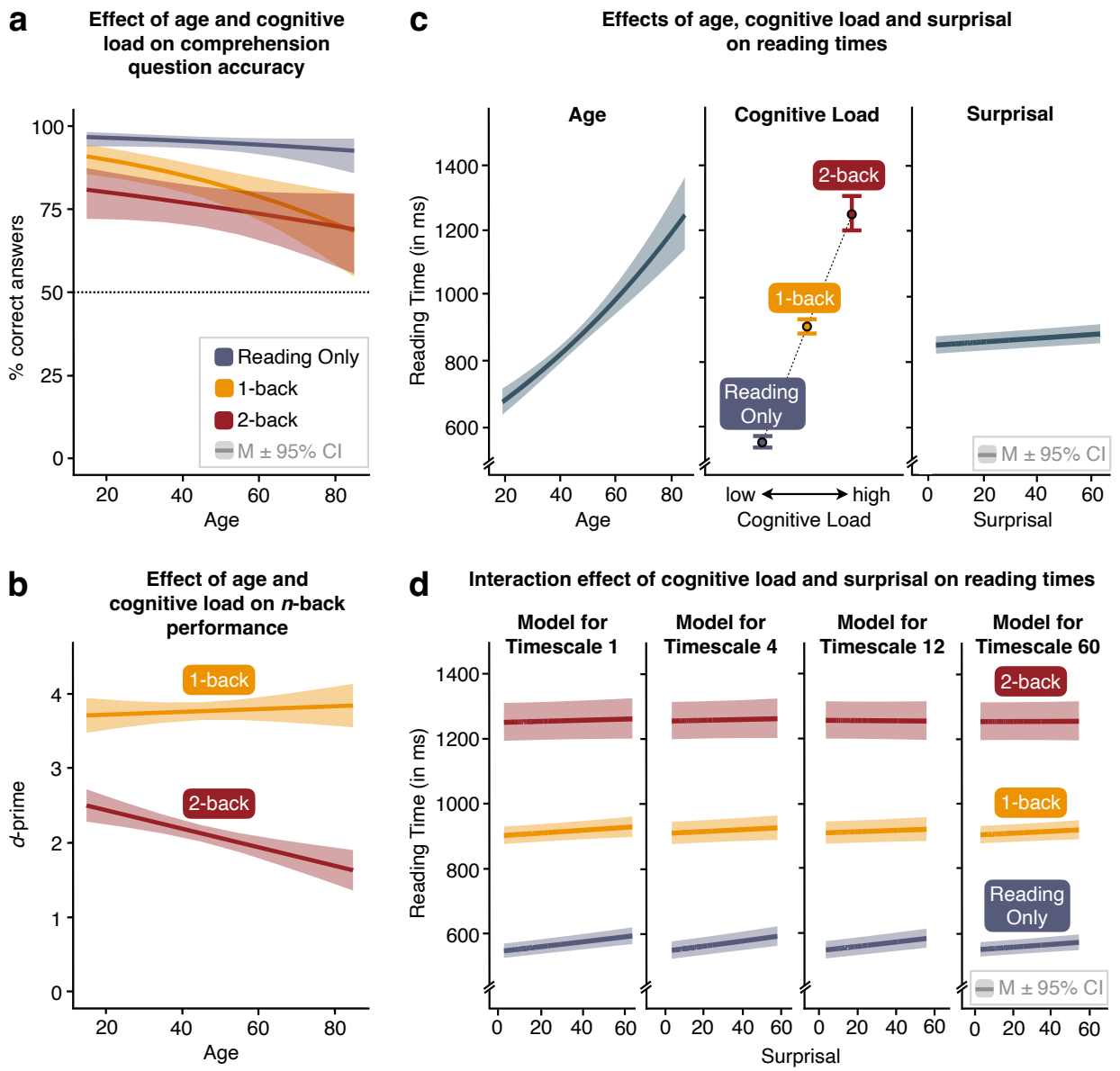

Figure 3. Marginal effects of predictors age, cognitive load and surprisal on task performance and reading time. Main effects of cognitive load and age on accuracy in the comprehension question task (panel a) and on n-back task performance (d-primes; panel b). Please note that we do not show *d*-primes for the Reading Only task as there was no n-back task in this condition. Reading time increased with increasing age, cognitive load and word surprisal across age groups (panel c), with cognitive load modulating the effect of surprisal on reading times on all timescales (panel d). Effects presented in panel c are based on results from the model for timescale 1. Please see table TS1 in the supplements for main effects in the models for timescales 4, 12 and 60.

Interaction of Surprisal and Cognitive Load is Modulated by Age

Finally, we tested whether the interaction between cognitive load and surprisal changes with advancing age: Across models, we found that age indeed modulated the influence of cognitive load on the effect of surprisal on reading times, as reflected by significant three-way interactions of surprisal, cognitive load and age.

To get a more nuanced understanding of this effect, we conducted simple slopes analyses (see Fig. 4): We found that in conditions with low cognitive load (condition *Reading Only*), surprisal significantly influenced reading times in all but the youngest participants. Put simply, when participants read a text without having to perform an additional n-back task, predictable words (i.e., low surprisal) yielded shorter reading times (see Fig. 4, first row). However, this effect narrowed to older adults (aged 60 and above) at shorter timescales when cognitive load was increased (*2-back Dual Task*; see Fig. 4, third row). Thus, when task demands were high, reading time was only significantly modulated by surprisal in adults aged 60 and over, whereas younger adults showed no significant effect of surprisal. Taken together, while all participants exhibited faster reading times during simple reading when surprisal was low, only older adults leveraged low surprisal to enhance their reading speed when task demands were high.

In the condition with intermediate cognitive load (*1-back Dual Task*), the pattern was less clear, with standard errors for the effects being relatively larger.

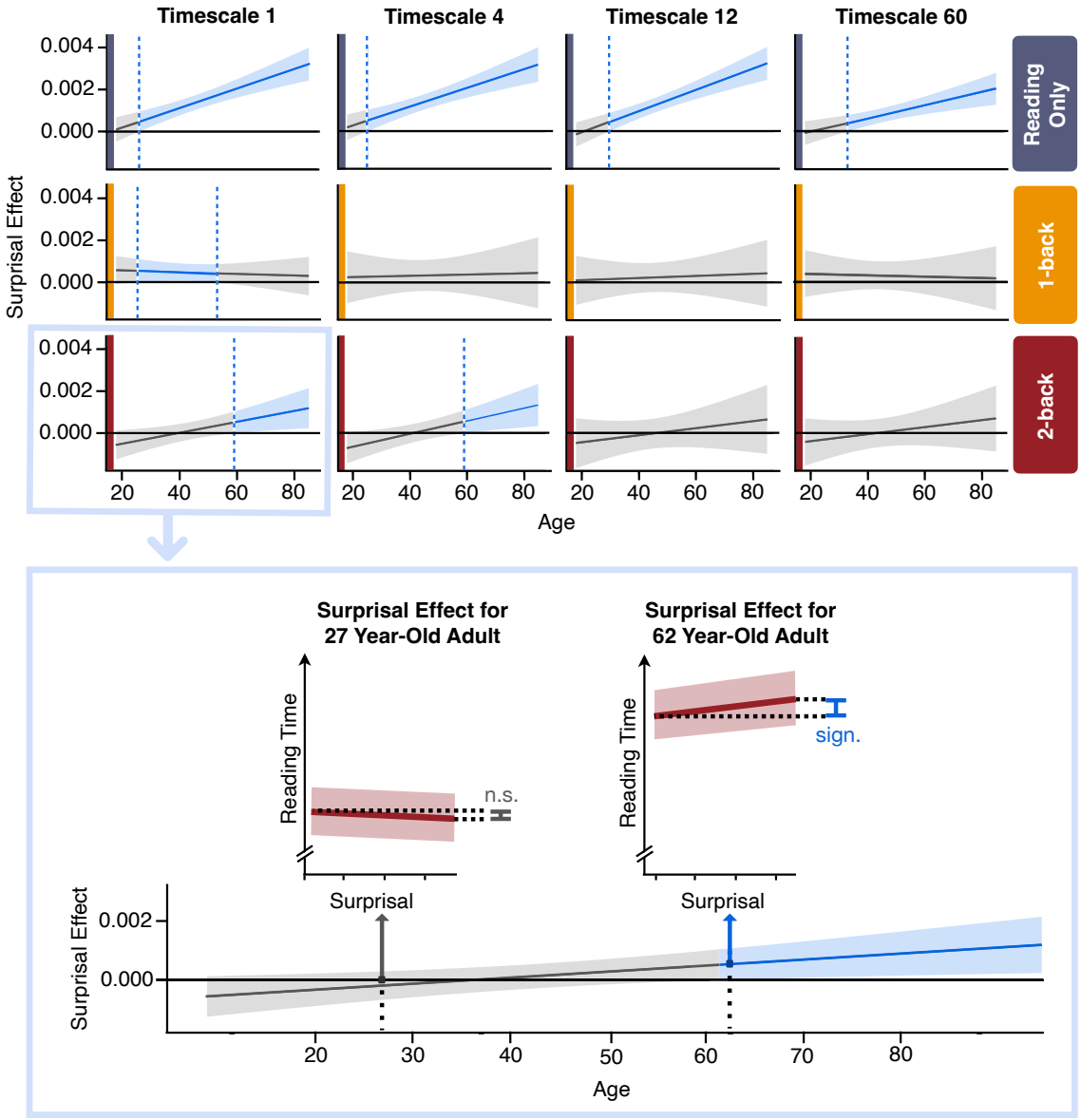

Figure 4. Results of the simple slopes analysis. In each column, we show the effect of surprisal on reading time across the whole age range in the conditions Reading Only (first row), 1-back Dual Task (second row) and 2-back Dual Task (third row). The stronger the surprisal effect for a certain age, the higher the value on the y-axis. Grey areas indicate age ranges for which we did not find an effect of surprisal on the respective time scale on reading time in the respective condition, whereas blue areas indicate we observed a significant surprisal effect (see inset at the bottom for a visualisation for how the simple slopes plots relate to marginal effects plots).

DISCUSSION

Do linguistic predictions during language comprehension draw on executive resources? With the present study, we asked how language predictions on four different timescales, ranging from single word to paragraph level, change with increasing cognitive load, and how this interaction is modulated by age. To do so, we paired a self-paced reading task with a secondary n-back task on the font colour of the words.

First, as expected, and validating our overall approach, high cognitive load was associated with an increase in reading times as well as a decrease in task performance across age groups. This is consistent with previous studies using n-back tasks (Lamichhane et al., 2020; Kwong See & Ryan, 1995).

Next, as hypothesised, higher word surprisal slowed reading down, even when controlling for word length and frequency. This held true irrespective of timescale on which word surprisal was being quantified. This effect of surprisal again is in line with previous studies, as more predictable words are generally read faster (Wilcox et al., 2023).

Finally, we explored the relationship between surprisal and cognitive load across the (cross-sectional) age span. We hypothesised that increasing cognitive load should gradually impair language prediction, which should be reflected by a diminished surprisal effect in conditions with high cognitive load. In line with our hypothesis, we found that across timescales the effect of word surprisal on reading times was modulated by cognitive load. Specifically, when cognitive load was high, the effect of surprisal on reading time was essentially absent.

Interestingly, this interaction between surprisal and cognitive load was significantly modulated by age: In contrast to younger adults, we found robust surprisal effects in older adults aged 60 and over even under conditions with high cognitive load.

Interpreting these effects of predictability is not quite as straightforward as it might initially appear. In general, generating language predictions can accelerate linguistic processing in cases where those predictions are accurate: Rather than processing each incoming linguistic stimulus throughout the entire processing hierarchy, features of upcoming stimuli can be activated prior to their occurrence. If a prediction aligns with the actual stimulus, processing can be facilitated, resulting in faster response times. However, if the prediction is erroneous, the subsequent prediction error has to be processed to inform upcoming predictions. This leads to slower response times compared to cases where predictions prove to be accurate (Goodkind & Bicknell,

2016; Wilcox et al., 2023; for reviews see Kuperberg & Jaeger, 2016; Ryskin and Nieuwland, 2023).

High language predictability does not exclude the occurrence of prediction errors altogether. Rather, such errors should be overall less frequent under high compared to low predictability. Consequently, on average, less processing time should be dedicated to integrating prediction errors if language predictability is high, resulting in lower average reading times. However, it is still unclear how the impact of prediction errors relates to processing where no language predictions are generated at all, and upcoming stimuli have to be processed across the entire processing hierarchy upon encountering them. This could answer the question if language predictions are always beneficial for reading time, just less beneficial when prediction errors occur, or if they are only beneficial if they prove to be accurate.

Taken together, our findings bear out the following interpretation: High predictability has a beneficial effect on reading time, whereas low predictability has a detrimental effect on reading time if compared to highly predictable words. In our study this was true across timescales.

In contrast to previous work on the influence of cognitive load on language comprehension (e.g. Fedorenko et al., 2006; Gordon et al., 2002; Rodd et al., 2010), we employed a *non-linguistic* secondary task to manipulate cognitive load, thereby aiming to engage domain-general executive resources. When investigating the interaction of cognitive load and language predictability on reading time, we found that increased cognitive load diminished the effect of predictability on all timescales, suggesting executive resources are involved in the generation of language prediction across timescales.

Predictive processing has not only been suggested to the foundation of language processing (e.g. Heilbron, 2022; Schmitt et al., 2021; Ryskin & Nieuwland, 2023), but it is also thought to be a core mechanism of the human brain (Bubic et al., 2010). Drawing parallels between domains can thus offer valuable insights into the mechanisms of language prediction: For instance, there is evidence from the visual domain showing that attending to the stimulus material is a prerequisite for predictive processing (Larsson & Smith, 2012; Richter & de Lange, 2019). This observation from the visual domain can potentially be extended to our findings, wherein language predictions were diminished if attention had to be divided between the reading material and a challenging non-linguistic secondary task.

However, attention alone cannot fully account for the differences in sensitivity to word predictability shown here, as we observed the predictability effect did not only diminish if a

secondary task was introduced, but also decreased with increasing cognitive load. In line with previous literature (e.g., Huettig & Janse, 2016; Ito et al., 2017; Fricke & Zirnstein, 2022), our results suggest that executive processes beyond attention, such as updating and maintenance of context information, shifting between tasks, and inhibition of irrelevant information (Friedman & Miyake, 2017), are integral to language prediction.

Further examining the interaction of cognitive load and language predictability across the adult lifespan revealed interesting differences: While older participants showed strong predictability effects even under high task demands, the effect vanished in the younger cohort despite their intact text comprehension. Our result aligns with previous work on age-related changes in linguistic predictions, indicating heightened sensitivity to unexpected lexical input in older adults (Cheimariou et al., 2019). This finding may reflect a commonly reported pattern of greater reliance on intact vocabulary and world knowledge with age in the face of declining executive functions (Martin et al., 2022; Salthouse et al., 2003). Here, we show that this is particularly the case when cognitive demands are high, which makes older adults rely more heavily on their refined system of linguistic predictions driven by their lifelong experience. This allows them to make more fine-tuned predictions, but also makes them more vulnerable to unexpected information. Our finding of age-related differences also raises the question whether language prediction is a default strategy for older, but not (yet) for younger adults.

Previous studies have found a larger vocabulary size associated with more rapid processing of language and improved language comprehension (Borovsky et al., 2012; Huettig & Pickering, 2019; Mainz et al., 2017; Matthews, 2018; Stæhr, 2008 & 2009). In line with this, older adults exhibit more advanced language processing abilities compared to children or younger adults due to their accumulated years of exposure to language and their increased vocabulary (Brysbaert et al., 2016; Ito & Sakai, 2021). This accumulated skill is thought to serve as a compensatory mechanism for any decline in working memory capacities or reduced executive functioning that they may experience with age (Bunzeck, et al., 2024; Reuter-Lorenz et al., 2021). Indeed, speech comprehension appears to remain largely intact in older adults (Burke, & Mackay, 1997; Shafto & Tyler, 2014).

Note, however, that the absence of a predictability effect in younger adults on the behavioural level does not necessarily imply a cessation of language prediction formation in younger adults when executive resources are limited. Conceivably, in our study, younger participants could have processed the n-back task and linguistic stimuli in parallel, potentially

masking the predictability effect due to the prolonged processing times in the n-back task. In contrast, older participants may have processed the *tasks* sequentially, leading to an additive effect on response times.

This question of the formation of language predictions in younger adults under increased task demands thus necessitates further investigation on the neural level. For instance, observing diminished neural markers of prediction – such as pre-stimulus alpha or beta desynchronisation in EEG or MEG (Bastos et al., 2012; Gastaldon et al., 2020; Rommers et al., 2017; Sedley et al., 2016; Wang et al., 2018; Terporten et al., 2019) – in conditions with increasing task load would suggest that younger individuals no longer form language predictions. Observing no influence of cognitive load on the presence of these markers, however, would suggest younger individuals still form language predictions even under increased task demands.

Additionally, it might be worth considering inter-individual differences regarding the *utilisation* of formed predictions: While in this study we assumed that formed predictions are utilised automatically, and that the observed differences in the effect of language predictability on reading times between age groups might thus be attributable to a lack of formation of predictions in younger participants, one could also argue that individuals might simply weigh their formed predictions differently.

Akin to the longer exposure to language in older individuals, one could hypothesise that older individuals have also had more time to accumulate experience regarding the *accuracy* of their predictions, and to refine their predictions through prediction errors. Consequently, older adults might rely more heavily on their predictions than younger adults (Chan et al., 2021; Moran et al., 2014). Additionally, due to age-related sensory decline (Verhaeghen et al., 2003), older adults might exhibit a stronger dependence on context-based predictions to process language, as incoming sensory information might be less informative (Rogers, 2017; Wingfield et al., 2015; Wolpe et al., 2016). A stronger reliance on language predictions could thus serve as a compensatory mechanism to facilitate language comprehension despite sensory decline in older adults.

CONCLUSION

In summary, the present study offers the following key insights: First, high language predictability as well as low task demands both have a beneficial effect on reading time. This held true across different timescales of predictability and a large age range. Second, we find that across

timescales, higher task demands diminish the positive effects of language predictability on reading time, suggesting that the formation of language predictions draws on resources of executive control. Third, we show that a limitation of domain-general executive resources impairs language prediction across timescales uniformly, indicating that processing costs for the formation of language predictions are relatively consistent across different timescales. Fourth, the data reveal age-related differences in the use of linguistic predictions: High predictability had a stronger and more robust leverage on reading efficiency in older than in younger adults, particularly under high cognitive load. These results spell out the dynamic interplay between language predictability and executive resources.

Acknowledgements

We thank Christian Koblitz and Marcel Blumenthal for their help with data acquisition in the lab.

Funding

This work was supported by German Research Foundation (DFG, OB 352/2-2 to JO, and HA 6314/4-2 to GH). GH was supported by Lise Meitner Excellence funding from the Max Planck Society and the European Research Council (ERC-2021-COG 101043747).

Author contributions

Author contributions are described using CRediT (Contributor Roles Taxonomy: https://credit.niso.org/).

Conceptualisation: M.S., S.M., G.H., J.O.; data curation: M.S., S.M.; formal analysis: M.S., S.M., S.T.; investigation: M.S., S.M.; methodology: M.S., S.M., S.T., L.-M.S., G.H., J.O.; project administration: M.S., S.M., visualisation: M.S.; writing - original draft: M.S., S.M., writing - review and editing: M.S., S.M., S.T., L.-M.S., G.H., J.O.; funding acquisition: G.H., J.O., supervision: G.H., J.O.

Declaration of AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT 3.5 in order to rephrase sentences. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Competing interests

The authors declare that they have no competing interests.

Data and code availability

All experimental and analysis scripts are publicly available on Github under the following link: https://github.com/MMarieSchuckart/EXNAT-1

The data will be published on OSF upon publication.

Preregistration

The present study was not preregistered.

References

Bamberger, R. & Vanecek, E. (1984). Lesen-Verstehen-Lernen-Schreiben. Die Schwierigkeitsstufen von Texten in deutscher Sprache. Wien: Jugend und Volk.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical Microcircuits for Predictive Coding. *Neuron, 76*(4), 695–711. doi: 10.1016/j.neuron.2012.10.038

Bell, A., Fairbrother, M. & Jones, K. (2018). Fixed and random effects models: making an informed choice. *Quality & Quantity*, *53,* 1051–1074. doi: 10.1007/s11135-018-0802-x

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. doi: 10.2307/2346101

Berger, A., & Kiefer, M. (2021). Comparison of Different Response Time Outlier Exclusion Methods: A Simulation Study. *Frontiers in psychology*, *12*, 675558. doi: 10.3389/fpsyg.2021.675558

Borovsky, A., Elman, J. L., Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology, 112*(4), 417–436. doi: 10.1016/j.jecp.2012.01.005

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. *Frontiers in Psychology, 7*. doi: 10.3389/fpsyg.2016.01116

Bubic, A., von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience, 4*. doi: 10.3389/fnhum.2010.00025

Bunzeck, N., Steiger, T. K., Krämer, U. M., Luedtke, K., Marshall, L., Obleser, J., & Tune, S. (2024). Trajectories and contributing factors of neural compensation in healthy and pathological aging. *Neuroscience and biobehavioral reviews*, *156*, 105489. doi: 10.1016/j.neubiorev.2023.105489

Burke, M. D., & Mackay, D. G. (1997). Memory, language, and ageing. *Philosophical Transactions of the Royal Society B, 352*(1363), 1845–1856. doi: 10.1098/rstb.1997.0170

Caucheteux, C., Gramfort, A., & King, JR. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behavior, 7*, 430–441. doi: 10.1038/s41562-022-01516-2

Chan, J.S., Wibral, M., Stawowsky, C., Brandl, M., Helbling, S., Naumer, M.J., Kaiser, J., & Wollstadt, P. (2021). Predictive Coding Over the Lifespan: Increased Reliance on Perceptual Priors in Older Adults—A Magnetoencephalography and Dynamic Causal Modeling Study. *Frontiers in Aging Neuroscience, 13*. doi: 10.3389/fnagi.2021.631599

Cheimariou, S., Farmer, T.A., & Gordon, J.K. (2019). Lexical prediction in the aging brain: The effects of predictiveness and congruency on the N400 ERP component. *Aging Neuropsychology and Cognition, 26*(67), 1–26. doi: 10.1080/13825585.2018.1529733

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204. doi: 10.1017/S0140525X12000477

Cosineau, D., & Chartier, S. (2010). Outliers detection and treatment: a review. International Journal of Psychological Research, 3(1), 58–67. doi: 10.21500/20112084.844

Dave, S., Brother, T.A., & Schwaab, T.Y. (2018). 1/*f* Neural Noise and Electrophysiological Indices of Contextual Prediction in Aging. *Brain research, 1691*, 34–43. doi: 10.1016/j.brainres.2018.04.007

Diacheck, E., I. Blank, M. Siegelman, J. Affourtit, and E. Federenko (2020). The domain-general multiple demand (MD) network does not support core aspects of language comprehension: A large-scale fMRI investigation. The Journal of Neuroscience, 40(23), 4536–4550. doi: 10.1523/jneurosci.2036-19.2020.

Donhauser, P. W., & Baillet, S. (2020). Two Distinct Neural Timescales for Predictive Speech Processing. *Neuron*, *105*(2), 385–393.e9. doi: 10.1016/j.neuron.2019.10.019

Ettinger, A., Linzen, T. & Marantz, A. (2014). The role of morphology in phoneme prediction: Evidence from MEG. *Brain And Language*, *129*, 14–23. doi: 10.1016/j.bandl.2013.11.004

Fedorenko, E., Gibson, E., & Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: Evidence against domain-specific working memory resources. *Journal of Memory and Language, 54*(4), 541–553. doi: 10.1016/j.jml.2005.12.006

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). *Mini-Mental State Examination (MMS, MMSE)* [Database record]. APA PsycTests. doi: 10.1037/t07757-000

Fricke, M., & Zirnstein, M. (2022). Predictive Processing and Inhibitory Control Drive Semantic Enhancements for Non-Dominant Language Word Recognition in Noise. *Languages, 7*(3), 239. doi: 10.3390/languages7030239

Friedman, N. P., & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex; a journal devoted to the study of the nervous system and behavior*, *86*, 186–204. doi: 10.1016/j.cortex.2016.04.023

Gastaldon, S., Arcara, G., Navarette, E., & Peressotti, F. (2020). Commonalities in alpha and beta neural desynchronizations during prediction in language comprehension and production. *Cortex, 133,* 328–345. doi: https://doi.org/10.1016/j.cortex.2020.09.026

Goldstein et al. (2022). Shared computational principles for language processing in humans and deep-language models. *Nature Neuroscience*, *25*, 369 - 380. doi: 10.1038/s41593-022-01026

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In A. Sayeed, C. Jacobs, T. Linzen, & M. van Schijndel (Eds.), *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)* (pp. 10-18). Association for Computational Linguistics (ACL).

Gordon, P. C., Hendrick, R., & Levine, W. H. (2002). Memory-Load Interference in Syntactic Processing. *Psychological Science*, *13*(5), 425–430. doi: 10.1111/1467-9280.00475

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. Paper presented at the Proceedings of the North American Chapter of the Association for Computational Linguistics on Language Technologies (NAACL '01), 2, 159 - 166.

Häuser, K. I., Demberg, V., & Kray, J. (2018). Surprisal modulates dual-task performance in older adults: Pupillometry shows age-related trade-offs in task performance and time-course of language processing. *Psychology and Aging, 33*(8), 1168–1180. doi: 10.1037/pag0000316

Heilbron, M., Armeni, K., Schoffelen, J., Hagoort, P. & De Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences of the United States of America, 119*(32). doi: 10.1073/pnas.2201968119

Huettig, F., & Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Language, Cognition and Neuroscience, 31*(1), 80–93. doi: 10.1080/23273798.2015.1047459

Huettig, F., & Pickering, M. J. (2019). Literacy Advantages Beyond Reading: Prediction of Spoken Language. *Trends in Cognitive Sciences, 23*(6), 464-475. doi: 10.1016/j.tics.2019.03.008

Idowu, M. I., & Szameitat, A. J. (2023). Executive function abilities in cognitively healthy young and older adults—A cross-sectional study. *Frontiers in Aging Neuroscience, 15*. doi: 10.3389/fnagi.2023.976915

Ito, A., Corley, M., & Pickering, M. J. (2017). A cognitive load delays predictive eye movements similarly during L1 and L2 comprehension. *Bilingualism: Language and Cognition, 21*(2), 251–264. doi: 10.1017/S1366728917000050

Ito, A., & Sakai, H. (2021). Everyday Language Exposure Shapes Prediction of Specific Words in Listening Comprehension: A Visual World Eye-Tracking Study. *Frontiers in Psychology, 12*. doi: 10.3389/fpsyg.2021.607474

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*(1), 32–59. Doi: 10.1080/23273798.2015.1102299

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13), 1–26. doi: 10.18637/jss.v082.i13.

Kwong See, S. T., & Ryan, E. B. (1995). Cognitive mediation of adult age differences in language performance. *Psychology and Aging, 10*(3), 458–468. doi: 10.1037/0882-7974.10.3.458

Lamichhane, B., Westbrook, A., Cole, M. W., & Braver, T. S. (2020). Exploring brain-behavior relationships in the N-back task. *NeuroImage, 212,* 116683. doi: 10.1016/j.neuroimage.2020.116683

Larsson, J., & Smith, A. T. (2012). fMRI Repetition Suppression: Neuronal Adaptation or Stimulus Expectation? *Cerebral Cortex*, 22(3), 567–576. doi: 10.1093/cercor/bhr119

Lerner, Y., Honey, C. J., Silbert, L. J. & Hasson, U. (2011). Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. *The Journal Of Neuroscience*, *31*(8), 2906–2915. doi: 10.1523/jneurosci.3684-10.2011

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126–1177. doi: 10.1016/j.cognition.2007.05.006.

Little, T. D. (2013). *Longitudinal Structural Equation Modeling (Methodology in the Social Sciences)*. New York, NY: The GuilfordPress.

Long, J.A. (2019). *interactions: Comprehensive, User-Friendly Toolkit for Probing Interactions*. R package version 1.1.0, https://cran.r-project.org/package=interactions

Lüdecke, D. (2023). *sjPlot: Data Visualization for Statistics in Social Science*. R package version 2.8.15, https://CRAN.R-project.org/package=sjPlot

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods, 49*(4), 1494–1502. doi: 10.3758/s13428-016-0809-y.

MacGregor, L., Gilbert, R. A., Balewski, Z., Mitchell, D. N., Erzinçlioğlu, S., Rodd, J. M., Duncan, J., Fedorenko, E. & Davis, M. H. (2022). Causal contributions of the Domain-General (Multiple demand) and the Language-Selective brain networks to perceptual and semantic challenges in speech comprehension. *Neurobiology of language*, *3*(4), 665–698. doi: 10.1162/nol_a_00081

Mainz, N., Shao, Z., Brysbaert, M., & Meyer, A. S. (2017). Vocabulary Knowledge Predicts Lexical Processing: Evidence from a Group of Participants with Diverse Educational Backgrounds. *Frontiers in Psychology, 8*. doi: 10.3389/fpsyg.2017.01164

Matthews, J. (2018). Vocabulary for listening: Emerging evidence for high and mid-frequency vocabulary knowledge. *System, 72,* 23–36. doi: 10.1016/j.system.2017.10.005

Martin, S., Williams, A.K., Saur, D., & Hartwigsen, G. (2022). Age-related reorganization of functional network architecture in semantic cognition. *Cerebral Cortex, 33*(1). doi: 10.1093/cercor/bhac387

Moran, R. J., Symmonds, M., Dolan, R. J., & Friston, K. J. (2014) The Brain Ages Optimally to Model Its Environment: Evidence from Sensory Learning over the Adult Lifespan. *PLoS Computational Biology, 10*(1), e1003422. doi: 10.1371/journal.pcbi.1003422

Nieuwland, M.S., et al. (2018). Large-scale sreplication study reveals a limit on probabilistic prediction in language comprehension. *eLife, 7*, e33468. doi: 10.7554/eLife.33468

Oh, B., Clark, C. & Schuler, W. (2022). Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in artificial intelligence*, *5*. doi: 10.3389/frai.2022.777963

Onnis, L., Lim, A., Cheung, S., & Huettig, F. (2022). Is the Mind Inherently Predicting? Exploring Forward and Backward Looking in Language Processing. *Cognitive Science, 46*(10), e13201. doi: 10.1111/cogs.13201

Payne, B. R., & Federmeier, K. D. (2018). Contextual constraints on lexico-semantic processing in aging: Evidence from single-word event-related brain potentials. *Brain Research, 1687,* 117–128. 10.1016/j.brainres.2018.02.021

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? Trends in Cognitive Sciences, 11(3), 105-110. doi: 10.1016/j.tics.2006.12.002

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI blog. URL:* https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

Rao, R., & Ballard, D. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience,* 2, 79–87. doi: 10.1038/4580

Reuter-Lorenz, P. A., Festini, S. B. & Jantz, T. K. (2021). Executive functions and neurocognitive aging. In K. W. Schaie & S. L. Willis (Hrsg.), *Handbook of the Psychology of Aging* (9th Ed., pp. 67–81). Academic Press. doi: 10.1016/B978-0-12-816094-7.00019-2

Richter, D., & de Lange, F. P. (2019). Statistical learning attenuates visual activity only for attended stimuli. *eLife, 8*, e47869. doi: 10.7554/eLife.47869

Rodd, J. M., Johnsrude, I. S., & Davis, M. H. (2010). The role of domain-general frontal systems in language comprehension: Evidence from dual-task interference and semantic ambiguity. *Brain and Language, 115*(3), 182–188. doi: 10.1016/j.bandl.2010.07.005

Rogers, C. (2017). Semantic Priming, Not Repetition Priming, is to Blame for False Hearing. Psychonomic Bulletin and Review, 24, 1194–1204. doi: 10.3758/s13423-016-1185-4

Rommers, J., Dickson, D. S., Norton, J. J. S., Wlotko, E. W., & Federmeier, K. D. (2017). Alpha and theta band dynamics related to sentential constraint and word expectancy. *Language, Cognition and Neuroscience, 32*(5), 576-589, doi: 10.1080/23273798.2016.1183799

Ryskin, R. & Nieuwland, M. S. (2023). Prediction during language comprehension: What is next? *Trends in Cognitive Sciences, 27*(11), 1032–1052. doi: 10.1016/j.tics.2023.08.003

Salthouse, T. A., Atkinson, T. M., & Berish, D. E. (2003). Executive Functioning as a Potential Mediator of Age-Related Cognitive Decline in Normal Adults. *Journal of Experimental Psychology: General, 132*(4), 566–594. doi: 10.1037/0096-3445.132.4.566

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, *2*(6), 110–114. doi: 10.2307/3002019

Schmitt, L., Erb, J., Tune, S., Rysop, A., Hartwigsen, G. & Obleser, J. (2021). Predicting speech from a cortical hierarchy of event-based time scales. *Science Advances, 7*(49). doi: 10.1126/sciadv.abi6070

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B. & Fedorenko, E. (2021). The neural architecture of language: integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences of the United States of America, 118*(45). doi: 10.1073/pnas.2105646118

Schweter, S. (2021). GPT-2 [dbmdz/german-gpt-2]. In *Hugging Face Model Hub*. Retrieved May 16, 2023, from https://huggingface.co/dbmdz/german-gpt2

Sedley, W., Gander, P. E., Kumar, S., Kovach, C. K., Oya, H., Kawasaki, H., Howard, M. A., III., & Griffiths, T. D. (2016). Neural signatures of perceptual inference. *eLife*, *5*, e11476. doi: 10.7554/eLife.11476

Shain, C., Blank, I. A., Fedorenko, E., Gibson, E., & Schuler, W. (2022). Robust effects of working memory demand during naturalistic language comprehension in language-selective cortex. *Journal of Neuroscience, 42*(39), 7412-7430. doi: 10.1523/JNEUROSCI.1894-21.2022

Shain, C., Meister, C., Pimentel, T. D., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, *121*(10). doi: 10.1073/pnas.2307876121

Shafto, M. A., & Tyler, L. K. (2014). Language in the aging brain: The network dynamics of cognitive decline and preservation. *Science, 346,* 583-587. doi: 10.1126/science.1254404

Smith, N. J. & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Speer, R., Chin, J., Lin, A., Jewett, S., & Nathan, L. (2018, October 3). LuminosoInsight/wordfreq: v2.2. Zenodo. doi: 10.5281/zenodo.1443582

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *The Language Learning Journal, 36*(2), 139-152, doi: 10.1080/09571730802389975

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, *31*(4), 577–607. doi: 10.1017/S0272263109990039

Terporten, R., Schoffelen, J.-M., Dai, B., Hagoort, P., & Kösem, A. (2019). The relation between alpha/beta oscillations and the encoding of sentence induced contextual information. *Scientific Reports, 9*(1), 20255. doi: 10.1038/s41598-019-56600-x

Tune, S., Alavash, M., Fiedler, L., & Obleser, J. (2021). Neural attentional-filter mechanisms of listening success in middle-aged and older individuals. *Nature Communications,* 12, 4533. doi: 10.1038/s41467-021-24771-9

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

Verhaegen, P., Borchelt, M., & Smith, J. (2003). Relation Between Cardiovascular and Metabolic Disease and Cognition in Very Old Age: Cross-Sectional and Longitudinal Findings From the Berlin Aging Study. *Health Psychology, 22*(6), 559–569. doi: 10.1037/0278-6133.22.6.559

Wang, L., Hagoort, P., & Jensen, O. (2018). Language prediction is reflected by coupling between frontal gamma and posterior alpha oscillations. *Journal of Cognitive Neuroscience, 30*(3), 432-447. doi: 10.1162/jocn_a_01190

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the Predictions of Surprisal Theory in 11 Languages. *Transactions of the Association for Computational Linguistics*, 11, 1451–1470. doi: 10.1162/tacl_a_00612

Wingfield, A., Amichetti, N. M. & Lash, A. (2015). Cognitive aging and hearing acuity: modeling spoken language comprehension. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.00684

Wolpe, N., Ingram, J., Tsvetanov, K., Geerlings, L., Kievit, R. A., Henson, R. N., Wolpert, D. M., Cam-CAN, & Rowe, J. B. (2016). Ageing increases reliance on sensorimotor prediction through structural and functional differences in frontostriatal circuits. *Nature Communications, 7,* 13034. doi: 10.1038/ncomms13034

Ylinen, S., Bosseler, A., Junttila, K. & Huotilainen, M. (2016). Predictive coding accelerates word recognition and learning in the early stages of language development. *Developmental Science*, *20*(6). doi: 10.1111/desc.12472

Xie, Z., Brodbeck, C., Chandrasekaran, B. (2023). Cortical Tracking of Continuous Speech Under Bimodal Divided Attention. *Neurobiology of Language, 4*(2), 318–343. doi: 10.1162/nol_a_00100