Executive Resources Shape the Impact of Language Predictability 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

Humans routinely anticipate upcoming language, but whether such predictions come at a cognitive cost remains debated. In this study, we demonstrate the resource-dependent nature of predictive mechanisms in language comprehension across the lifespan: Experimentally limiting executive resources through a concurrent task reduces the effect of language predictability on reading time. Participants (N=175, replication N=96) read short articles presented word-by-word while completing a secondary font colour n-back task, thus varying cognitive demand. Language predictability was indexed by word surprisal as derived from a pre-trained large language model (GPT-2). Across two independent samples, our findings reveal that language predictions are not cost-free: They draw on executive control resources, and this dependency becomes more pronounced with age (18–85 years). These results help resolve the debate over cognitive demands in language comprehension and highlight prediction as a dynamic, resource-dependent process across the lifespan.

INTRODUCTION

We constantly rely on our ability to swiftly yet accurately process linguistic input. When reading a book, watching television, or navigating a car through busy traffic while following instructions, language prediction is considered a catalyst that enhances the efficiency of linguistic processing [1–5]. However, due to the inherent flexibility and richness of natural language, upcoming words can rarely be predicted from context with complete certainty [6]. Instead, linguistic features are thought to be pre-activated broadly rather than following an all-or-nothing principle, as there is evidence for predictability effects even for moderately- or low-restraint contexts across the entire predictability span [7–10]. This generation of graded predictions is sometimes described as being passive and cost-free [11]. However, it is still under debate whether maintaining such an elaborate process really incurs no cognitive cost.

Graded language predictions necessitate the active generation of hypotheses on upcoming words as well as the integration of prediction errors to inform future predictions [1,5]. Supporting this, recent evidence suggests that language predictions may indeed impose processing demands. Shain et al. (2024) found that reading time increases with decreasing word predictability, with even small drops in predictability of highly expected words leading to significant processing costs. These findings suggest that language predictions are not entirely automatic or effortless. This aligns with numerous neuroimaging studies arguing for a strong interaction between language-specific and domain-general executive brain regions [12–15], particularly in situations that are cognitively demanding [16–20].

In this context, domain-general executive resources refer to higher-level cognitive control processes, such as working memory, inhibitory control, and cognitive flexibility, that are crucial for managing and coordinating behaviour across a wide range of tasks and modalities [21–23]. These processes are supported by the multiple-demand (MD) system, a fronto-parietal network that is recruited in various cognitively challenging situations [22]. However, while numerous studies support the claim that language prediction relies on such domain-general executive resources, a body of research suggests the opposite (e.g., [24–27]). This raises the unresolved question to what extent such resources are taxed by predictive processes [5,28–30].

An interesting test case for the dynamic interplay between language-specific and domain-general systems is cognitive ageing, as advancing age has been shown to be associated with sensory and executive decline [31–33]. In the current study, we thus ask: How does language prediction change when executive resources are limited – both intrinsically due to advanced age, and extrinsically through increased task demands?

The age-related change in cognitive resources is reflected by longer linguistic processing times, especially in situations with high cognitive effort, such as dual-task processing [34,35]. However, previous research presents conflicting results on how cognitive ageing affects the use of linguistic predictions during language comprehension. Behavioural studies suggest a decline in using context to make semantic predictions with age [36] while EEG studies present a more nuanced picture. Some studies indicate heightened neural sensitivity to unexpected information in older adults [37], while others report no significant age-related differences in neural response [38,39].

Furthermore, it is unclear how the use of language predictions across the adult lifespan might be affected by increasing cognitive demands. According to the compensation-related utilisation of

neural circuits (CRUNCH) hypothesis [40], older adults might earlier reach a point where their cognitive load capacity is fully exhausted than young adults, leading to a performance decline. If language prediction draws on executive resources, its effect on reading time might thus diminish with increasing cognitive load due to shared cognitive resources. However, this effect might re-emerge once capacity limits are reached, causing tasks to be processed sequentially, which would result in an increased effect of language prediction paralleled by slower performance.

Here, we explored the role of executive control in language prediction in two large cohorts across the adult lifespan. Using a novel dual-task paradigm that couples natural reading with an n-back task that taxes executive resources, we tested the following hypotheses: First, both increased cognitive load and reduced word predictability (i.e., increased surprisal) should be reflected by longer reading time (Fig. 1a and b), alongside a decrease in text comprehension and n-back task performance. Second, we expected that the formation of language predictions should be contingent upon the availability of executive resources. Most importantly, a gradual limitation of these resources due to increased task demands should result in diminished effects of language predictability on reading time (interaction between cognitive load and surprisal; Fig. 1c).

Lastly, we explored how these previously described effects would be modulated by age (Fig. 1d). Note that the literature allows for contradicting hypotheses: On the one hand, if language predictions are impaired under limited executive resources, older individuals should rely less on language predictions due to overall decreased executive resources [31,32]. This should be reflected by diminished predictability effects. On the other hand, given the presumed stability of language comprehension across the lifespan [41], older adults might instead rely more heavily on language predictions, thus fully compensating for any impairments in reading comprehension caused by sensory and executive decline [31–33]. In this case, we should see strong surprisal effects independent of age, or even stronger surprisal effects in older than younger adults.

The two large-sample reading-time studies presented here help resolve the ongoing discussion on the role of executive resources in language predictions with three key findings: First, across our large age range, we found a general increase in reading time with both low language predictability and high task demands. Second, higher task demands reduce the influence of language predictability on reading time, indicating that linguistic prediction relies on executive control resources. Third, predictability had a more pronounced effect on reading time in older adults compared to younger ones. These findings highlight the dynamic interaction between language predictability and executive resources.

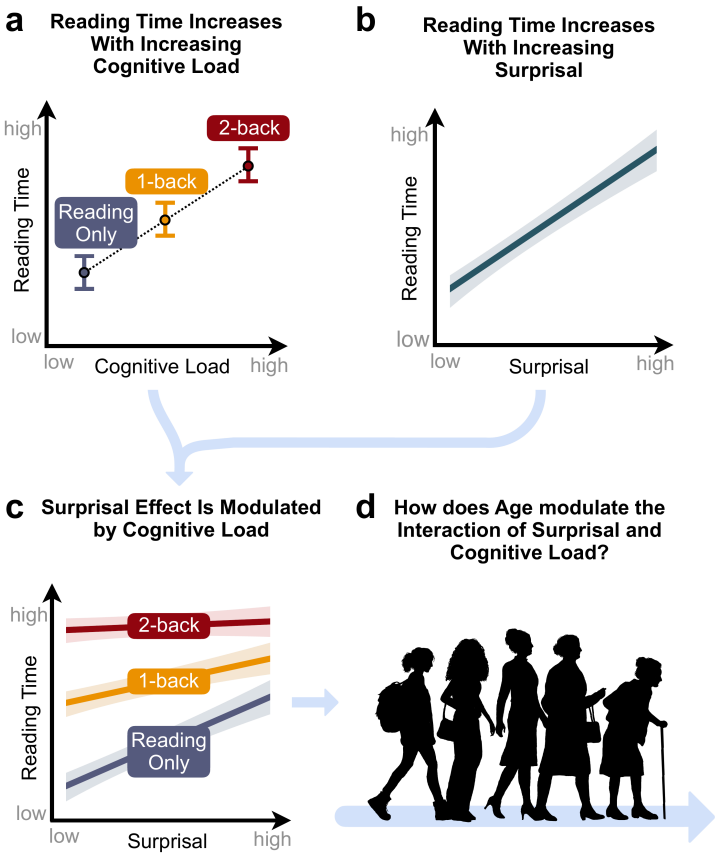

Figure 1. Visualisation of hypotheses. We expected main effects on reading time of (a) cognitive load and (b) surprisal, as well as (c) an interaction of surprisal and cognitive load. Additionally, (d) we explored how these effects are modulated by age.

RESULTS

We report data from 175 participants (M = 44.9 ± 17.9, 18-85 years, 51% female) who were either tested online (N_{online} = 80) or in the laboratory ($N_{laboratory}$ = 95) in one session. Moreover, we conducted an internal, pre-registered replication study involving another 96 participants (M = 39.8 ± 14.0, 18-70 years, 51% female) tested online to replicate our main findings (see Fig. S1 for age distributions of all samples). During the experiment, participants engaged in a self-paced reading task. They read 300-word newspaper articles, presented word-by-word in various font colours (Fig. 2a). The task was either performed in isolation (Reading Only) or paired with a competing n-back task on the words' font colour (1-back and 2-back Dual Task). Participants were instructed to read the texts carefully, as content-related multiple-choice questions were asked after each text (see *Methods* for details).

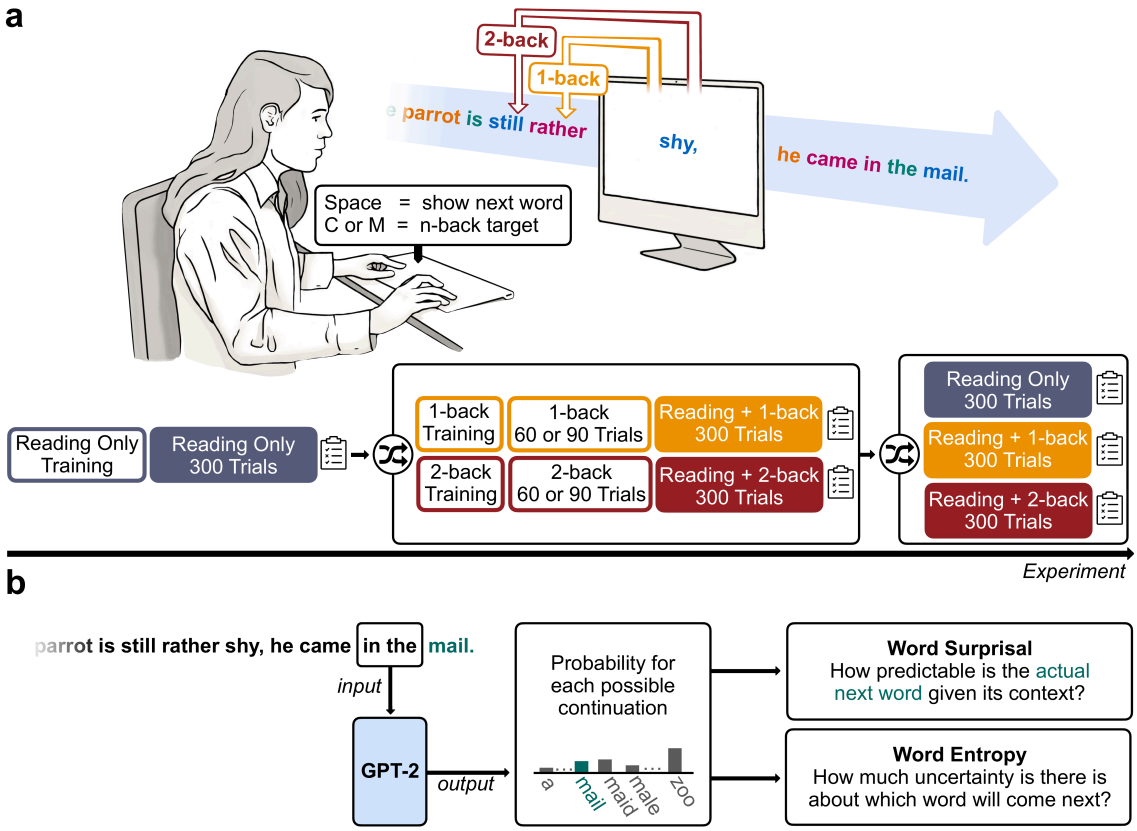

Figure 2. Experimental design and quantification of predictability as word surprisal using a large language model (GPT-2). (a) Participants were asked to perform a self-paced reading task (Reading Only) which was complemented in some blocks by a secondary n-back task on the font colour of the words (Reading + 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 one surprisal score for each word in the reading material by using context chunks of 2 words as prompts for next-word-predictions in GPT-2. The resulting probability for the actual next word in the text (here: "mail", marked in teal) was then transformed into a surprisal score, which reflected how predictable the respective word was given the context. Additionally, based on the distribution of probabilities for all possible continuations, we computed an entropy score, which reflects the uncertainty in predicting the next word. Please note that the example sentence used here has been translated to English for better comprehensibility, while the original text materials were in German.

Increased Cognitive Load and Older Age Reduce Task Performance

Results from a generalised mixed model (GMM) for text comprehension accuracy showed that participants read the texts carefully and answered most comprehension questions correctly, with accuracies of 93% ± 14% (mean ± SD) in the Reading Only condition, and 80% ± 25% in the 1-back and 73% ± 28% in the 2-back Dual Task conditions (Fig. S2b). Increases in cognitive load ($OR_{1\text{-}back} = 0.253$, $z_{1\text{-}back} = -8.928$, $p_{1\text{-}back} < 0.001$, $OR_{2\text{-}back} = 0.156$, $z_{2\text{-}back} = -12.403$, $p_{2\text{-}back} < 0.001$) and in age ($OR = 0.986$, $z = -2.676$, $p = 0.013$) were associated with poorer performance (see *Methods* for model details; Fig. 3a, Table S1).

Similarly, d' values demonstrated good performance in the n-back task, with mean d' of 3.77 ± 0.8 in the 1-back and 2.12 ± 0.87 in the 2-back Dual Task condition (Fig. S2a). The overall high d' values observed here can be attributed to the low target ratio in the experiment, resulting in high correct rejection and low false alarm rates. A linear mixed-effects model revealed n-back performance declined with cognitive load (β = -0.014, t(170.5) = -3.961, p < 0.001) and with age (β = -0.005, t(170.17) = 25.462, p > 0.001; Fig. 3b, Table S1).

Reading Time Increases with Age, Surprisal, and Cognitive Load

To operationalise language predictability, we calculated word surprisal scores using a 12-layered generative pre-trained transformer model (GPT-2; [42]; Fig. 2b). Word surprisal quantifies the predictability of the current word given its preceding context [10,43,44]. We chose a context length of two words, as constraining the context has been shown to increase GPT-2's psychometric predictive power, making its next-word predictions more human-like [45]. In addition to each word's surprisal, we also computed word entropy, which reflects the uncertainty in predicting the next word. Thus, while word surprisal indicates the predictability of each word, word entropy reflects the uncertainty underlying its prediction.

We used linear mixed-effects regression (LMM) to assess the effect of word surprisal on reading time and its interaction with cognitive load and age (see Methods for model details). Our model as a whole was able to explain 65% of variance (conditional *R2*) in single-word reading time from surprisal, n-back load, n-back performance and other linguistic and demographic variables, and up to 81% when additionally considering the variability across subjects (marginal *R2*; see Table S2).

As hypothesised, we observed significantly longer reading time with advancing age, independent of cognitive load condition (β = 0.009, t(178.46) = 9.199, p < 0.001). For illustration, each additional year of age increased reading time by roughly 1%, respectively (see Fig. 3c, top; Table 1; Fig. S3a).

To account for the potential influence of verbal intelligence and education on the observed age effects, we compared three additional LMMs: a baseline model without these predictors, and two additional models, one including a verbal intelligence predictor and the other including an education predictor (please see *Methods* section "*Control Analysis: Assessing Potential Effects of Verbal Intelligence and Education*").

Adding verbal intelligence as a predictor to our model did not significantly improve the model fit (χ^2(1) = 0.769, p = 0.381). This implies that verbal intelligence cannot account meaningfully for the differences in reading time found between younger and older participants.

When including years of education as a predictor, we observed a significant effect on reading time (β = 0.009, t = 9.199, p < 0.001), which led to a modest yet significant improvement in model fit (χ^2(1) = 4.209, p = 0.0402, $AIC_{model_baseline}$ = 37580, $AIC_{model_education}$ = 37578). However, the 3-way interaction of age, surprisal and cognitive load we found in our original model remained significant ($\beta_{1\text{-}back}$ = -0.00007, $t_{1\text{-}back}$(88.61) = -5.229, $p_{1\text{-}back}$ < 0.001, $\beta_{2\text{-}back}$ = -0.00004, $t_{2\text{-}back}$(88.90) = -2.784, $p_{2\text{-}back}$ = 0.007), suggesting that while education has a significant effect on reading time, it cannot account for the age-related effects observed in our models.

Analogous to the effect of age, increased cognitive load was associated with significantly longer reading time ($\beta_{1\text{-}back}$ = 0.474, $t_{1\text{-}back}$(176.18) = 34.046, $p_{1\text{-}back}$ < 0.001, $\beta_{2\text{-}back}$ = 0.792, $t_{2\text{-}back}$(173.76) = 30.339, $p_{2\text{-}back}$ < 0.001; Table 1 and Fig. S2 and S3), indicating participants read more slowly when faced with a more challenging task. Even after excluding the Reading Only condition from the model and comparing only the two equally attention-demanding dual-task conditions 1-back and 2-back (control analysis), hereby controlling for attentional switching costs, this effect still held true (β = 0.339, t(178.98) = 16.221, p < 0.001; Table S4).

Moreover, in line with our hypothesis, we observed a consistent increase in reading time with higher surprisal (β = 0.002, t(2361.37) = 11.321, p < 0.001; Table 1 and Fig. 3c, bottom). Highly predictable words were read more quickly. Specifically, for a change in surprisal by one standard deviation, reading time increased by about 2.1%. For words and individuals matched in all other regards, this translates to a mean reading-time difference of approximately 118 ms between words with the minimum and maximum surprisal values in our dataset.

Cognitive Load Reduces the Impact of Surprisal

In line with our hypotheses, cognitive load significantly modulated the effect of surprisal on reading time ($\beta_{1\text{-}back}$ = -0.001, $t_{1\text{-}back}$ (287959.11) = -6.772, *p1-back* < 0.001, $\beta_{2\text{-}back}$ = -0.001, $t_{2\text{-}back}$(288294.96) = -7.681, $p_{2\text{-}back}$ < 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 both the 1-back and 2-back Dual-Task condition, where cognitive load was increased (Table 1 and Fig. 3d, left).

Age Modulates the Effect of Cognitive Load

While increased task demands were associated with prolonged reading time across participants, we found this effect became less pronounced with advancing age ($\beta_{1\text{-}back}$ = -0.003, $t_{1\text{-}back}$(171.99) = -3.606, $p_{1\text{-}back}$ < 0.001; $\beta_{2\text{-}back}$ = -0.002, $t_{2\text{-}back}$(170.79) = -1.690, $p_{2\text{-}back}$ = 0.097; Table 1 and Fig. 3d, middle). To illustrate, when comparing the reading time in the 2-back Dual Task and the Reading Only condition, we found an increase of 130.48% in an average young (i.e., 27 years old) and an increase of 111.30% in an average older (i.e., 63 years old) participant.

Age Modulates the Effect of Surprisal

Older age was associated with stronger surprisal effects (β = 0.00004, t(287771.27) = 9.287, p <0.001; Table 1 and Fig. 3d, right), indicating that highly unpredictable (i.e. surprising) words were associated with a significantly longer reading time, especially in older adults. This suggests that, as individuals age, the effect of word predictability on their reading time becomes increasingly pronounced instead of remaining constant or diminishing.

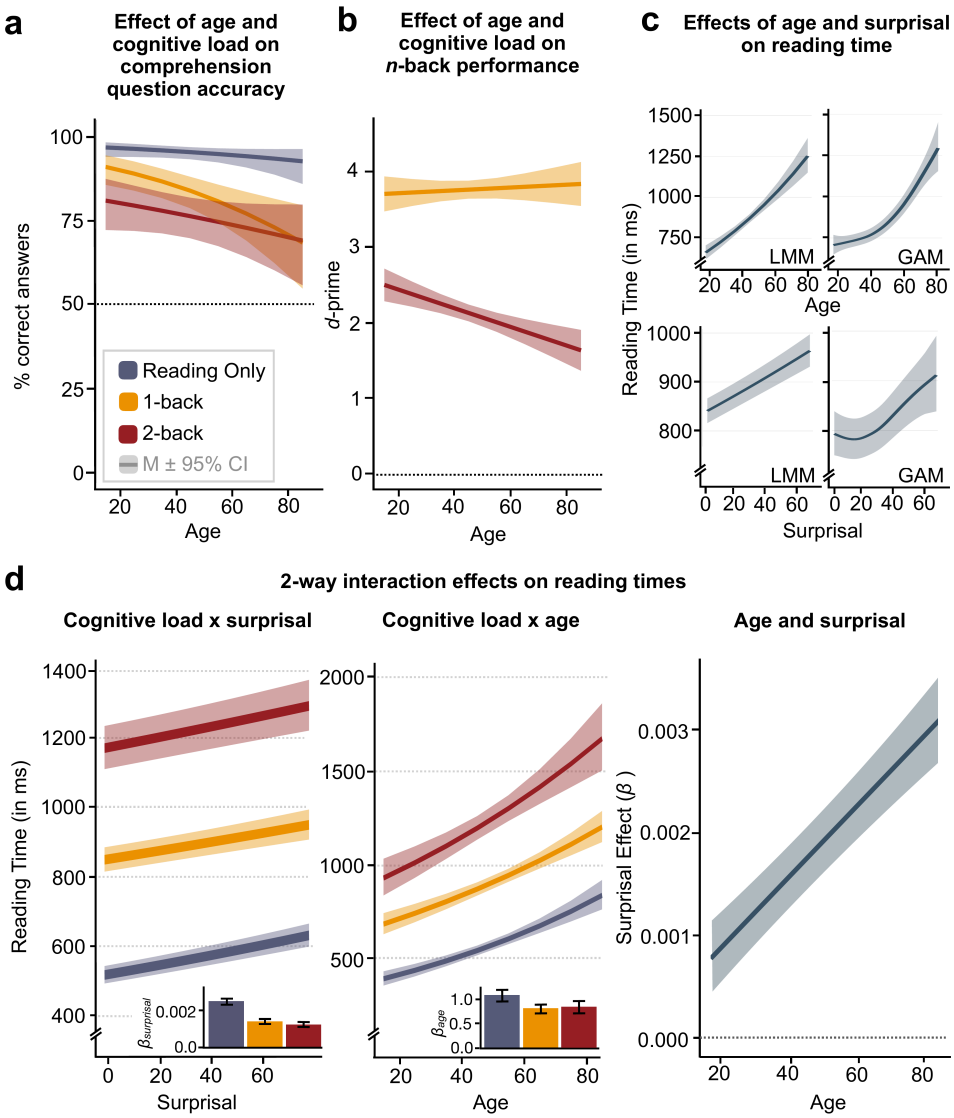

Figure 3. **Estimated marginal effects of predictors age, cognitive load and surprisal on task performance and reading time**. Main effects of cognitive load and age on accuracy in the comprehension question task (panel **a**) and on n-back task performance (d-primes; panel **b**). Please note that we 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 and word surprisal (panel **c**, left: results from linear mixed model, LMM, right: results from generalised additive model, GAM – for an explanation see section *Modelling Potential Non-Linear Distributions*). In panel d, we show the 2-way interaction of cognitive load and surprisal (left) and cognitive load and age (middle). In both cases, effects were strongest in the Reading Only condition (see barplot insets). Additionally, we show how age modulates the effect of surprisal on reading time (panel **c**, right). For raw and predicted individual trajectories, please see Fig. S2 and S3 in the Supplementary Material. Estimated marginal effects were adjusted for "Reading Only" as the reference level.

Age Modulates the Interaction of Surprisal and Cognitive Load

Finally, we tested whether the observed interaction between cognitive load and surprisal changes with advancing age: We found that age indeed modulated the joint influence of cognitive load and surprisal on reading time, as reflected by a significant three-way interaction of surprisal, cognitive load and age ($\beta_{1\text{-back - Reading Only}}$ = -0.00011, $t_{1\text{-back - Reading Only}}(287807.34)$ = -12.2661, $p_{1\text{-back - Reading Only}} < 0.001$, $\beta_{2\text{-back - Reading Only}}$ = -0.00008, $t_{2\text{-back - Reading Only}}(287771.65)$ = -8.484, $p_{2\text{-back - Reading Only}} < 0.001$; Table 1).

Even after excluding the Reading Only condition from the model and contrasting only the dual-task conditions 1-back and 2-back (control analysis), results still showed a significant three-way interaction of age, cognitive load and surprisal ($\beta_{2\text{-back - 1-back}}$ = 0.00003, $t(188203.53)$ = 3.373, p = 0.001; Table S4). Please note the change in reference level caused by the exclusion of the Reading Only condition, which causes a reversal in the direction of the three-way interaction between age, cognitive load, and surprisal.

To get a more nuanced understanding of age-related difference in the effect of cognitive load and surprisal on reading time, we conducted a simple slopes analysis for our original model (Fig. 4, Fig. S4): Under low cognitive load (condition Reading Only), surprisal significantly influenced reading time in all but the youngest participants. Put simply, when participants read a text without having to perform an additional n-back task, predictable (i.e., low-surprisal) words yielded significantly shorter reading time than unpredictable (i.e., high-surprisal) words (Fig. 4a, left panel, first plot). This effect was most pronounced in the oldest participants ($\beta_{85\text{ - 18 years}}$ = 0.006, $p < 0.001$).

As task demands increased, we observed a reversal of this pattern, with younger participants exhibiting stronger surprisal effects than older participants (condition 1-back; Fig. 4a, left panel, second plot). Notably, under increased cognitive load, even the youngest participants started showing significant surprisal effects ($\beta_{85\text{ - 18 years}}$ = -0.0009, p = 0.04).

Finally, in the most demanding condition (2-back), the pattern shifted again, with older adults again showing stronger surprisal effects than their younger counterparts ($\beta_{85\text{ - 18 years}}$ = 0.001, p = 0.003; Fig. 4a, left panel, third plot).

Comparing surprisal effects between cognitive load conditions revealed that older adults showed the most pronounced reduction in surprisal effects as cognitive load increased (Fig. 4b), which suggests they were more vulnerable to increased task demands than younger participants (18 year-old: $\beta_{1\text{-back - Reading Only}}$ = -0.003, $p < 0.001$, $\beta_{2\text{-back - Reading Only}}$ = -0.003, $p < 0.001$, $\beta_{2\text{-back - 1-back}}$ = 0.0004, p = 0.06; 85 year-old: $\beta_{1\text{-back - Reading Only}}$ = -0.011, $p < 0.001$, $\beta_{2\text{-back - Reading Only}}$ = -0.008, $p < 0.001$, $\beta_{2\text{-back - 1-back}}$ = 0.003, $p < 0.001$).

Modelling Potential Non-Linear Contributions

To account for the possibility that age, surprisal, and their interaction with cognitive load might demonstrate non-linear effects on reading time, we fitted a generalised additive mixed-effects model (GAM) to our data, using the same model structure as for the linear regression model (see *Methods* for model details). Results from the GAM revealed overall similar effects relative to the LMM. The effective degrees of freedom (EDF) of all continuous predictors were estimated above 1, confirming their non-linearity. Similar to the LMM, both our predictors of interest, surprisal and age, demonstrated significant

effects on reading time (Fig. 3c, right), with surprisal showing a more non-linear trajectory than age (EDF for surprisal: 4.107, EDF for age: 3.028, both $p < 0.001$). The smoothing splines for the three-way interaction of age, surprisal, and cognitive load showed significant effects for all levels of cognitive load (Fig. S4 and S5). Similar to the LMM, we found the strongest effect for the Reading Only condition (EDF for Reading Only: 10.248, $p < 0.001$, EDF for 1-back: 2.017, $p = 0.036$, EDF for 2-back: 4.89, $p = 0.024$). The full model results are reported in Supplementary Table S5 and Figure S4.

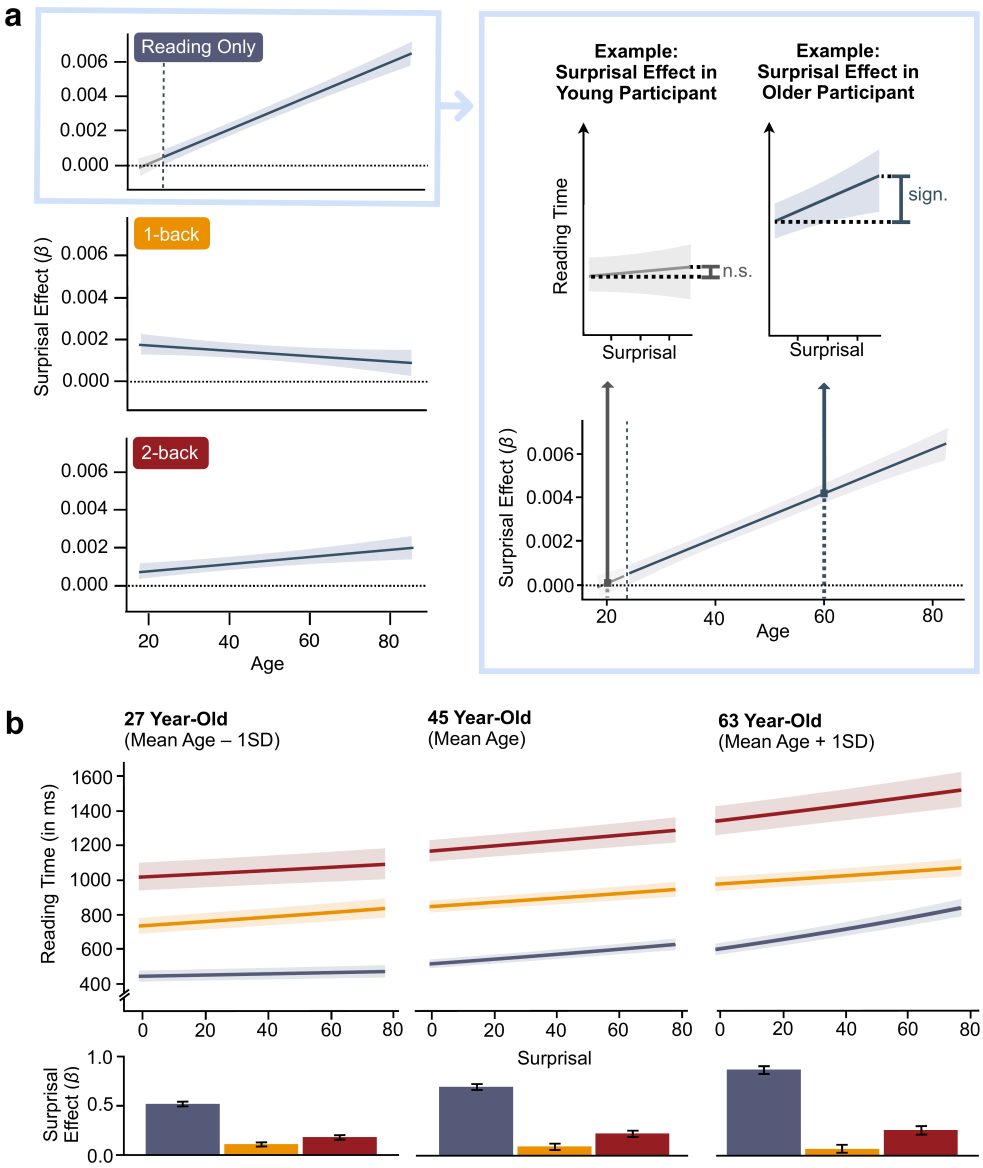

Figure 4. Results of the simple slopes analysis and exemplary marginal effects plots for three different ages. In the Johnson-Neyman plot [46] on the left side of panel (**a**), we show the effect of surprisal on reading time across the whole age range separated by cognitive load condition: *Reading Only* (top; blue), *1-back Dual Task* (middle; yellow) and *2-back Dual Task* (bottom; red). 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 reading time in the respective condition, whereas blue areas indicate a significant surprisal effect (see inset on the right for a visualisation of a non-significant effect in a younger participant and a significant effect in an older participant). In panel (**b**), we show the predicted surprisal effect in each cognitive load for an average young (average age - 1SD), middle-aged (average age) and older participant (average age + 1SD). The bar plots illustrate the predicted effects of surprisal on reading time across the three cognitive load conditions for those three average participants.

Table 1. Main results for model for reading time (N = 175).

	Predictors	*Estimate*	*Std. Error*	*CI*	*t*	*df*	*p*	
Main Effects	surprisal	0.001707	0.000151	0.001411 – 0.002002	11.320677	2361.37	1.368×10^{-28}	*
	age	0.009113	0.000991	0.007158 – 0.011068	9.199100	178.46	1.751×10^{-16}	*
	cognitive load [1-back vs. Reading Only]	0.473800	0.013916	0.446336 – 0.501264	34.046321	176.18	8.399×10^{-79}	*
	cognitive load [2-back vs. Reading Only]	0.791540	0.026090	0.740046 – 0.843034	30.338989	173.76	7.320×10^{-71}	*
2-way Interactions	surprisal x age	0.000035	0.000004	0.000027 – 0.000042	9.287151	287771.27	3.481×10^{-20}	*
	surprisal x cognitive load [1-back vs. Reading Only]	-0.001093	0.000161	-0.001409 – -0.000776	-6.771521	287959.11	2.043×10^{-11}	*
	surprisal x cognitive load [2-back vs. Reading Only]	-0.001255	0.000163	-0.001575 – -0.000935	-7.681261	288294.96	2.709×10^{-14}	*
	age x cognitive load [1-back vs. Reading Only]	-0.002798	0.000776	-0.004330 – -0.001267	-3.606479	171.99	5.135×10^{-4}	*
	age x cognitive load [2-back vs. Reading Only]	-0.002458	0.001454	-0.005329 – 0.000412	-1.690400	170.79	9.681×10^{-2}	
3-way Interactions	surprisal x age x cognitive load [1-back vs. Reading Only]	-0.000111	0.000009	-0.000129 – -0.000094	-12.266076	287807.34	3.748×10^{-34}	*
	surprisal x age x cognitive load [2-back vs. Reading Only]	-0.000078	0.000009	-0.000096 – -0.000060	-8.483676	287771.65	4.384×10^{-17}	*
Model Fit	Intra-class Correlation (ICC)	0.46						
	Marginal R^2 / Conditional R^2	0.643 / 0.807						

Note. All continuous predictors were centred. Degrees of freedom for *p*-values, standard errors and confidence intervals (CI) were computed using Satterthwaite's approximation. All *p*-values reported here are FDR-corrected and were computed using ANOVAs with type III sum of squares. Results that are significant on an alpha-level of 0.05 are marked with a star.

Interaction of Age, Surprisal, and Cognitive Load Generalises to New Sample

To probe the reliability of our findings, we conducted an exact online replication of the original experiment. When comparing the results of the online participants of the original sample and the online replication sample, we find similar patterns, as illustrated in Figure 5. We again found a significant 3-way interaction of age, cognitive load and surprisal ($\beta_{1\text{-}back}$ = -0.00002, $t_{1\text{-}back}$(161446.16) = -1.605, $p_{1\text{-}back}$ = 0.1386, $\beta_{2\text{-}back}$ = -0.00006, $t_{2\text{-}back}$(161861.30) = -3.871, $p_{2\text{-}back}$ < 0.001; see Table S3, Fig. 5, and Fig. S6), albeit using a smaller sample size of 96 instead of 175 participants and a slightly different age distribution. Notably, the three-way interaction showed similar effect sizes and directions of effects in the replication sample relative to our original online sample.

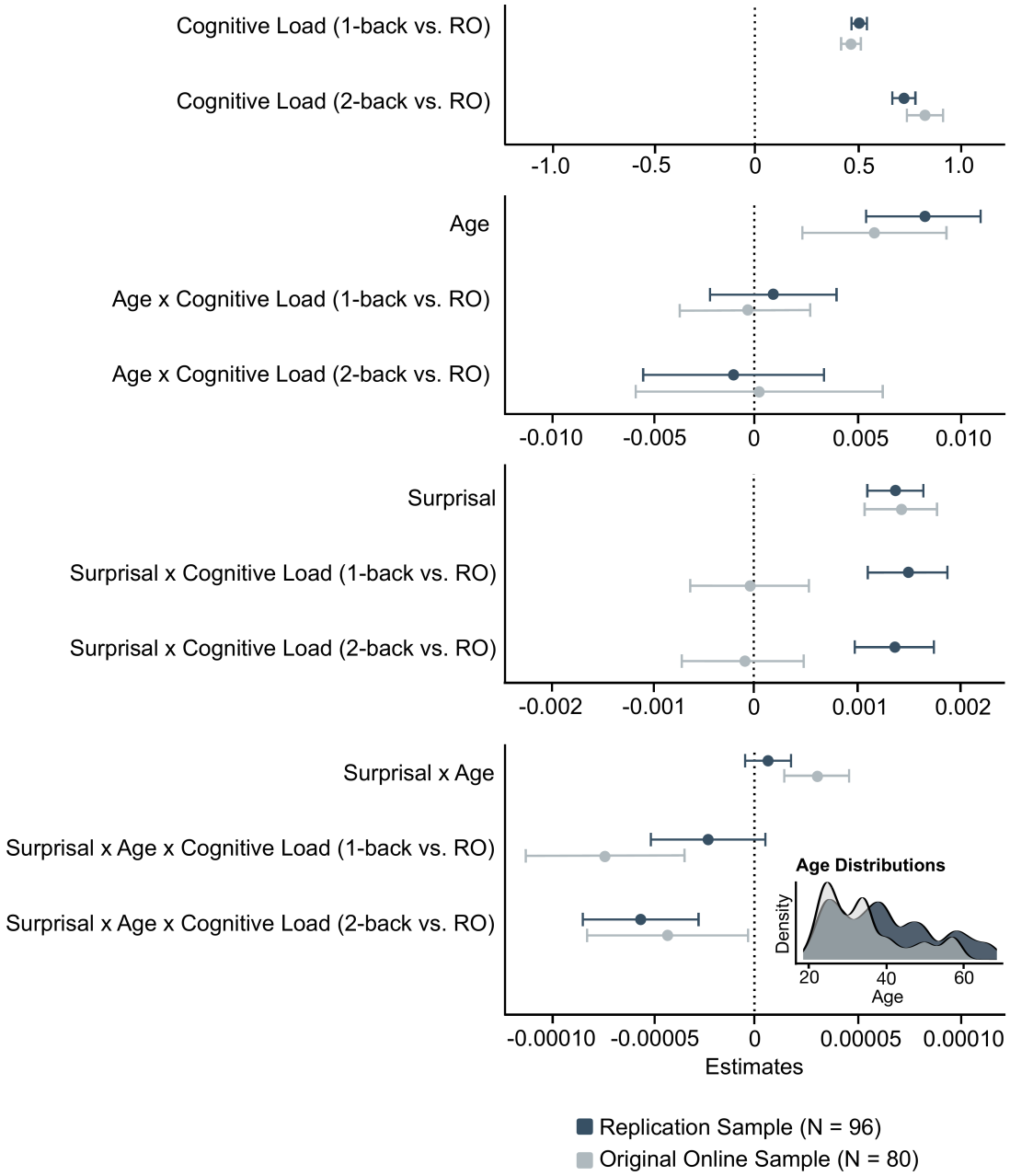

Figure 5. Results of the internal online replication in comparison with the results of the online sample of the original study. Shown are estimates ± CI for all main effects of surprisal, age, and cognitive load as well as their 2-way and 3-way interactions. RO: Reading only. For full results please see Table S3. For a comparison of age distributions in the original online and lab sample and the online replication sample, please see Fig. S1. Please note that effects are grouped by their magnitude.

Modelling Cumulative Effects of Surprisal on Reading Time

As noted above (see section *Reading Time Increases with Age, Surprisal, and Cognitive Load*), reading time increased as a function of word surprisal, with a mean difference of 118 ms between the most and least predictable word in our text material. This corresponds to approximately 22.91% of the average per-word reading time in the Reading Only condition (M = 522.04 ± 275.927 ms), highlighting a substantial effect of surprisal – especially considering that all other predictors were held constant when estimating the effect of surprisal. The substantial effect of surprisal is particularly notable given that the texts were edited for ease of comprehension and contained relatively low surprisal values overall (M = 18.165 ± 7.523), indicating that the observed reading time differences between high- and low-surprisal words likely underestimate the potential effect size in more complex texts with higher surprisal variability.

Notably, in everyday life, we typically encounter sequences of words, ranging from short phrases to texts of hundreds or thousands of words. Consequently, small word-level effects can accumulate and yield substantial processing differences over time (cf. [47]). Thus, to quantify this cumulative effect of surprisal, we predicted reading time for two average participants aged 27 (M - 1SD) and 63 (M + 1SD) years for a short example sentence. Predictions were conducted for the easiest cognitive load condition Reading Only, in which surprisal effects were most pronounced, and for the most challenging condition 2-back Dual Task, where surprisal effects were diminished. The example sentence comprised 14 words and had relatively low surprisal values (M = 15.24 ± 7.65; even slightly lower than in our original text material), implying that the cumulative effects of surprisal shown here are rather conservative estimations.

In the condition with the largest surprisal effects (Reading Only), surprisal led to a total cumulative increase in reading time of 73.6 ms for younger and 648 ms for older participants over the course of the sentence. In the more challenging 2-back Dual Task condition, we observed a total increase of 199 ms in younger and 485 ms in older participants (see Fig. 6b for cumulated effects of surprisal; see Fig. 6a for predicted reading time incorporating all effects for a comparison).

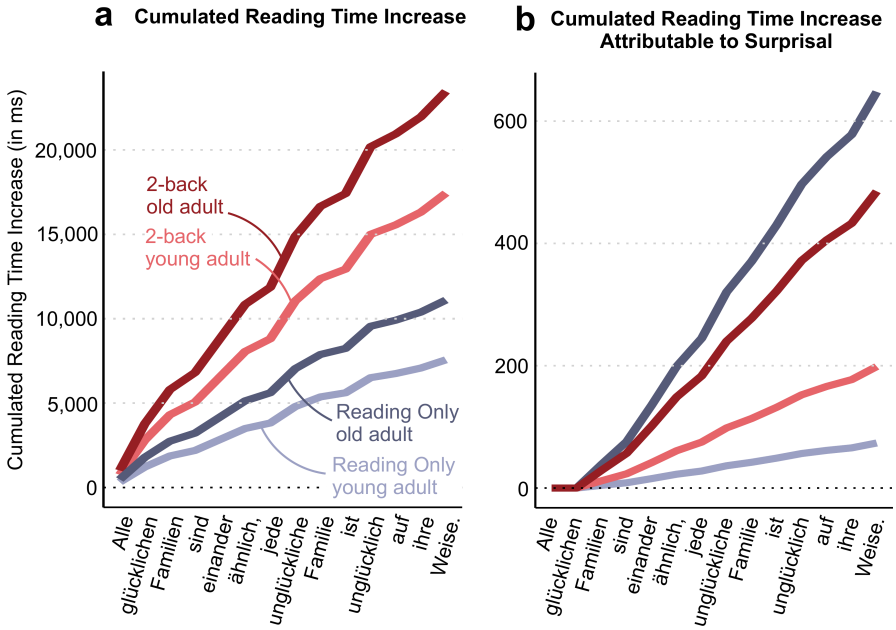

Figure 6. **Cumulative Effect of Surprisal on Reading Time.** To illustrate the cumulative effect of surprisal on reading time over the course of a text, we predicted reading times for an average younger (27 years, M - 1SD) and average older (63 years, M + 1SD) participant in the easy Reading Only condition (blue) and the most challenging condition 2-back (Dual Task; red) and computed the cumulative sum for a short example sentence. Panel **a** illustrates how reading time gradually increases in total over the course of the sentence, with all predictors being held constant at their average, except for the predictors age, cognitive load and word length. In panel **b**, we again show cumulative reading times, this time isolating the effect of surprisal. Please note that surprisal values are zero for the first two words, as our GPT-2 model estimates surprisal based on the two preceding words, which are unavailable at the beginning of the sentence. The example sentence used in both panels is the German translation of the opening line of *Anna Karenina*, "Happy families are all alike, every unhappy family is unhappy in its own way" [48].

DISCUSSION

Linguistic predictions are a powerful feature of language comprehension. But do they really come 'for free', or how much do they draw on executive resources? With the present study, we asked how language predictions 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 time as well as a decrease in task performance across age groups. This is consistent with previous studies using n-back tasks [49,50].

Next, as hypothesised, higher word surprisal slowed down reading, even when controlling for word length and frequency as well as prediction entropy. This effect of surprisal replicates findings from previous studies showing that highly unpredictable words are generally associated with a longer reading time [9,10,51–54].

Finally, we explored the relationship between surprisal and cognitive load across the (cross-sectional) adult lifespan. We hypothesised that increasing cognitive load should gradually impair the building of language predictions, which should surface in a diminished surprisal effect on reading time in conditions with high cognitive load. In line with our hypothesis, we found that the effect of word surprisal on reading time was modulated by cognitive load. Specifically, when cognitive load was high, the effect of surprisal on reading time was significantly diminished.

Interestingly, this interaction between surprisal and cognitive load was modulated by age. While age generally increased the reliance on language prediction, it also increased the susceptibility of this strategy to changes in available executive resources: Older adults showed the strongest relative reduction of the surprisal effect with increasing cognitive load. However, under high load, older adults still showed the strongest surprisal effect in absolute terms (Figure 4b). In a direct replication of the original experiment, we reproduced this finding, further confirming the reliability of our results.

Disentangling Effects of Attention Versus 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 word predictability. We take this as evidence that executive resources are involved in the generation of language predictions.

Predictive processing has not only been suggested to be foundational to language comprehension [5,9,51], but it is also thought to be a core mechanism of the human brain [4,55,56]. 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 [57,58]. 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. The predictability effect did not only diminish when a secondary task was introduced but also decreased with increasing cognitive load even when attentional switching costs were held constant between conditions (i.e., when comparing the 1-back to 2-back load conditions). In line with previous literature [59–61], our results thus suggest that executive processes beyond attention, such as updating and maintenance of context information, shifting between tasks, and inhibition of irrelevant information [21,23], are integral to language prediction.

Language Prediction as a Compensatory Mechanism in Older Age

Further examining the effect of language predictability across the adult lifespan revealed interesting age differences. Namely, older adults showed stronger language predictability effects than younger participants. This effect held true even when controlling for potential differences in verbal intelligence and education between participants. Our result aligns with previous work on age-related changes in linguistic predictions, indicating heightened sensitivity to unexpected lexical input in older adults [37].

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 [13,32]. Here, we show that even under increased cognitive load, older adults still rely heavily on their refined system of linguistic predictions driven by their lifelong experience. This allows them to make more fine-tuned predictions but also renders them more vulnerable to unexpected information (i.e., high-surprisal words) than their younger counterparts.

Previous studies have found a larger vocabulary size associated with more rapid processing of language and improved language comprehension [62–67]. 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 [68,69]. This accumulated skill is thought to serve as a compensatory mechanism for decline in working memory capacities or reduced executive functioning with age [70,71]. Indeed, speech comprehension appears to remain largely intact in older adults [41,72].

As a caveat, one should not disregard the possibility of age-related differences between younger and older adults in utilising formed predictions. While we assumed that formed predictions are utilised automatically, and that the observed differences in the effect of language predictability on reading time between individuals might thus be attributable to a difference in executive resources, one could also argue that individuals might simply weigh their formed predictions differently.

Akin to the longer exposure to language in older individuals, it is reasonable to assume 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 [73,74]. Additionally, due to age-related sensory decline [33], older adults might exhibit a stronger dependence on context-based predictions to process language, as incoming sensory information might be less informative [75–77]. A stronger reliance on language predictions could thus serve as a compensatory mechanism to facilitate language comprehension despite sensory decline in older adults.

How Can Limited Executive Resources Affect Language Prediction?

As shown in Figure 4, the relationship between word predictability and age depends on cognitive load. Under low load, predictability effects on reading time increased with age; older adults showed robust effects, while younger adults showed none. Under intermediate load (1-back), this pattern reversed, with younger adults showing stronger predictability effects than older adults.

This reversal begs for an explanation, and we deem it most likely to reflect differences in how executive resources are deployed. In low-load settings, young adults may process both predictable and unpredictable words efficiently, minimizing observable surprisal effects. The absence of a predictability effect in this group should therefore not be taken as evidence against predictive processing. Rather, it may indicate that prediction is less necessary when processing is fast and flexible. Under intermediate load, executive resources are partially taxed, revealing underlying prediction processes in young adults. At higher loads, predictability effects diminish again, suggesting resource constraints impair predictive processing.

In older adults, however, predictability effects decline already at intermediate load, consistent with the CRUNCH model [40], which posits that cognitive capacity limits are reached earlier in aging. At this point, resources are insufficient to maintain predictive processing while also performing the secondary task. Behaviourally, this may result in fluctuating performance or trial-wise switching between the two tasks. As load increases further, older adults continue to show reduced, though still measurable, predictability effects – indicating sustained but strained processing.

Taken together, the data suggest that both age groups experience a reduction in predictive processing when executive resources are limited, but the "crunch point" is reached earlier in older adults.

Limitations and Future Directions

One intriguing question that remains is how n-back performance and language surprisal interrelate. It is plausible that when texts are highly predictable (i.e., when surprisal is very low on average), the cognitive load associated with language processing is reduced. This reduction could free up domain-general executive resources, thereby enhancing n-back performance. Conversely, when surprisal is high, the increased demands of processing less predictable language may compromise working memory updating and lead to poorer n-back performance. As we assessed n-back performance at block-level and maintained equal predictability across texts, analysing trial-level effects of surprisal on n-back task performance was not possible. Future studies could address this limitation by systematically examining the relationship between surprisal and n-back performance at a more granular level.

Another important area for future research involves exploring potential age-related differences in task strategies within dual-task designs: We previously hypothesised that the differing effects of cognitive load on surprisal-driven reading times across age groups may reflect compensatory strategies in older adults. Given declines in executive control and working memory [31–33], older adults may prioritise language processing over multitasking under high load, whereas younger adults might distribute resources more flexibly. Future studies should thus examine how age affects our response strategies in dual tasks.

CONCLUSION

In summary, the present study contributes to resolving the debate about the cognitive cost of predictive language processing. The data offer the following key insights:

First, low language predictability as well as high task demands both have a detrimental effect on reading time. This holds true across a large age range. Second, we find that higher task demands diminish the effects of language predictability on reading time, replicably demonstrating that language prediction draws on resources of executive control. Third, the data reveal age-related differences in the use of linguistic predictions: High predictability had more leverage on reading efficiency in older than in younger adults but was also more sensitive to available executive resources.

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 = 44.9 ± 17.9, 18-85 years) with a balanced gender distribution of 51% female, 47% male, and 2% 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; Ishihara, 1987). Exclusion criteria were a history of psychiatric or neurological disorders, drug abuse, dyslexia, illiteracy or any impairments in language processing. 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 [78] 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.

Ethics Statement

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. Participants provided their written informed consent prior to participation, and received financial compensation (12€/h).

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 texts 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 (Fig. 2a). 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. 2a), in which participants were instructed to press a target button on the keyboard ("C" for right-handed, "M" 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. 2a), 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.

All experiments were implemented using *lab.js* [79]. Online studies were hosted on *OpenLab* – a server-side platform designed specifically for *lab.js* experiments [80] – and data were saved on *OSF* [81].

Generation of Word Surprisal and Entropy Scores

The predictability of each word was operationalised via word surprisal (Fig. 2b), which reflects the predictability of the current word given its preceding context [10,43,44]. A word's surprisal score is defined as the negative logarithm of the word's probability given its context [43]:

If a word w_n has a high surprisal score, its occurrence given its preceding context $w_1, w_2, \ldots, w_{n-1}$ has a low probability, rendering it highly unpredictable (i.e. surprising).

In addition to each word's surprisal, we also computed the entropy of the probability distribution for each predicted word given its context (Fig. 2b), which reflects the uncertainty in predicting the next word. The entropy is defined as the negative sum of the product of the probability of each word in the vocabulary and its respective logarithm probability [82], or – put simply – as the average surprisal of all possible continuations in the vocabulary [83]:

If entropy is low, only one or a few possible words in the vocabulary are assigned high probabilities of being the actual next word, hence indicating low uncertainty about which word will come next. This is usually the case if the previous context is very restricting. Conversely, if a vast amount of words in the vocabulary would be suited as continuations for the given context, and the probability distribution across words is fairly uniform, word entropy – and thus uncertainty about which word will come next –

is high. Taken together, word surprisal signifies the predictability of each word whereas word entropy signifies the uncertainty underlying its prediction.

In the current study, we computed surprisal scores as well as one entropy score for each word in the experimental texts (mean surprisal: 18.165 ± 7.523, mean entropy: 4.067 ± 0.932; see further information in the supplementary material). We chose a context length of two words, as constraining the context has been shown to increase GPT-2's psychometric predictive power, making its next-word predictions more human-like [45].

To generate word entropy and surprisal scores, we used a 12-layered GPT-2 model [42], 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 [84]. Scores were calculated using Python version 3.10.12 [85].

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 nth 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. This cutoff was selected arbitrarily, based on the observation that participants tested in the lab did not exhibit trial durations exceeding 5000 ms. Therefore, we assume that participants tested online may have been distracted and less focused on the experiment during trial durations of this length.

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 time: First, reading times were transformed using the Proportion of Maximum Scaling method (POMS; [86]). We POMS-transformed the data on block level to account for potential differences in reading time distributions between blocks. The square root of each value was then taken to ensure a symmetric

distribution. 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 [87,88]. 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 both in n-back task performance – as indicated by reduced d-primes – and the 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 as well as their 2-way and 3-way interactions on log-transformed reading times using a LMM. The model 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.

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, the first two 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. [89,90]). 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 their 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; [91]), word length, as well as the position of block and trial in the course of the experiment as main effects. Furthermore, we included entropy as a fixed effect to account for the uncertainty in the prediction of the next word. Surprisal and entropy values were weakly correlated ($r = 0.29$, $p < 0.001$). To account for the delay in reaction time associated with n-back responses, we included n-back reaction as a binary predictor in our models.

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.

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 + entropy +
word frequency + word length (without punctuation) +
n-back reaction + block number + trial number +
surprisal * age * cognitive load +
(1 + cognitive load | ID) + (1 | text number) + (1 | word) + (1 | colour)

Note. Model structure. 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. This analysis allows exploring the interaction of two continuous predictors, in our case quantifying the slope of the surprisal effect 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, including control analyses and the internal replication, *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 and GAM [92,93]. All effects reported are significant on an alpha level of 0.05 after FDR-correction for multiple comparisons [94].

All analyses were carried out in *R* version 4.2.2 [95] using the packages *gratia*, *interactions, lmerTest, lme4, mgcv*, *modelbased,* and *sjPlot [96–101]*.

Control Analysis: Dissociating Cognitive Control from Attention

To disentangle attentional and cognitive load effects, we modelled reading times using an additional linear mixed model of the same structure as described before but contrasting only the 2-back Dual Task condition with the less demanding 1-back Dual Task condition. The two Dual Task conditions only differ in cognitive load, but not attentional switching costs, which means any effects of cognitive load can be attributed to the cognitive load manipulation, with attentional demands held constant.

Control Analysis: Assessing Potential Effects of Verbal Intelligence and Education

To ensure potential effects of verbal intelligence or education did not unduly influence our findings, we analysed data from 95 lab participants who reported their formal education in years (M = 18.0 ± 3.343, range = 11–30) and completed a lexical decision task – the Spot-the-Word test [102] –, where they were asked to identify the word in pairs of words and non-words. Each participant's score on this test (M = 32.021 ± 2.993, range = 21–37) provided a measure of their verbal intelligence.

To assess the potential effect of education and verbal intelligence on reading time, we fitted three additional LMMs: The first model mirrored the structure of the original LMM used to analyse log-transformed reading times (see section *Statistical Analysis of Reading Times* for the model structure), with one key modification: The predictor for recording location was excluded, as all participants were tested in a single location. The remaining two models followed the same structure, with the inclusion of centred education scores and centred Spot-the-Word test scores as additional predictors to account for education and verbal intelligence, respectively. We then statistically compared the baseline LMM with the two other LMMs using an ANOVA to determine whether verbal intelligence or education significantly improved the model fits.

Control Analysis: Modelling Non-Linear Effects of Age

We also fitted a generalised additive model (GAM) to our data to allow for non-linear relationships of the predictor variables with reading time, as it has been shown that reaction time is oftentimes modulated by predictor variables in a non-linear way [103,104].

Specifically, we employed a GAM to account for a potential non-linear relationship of age and surprisal with reading time. To this end, all continuous predictors were fitted with thin-plate regression splines and the interaction of surprisal, age, and cognitive load was fitted via a tensor product smooth with individual curves for each level of cognitive load. The number of basis dimensions for each smoothing spline was checked via model diagnostics available in *mgcv* after the first model set-up and appropriately updated to reach a k-index > 1.01 and $p > 0.05$ to avoid oversmoothing. The random effect structure was set up similarly to the LMM.

It is important to note that the outcomes of GAMs and LMMs can differ: GAMs are particularly adept at identifying localized, nonlinear changes in predictor effects on reading time that may be overlooked by LMMs. As a result, the effects obtained from LMMs and those derived from GAMs are based on distinct metrics, which complicates direct comparisons between the two approaches.

Internal Replication

To ensure the reliability of our findings, we conducted an internal replication of the previously described experiment. This internal replication was preregistered on OSF (doi: 10.17605/OSF.IO/SU6VX).

Replication Sample

As outlined in the preregistration, we conducted an online study with a sample of 100 participants. We excluded data from four participants from further analysis, either due to technical issues (N = 2) or because they reported having been distracted during the task (N = 2). The resulting final sample comprised 96 participants aged 18 to 70 years (M = 39.750 ± 13.996 years) with a balanced gender distribution of 51% female, 48% male, and 1% non-binary identifying participants. As in the original experiment, all participants were native German speakers with normal or corrected-to-normal vision and intact colour vision without 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.

Ethics Statement

The study was conducted in accordance with the Declaration of Helsinki and was approved by the local ethics committee of the University of Lübeck. All participants were recruited online through Prolific and only tested upon having provided their written informed consent. Participants received a financial compensation of 12€/h.

Replication Analyses

Analogous to the original experiment, we first cleaned the reading time data of trials exceeding a duration of 5000 ms as well as outliers (see section *Preprocessing*), which affected 4.036% of all trials from the main blocks with an average of 72.656 ± 11.089 excluded trials per participant. The structure of the statistical model for the replication analysis was analogous to the model for the original analysis of reading times (see section *Statistical Analysis of Reading Times*), except for the predictor of recording location, which was excluded as we only analysed data collected online. To compare results from the original experiment and the replication, we fitted the model once using the data of the online sample from the original experiment and once using the new online replication datasets.

Given that we simplified the analysis approach in the original study after having preregistered the replication, we deviated from the analysis plan described in the preregistration and made the same modifications here, resulting in the use of word surprisal for only one context length instead of four and, consequently, only one LMM instead of several.

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

M.S.: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, visualisation, writing: original draft, review & editing

S.M.: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, writing: original draft, review & editing

S.T.: formal analysis, methodology, writing: review & editing

L-M.S.: methodology, writing: review & editing

G.H.: Conceptualization, funding acquisition, supervision, methodology, writing: review & editing

J.O.: Conceptualization, methodology, funding acquisition, supervision, writing: review & editing

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

During the preparation of this work the authors used ChatGPT 3.5 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 Availability

All experimental and analysis scripts are publicly available in the project's Github repository https://github.com/MMarieSchuckart/EXNAT-1. Raw data as well as model outputs, results, and model comparisons will be made available upon publication in the project's OSF repository: https://osf.io/dtwjk/.

Preregistration

The internal replication of the main experiment was preregistered on OSF (doi: 10.17605/OSF.IO/SU6VX).

REFERENCES

1. Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci [Internet]. 2013 Jun;36(3):181–204. Available from: http://dx.doi.org/10.1017/S0140525X12000477

2. Pickering MJ, Garrod S. Do people use language production to make predictions during comprehension? Trends Cogn Sci [Internet]. 2007 Mar;11(3):105–10. Available from: http://dx.doi.org/10.1016/j.tics.2006.12.002

3. Onnis L, Lim A, Cheung S, Huettig F. Is the mind inherently predicting? Exploring forward and backward looking in language processing. Cogn Sci [Internet]. 2022 Oct;46(10):e13201. Available from: http://dx.doi.org/10.1111/cogs.13201

4. Rao RP, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci [Internet]. 1999 Jan;2(1):79–87. Available from: http://dx.doi.org/10.1038/4580

5. Ryskin R, Nieuwland MS. Prediction during language comprehension: what is next? Trends Cogn Sci [Internet]. 2023 Nov;27(11):1032–52. Available from: https://doi.org/10.1016/j.tics.2023.08.003

6. Rubenstein H, Aborn M. Learning, prediction, and readability. J Appl Psychol [Internet]. 1958 Feb;42(1):28–32. Available from: http://dx.doi.org/10.1037/h0039808

7. Boston MF, Hale J, Kliegl R, Patil U, Vasishth S. Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. J Eye Mov Res [Internet]. 2008 Sep 8;2(1). Available from: http://dx.doi.org/10.16910/jemr.2.1.1

8. Roland D, Yun H, Koenig JP, Mauner G. Semantic similarity, predictability, and models of sentence processing. Cognition [Internet]. 2012 Mar;122(3):267–79. Available from: http://dx.doi.org/10.1016/j.cognition.2011.11.011

9. Schmitt LM, Erb J, Tune S, Rysop AU, Hartwigsen G, Obleser J. Predicting speech from a cortical hierarchy of event-based time scales. Sci Adv [Internet]. 2021 Dec 3;7(49):eabi6070. Available from: http://dx.doi.org/10.1126/sciadv.abi6070

10. Smith NJ, Levy R. The effect of word predictability on reading time is logarithmic. Cognition [Internet]. 2013 Sep;128(3):302–19. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0010027713000413

11. Luke SG, Christianson K. Limits on lexical prediction during reading. Cogn Psychol [Internet]. 2016 Aug;88:22–60. Available from: http://dx.doi.org/10.1016/j.cogpsych.2016.06.002

12. Geranmayeh F, Chau TW, Wise RJS, Leech R, Hampshire A. Domain-general subregions of the medial prefrontal cortex contribute to recovery of language after stroke. Brain [Internet]. 2017 Jul 1;140(7):1947–58. Available from: http://dx.doi.org/10.1093/brain/awx134

13. Martin S, Saur D, Hartwigsen G. Age-dependent contribution of domain-general networks to semantic cognition. Cereb Cortex [Internet]. 2022 Feb 8;32(4):870–90. Available from: http://dx.doi.org/10.1093/cercor/bhab252

14. Sliwinska MW, Violante IR, Wise RJS, Leech R, Devlin JT, Geranmayeh F, et al. Stimulating multiple-demand cortex enhances vocabulary learning. J Neurosci [Internet]. 2017 Aug 9;37(32):7606–18. Available from: http://dx.doi.org/10.1523/jneurosci.3857-16.2017

15. Wingfield A, Grossman M. Language and the aging brain: patterns of neural compensation revealed by functional brain imaging. J Neurophysiol [Internet]. 2006 Dec;96(6):2830–9. Available from: http://dx.doi.org/10.1152/jn.00628.2006

16. Erb J, Henry MJ, Eisner F, Obleser J. The brain dynamics of rapid perceptual adaptation to adverse listening conditions. J Neurosci [Internet]. 2013 Jun 26;33(26):10688–97. Available from: http://dx.doi.org/10.1523/JNEUROSCI.4596-12.2013

17. Vaden KI Jr, Kuchinsky SE, Cute SL, Ahlstrom JB, Dubno JR, Eckert MA. The cingulo-opercular network provides word-recognition benefit. J Neurosci [Internet]. 2013 Nov 27;33(48):18979–86. Available from: https://doi.org/10.1523/JNEUROSCI.1417-13.2013

18. Vaden KI Jr, Kuchinsky SE, Ahlstrom JB, Dubno JR, Eckert MA. Cortical activity predicts which older adults recognize speech in noise and when. J Neurosci [Internet]. 2015 Mar 4;35(9):3929–37. Available from: http://dx.doi.org/10.1523/JNEUROSCI.2908-14.2015

19. Vaden KI Jr, Kuchinsky SE, Ahlstrom JB, Teubner-Rhodes SE, Dubno JR, Eckert MA. Cingulo-opercular function during word recognition in noise for older adults with hearing loss. Exp Aging Res [Internet]. 2016;42(1):67–82. Available from: http://dx.doi.org/10.1080/0361073X.2016.1108784

20. Peelle JE. Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. Ear Hear [Internet]. 2018;39(2):204–14. Available from: http://dx.doi.org/10.1097/AUD.0000000000000494

21. Alvarez JA, Emory E. Executive function and the frontal lobes: a meta-analytic review. Neuropsychol Rev [Internet]. 2006 Mar;16(1):17–42. Available from: http://dx.doi.org/10.1007/s11065-006-9002-x

22. Duncan J. The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. Trends Cogn Sci [Internet]. 2010 Apr;14(4):172–9. Available from: https://doi.org/10.1016/j.tics.2010.01.004

23. Friedman NP, Miyake A. Unity and diversity of executive functions: Individual differences as a window on cognitive structure. Cortex [Internet]. 2017 Jan;86:186–204. Available from: http://dx.doi.org/10.1016/j.cortex.2016.04.023

24. Ryskin R, Levy RP, Fedorenko E. Do domain-general executive resources play a role in linguistic prediction? Re-evaluation of the evidence and a path forward. Neuropsychologia [Internet]. 2020 Jan;136(107258):107258. Available from: https://doi.org/10.1016/j.neuropsychologia.2019.107258

25. Shain C, Blank IA, van Schijndel M, Schuler W, Fedorenko E. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. Neuropsychologia [Internet]. 2020 Feb 17;138(107307):107307. Available from: http://dx.doi.org/10.1016/j.neuropsychologia.2019.107307

26. Wehbe L, Blank IA, Shain C, Futrell R, Levy R, von der Malsburg T, et al. Incremental language comprehension difficulty predicts activity in the language network but not the multiple demand

network. Cereb Cortex [Internet]. 2021 Jul 29;31(9):4006–23. Available from: https://doi.org/10.1093/cercor/bhab065

27. Diachek E, Blank I, Siegelman M, Affourtit J, Fedorenko E. The domain-general multiple demand (MD) network does not support core aspects of language comprehension: A large-scale fMRI investigation. J Neurosci [Internet]. 2020 Jun 3;40(23):4536–50. Available from: http://dx.doi.org/10.1523/JNEUROSCI.2036-19.2020

28. MacGregor LJ, Gilbert RA, Balewski Z, Mitchell DJ, Erzinçlioğlu SW, Rodd JM, et al. Causal contributions of the domain-general (multiple demand) and the language-selective brain networks to perceptual and semantic challenges in speech comprehension. Neurobiol Lang (Camb) [Internet]. 2022 Dec 16;3(4):665–98. Available from: http://dx.doi.org/10.1162/nol_a_00081

29. Shain C, Blank IA, Fedorenko E, Gibson E, Schuler W. Robust effects of working memory demand during naturalistic language comprehension in language-selective cortex. J Neurosci [Internet]. 2022 Sep 28;42(39):7412–30. Available from: https://www.jneurosci.org/content/42/39/7412.abstract

30. Xie Z, Brodbeck C, Chandrasekaran B. Cortical tracking of continuous speech under bimodal divided attention. Neurobiol Lang (Camb) [Internet]. 2023 Apr 11;4(2):318–43. Available from: http://dx.doi.org/10.1162/nol_a_00100

31. Idowu MI, Szameitat AJ. Executive function abilities in cognitively healthy young and older adults-A cross-sectional study. Front Aging Neurosci [Internet]. 2023 Feb 8;15:976915. Available from: http://dx.doi.org/10.3389/fnagi.2023.976915

32. Salthouse TA, Atkinson TM, Berish DE. Executive functioning as a potential mediator of age-related cognitive decline in normal adults. J Exp Psychol Gen [Internet]. 2003 Dec;132(4):566–94. Available from: http://dx.doi.org/10.1037/0096-3445.132.4.566

33. Verhaeghen P, Steitz DW, Sliwinski MJ, Cerella J. Aging and dual-task performance: a meta-analysis. Psychol Aging [Internet]. 2003 Sep;18(3):443–60. Available from: http://dx.doi.org/10.1037/0882-7974.18.3.443

34. Liu X, Chin J, Payne BR, Fu WT, Morrow DG, Stine-Morrow EAL. Adult age differences in information foraging in an interactive reading environment. Psychol Aging [Internet]. 2016 May;31(3):211–23. Available from: http://dx.doi.org/10.1037/pag0000079

35. Smiler AP, Gagne DD, Stine-Morrow EAL. Aging, memory load, and resource allocation during reading. Psychol Aging [Internet]. 2003 Jun;18(2):203–9. Available from: http://dx.doi.org/10.1037/0882-7974.18.2.203

36. Häuser KI, Demberg V, Kray J. Effects of aging and dual-task demands on the comprehension of less expected sentence continuations: Evidence from pupillometry. Front Psychol [Internet]. 2019 Mar 29;10:709. Available from: http://dx.doi.org/10.3389/fpsyg.2019.00709

37. Cheimariou S, Farmer TA, Gordon JK. Lexical prediction in the aging brain: The effects of predictiveness and congruency on the N400 ERP component. Neuropsychol Dev Cogn B Aging Neuropsychol Cogn [Internet]. 2019 Sep;26(5):781–806. Available from: http://dx.doi.org/10.1080/13825585.2018.1529733

38. Dave S, Brothers TA, Swaab TY. 1/f neural noise and electrophysiological indices of contextual prediction in aging. Brain Res [Internet]. 2018 Jul 15;1691:34–43. Available from: http://dx.doi.org/10.1016/j.brainres.2018.04.007

39. Payne BR, Federmeier KD. Contextual constraints on lexico-semantic processing in aging: Evidence from single-word event-related brain potentials. Brain Res [Internet]. 2018 May;1687:117–28. Available from: http://dx.doi.org/10.1016/j.brainres.2018.02.021

40. Reuter-Lorenz PA, Cappell KA. Neurocognitive aging and the compensation hypothesis. Curr Dir Psychol Sci [Internet]. 2008 Jun;17(3):177–82. Available from: http://dx.doi.org/10.1111/j.1467-8721.2008.00570.x

41. Shafto MA, Tyler LK. Language in the aging brain: the network dynamics of cognitive decline and preservation. Science [Internet]. 2014 Oct 31;346(6209):583–7. Available from: http://dx.doi.org/10.1126/science.1254404

42. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners [Internet]. 2019. Available from: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

43. Hale J. A probabilistic earley parser as a psycholinguistic model. In: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001 - NAACL '01 [Internet]. Morristown, NJ, USA: Association for Computational Linguistics; 2001. Available from: http://dx.doi.org/10.3115/1073336.1073357

44. Levy R. Expectation-based syntactic comprehension. Cognition [Internet]. 2008 Mar;106(3):1126–77. Available from: http://dx.doi.org/10.1016/j.cognition.2007.05.006

45. Kuribayashi T, Oseki Y, Brassard A, Inui K. Context limitations make neural language models more human-like [Internet]. arXiv [cs.CL]. 2022. Available from: http://dx.doi.org/10.48550/ARXIV.2205.11463

46. Johnson PO, Neyman J. Tests of certain linear hypotheses and their application to some educational problems. Statistical Research Memoirs. 1936;1:57–93.

47. Funder DC, Ozer DJ. Evaluating effect size in psychological research: Sense and nonsense. Adv Methods Pract Psychol Sci [Internet]. 2019 Jun;2(2):156–68. Available from: https://doi.org/10.1177/2515245919847202

48. Tolstoy L. Anna Karenina [Internet]. 1878. Available from: https://www.gutenberg.org/cache/epub/1399/pg1399-images.html

49. Lamichhane B, Westbrook A, Cole MW, Braver TS. Exploring brain-behavior relationships in the N-back task. Neuroimage [Internet]. 2020 May 15;212(116683):116683. Available from: http://dx.doi.org/10.1016/j.neuroimage.2020.116683

50. Kwong See ST, Ryan EB. Cognitive mediation of adult age differences in language performance. Psychol Aging [Internet]. 1995 Sep;10(3):458–68. Available from: http://dx.doi.org/10.1037/0882-7974.10.3.458

51. Heilbron M, Armeni K, Schoffelen JM, Hagoort P, de Lange FP. A hierarchy of linguistic predictions during natural language comprehension. Proc Natl Acad Sci U S A [Internet]. 2022 Aug 9;119(32):e2201968119. Available from: http://dx.doi.org/10.1073/pnas.2201968119

52. Monsalve IF, Frank SL, Vigliocco G. Lexical surprisal as a general predictor of reading time. In: Proceedings of the 13th Conference of the European Chapter. France: Association for Computational Linguistics; 2012. p. 398–408.

53. Shain C, Meister C, Pimentel T, Cotterell R, Levy R. Large-scale evidence for logarithmic effects of word predictability on reading time. Proc Natl Acad Sci U S A [Internet]. 2024 Mar 5;121(10):e2307876121. Available from: http://dx.doi.org/10.1073/pnas.2307876121

54. Wilcox EG, Pimentel T, Meister C, Cotterell R, Levy RP. Testing the predictions of surprisal theory in 11 languages. Trans Assoc Comput Linguist [Internet]. 2023 Dec 14;11:1451–70. Available from: http://dx.doi.org/10.1162/tacl_a_00612

55. Bubic A, von Cramon DY, Schubotz RI. Prediction, cognition and the brain. Front Hum Neurosci [Internet]. 2010 Mar 22;4:25. Available from: http://dx.doi.org/10.3389/fnhum.2010.00025

56. Friston K. The free-energy principle: a unified brain theory? Nat Rev Neurosci [Internet]. 2010 Feb;11(2):127–38. Available from: http://dx.doi.org/10.1038/nrn2787

57. Larsson J, Smith AT. fMRI repetition suppression: neuronal adaptation or stimulus expectation? Cereb Cortex [Internet]. 2012 Mar;22(3):567–76. Available from: http://dx.doi.org/10.1093/cercor/bhr119

58. Richter D, de Lange FP. Statistical learning attenuates visual activity only for attended stimuli. Elife [Internet]. 2019 Aug 23;8. Available from: http://dx.doi.org/10.7554/eLife.47869

59. Huettig F, Janse E. Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. Lang Cogn Neurosci [Internet]. 2016 Jan 2;31(1):80–93. Available from: http://dx.doi.org/10.1080/23273798.2015.1047459

60. Ito A, Corley M, Pickering MJ. A cognitive load delays predictive eye movements similarly during L1 and L2 comprehension. Biling (Camb Engl) [Internet]. 2018 Mar;21(2):251–64. Available from: http://dx.doi.org/10.1017/s1366728917000050

61. Fricke M, Zirnstein M. Predictive processing and inhibitory control drive semantic enhancements for non-dominant language word recognition in noise. Languages [Internet]. 2022 Sep 15;7(3):239. Available from: http://dx.doi.org/10.3390/languages7030239

62. Borovsky A, Elman JL, Fernald A. Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. J Exp Child Psychol [Internet]. 2012 Aug;112(4):417–36. Available from: http://dx.doi.org/10.1016/j.jecp.2012.01.005

63. Huettig F, Pickering MJ. Literacy advantages beyond reading: Prediction of spoken language. Trends Cogn Sci [Internet]. 2019 Jun;23(6):464–75. Available from: http://dx.doi.org/10.1016/j.tics.2019.03.008

64. Mainz N, Shao Z, Brysbaert M, Meyer AS. Vocabulary knowledge predicts lexical processing: Evidence from a group of participants with diverse educational backgrounds. Front Psychol [Internet]. 2017 Jul 13;8:1164. Available from: http://dx.doi.org/10.3389/fpsyg.2017.01164

65. Matthews J. Vocabulary for listening: Emerging evidence for high and mid-frequency vocabulary knowledge. System [Internet]. 2018 Feb;72:23–36. Available from: http://dx.doi.org/10.1016/j.system.2017.10.005

66. Stæhr LS. Vocabulary knowledge and advanced listening comprehension in English as a foreign language. Stud Second Lang Acquis [Internet]. 2009 Dec;31(04):577. Available from: http://dx.doi.org/10.1017/s0272263109990039

67. Stæhr LS. Vocabulary size and the skills of listening, reading and writing. Lang Learn J [Internet]. 2008 Dec;36(2):139–52. Available from: http://dx.doi.org/10.1080/09571730802389975

68. Brysbaert M, Stevens M, Mandera P, Keuleers E. 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. Front Psychol [Internet]. 2016 Jul 29;7:1116. Available from: http://dx.doi.org/10.3389/fpsyg.2016.01116

69. Ito A, Sakai H. Everyday language exposure shapes prediction of specific words in listening comprehension: A visual world eye-tracking study. Front Psychol [Internet]. 2021 Feb 9;12:607474. Available from: http://dx.doi.org/10.3389/fpsyg.2021.607474

70. Bunzeck N, Steiger TK, Krämer UM, Luedtke K, Marshall L, Obleser J, et al. Trajectories and contributing factors of neural compensation in healthy and pathological aging. Neurosci Biobehav Rev [Internet]. 2024 Jan;156(105489):105489. Available from: http://dx.doi.org/10.1016/j.neubiorev.2023.105489

71. Reuter-Lorenz PA, Festini SB, Jantz TK. Executive functions and neurocognitive aging. In: Handbook of the Psychology of Aging [Internet]. Elsevier; 2021. p. 67–81. Available from: http://dx.doi.org/10.1016/b978-0-12-816094-7.00019-2

72. Burke DM, Mackay DG. Memory, language, and ageing. Philos Trans R Soc Lond B Biol Sci [Internet]. 1997 Dec 29;352(1363):1845–56. Available from: http://dx.doi.org/10.1098/rstb.1997.0170

73. Chan JS, Wibral M, Stawowsky C, Brandl M, Helbling S, Naumer MJ, et al. Predictive coding over the lifespan: Increased reliance on perceptual priors in older adults-A magnetoencephalography and dynamic causal modeling study. Front Aging Neurosci [Internet]. 2021 Apr 9;13:631599. Available from: http://dx.doi.org/10.3389/fnagi.2021.631599

74. Moran RJ, Symmonds M, Dolan RJ, Friston KJ. The brain ages optimally to model its environment: Evidence from sensory learning over the adult lifespan. PLoS Comput Biol [Internet]. 2014 Jan 23;10(1):e1003422. Available from: http://dx.doi.org/10.1371/journal.pcbi.1003422

75. Rogers CS. Semantic priming, not repetition priming, is to blame for false hearing. Psychon Bull Rev [Internet]. 2017 Aug;24(4):1194–204. Available from: http://dx.doi.org/10.3758/s13423-016-1185-4

76. Wingfield A, Amichetti NM, Lash A. Cognitive aging and hearing acuity: modeling spoken language comprehension. Front Psychol [Internet]. 2015 Jun 11;6:684. Available from: http://dx.doi.org/10.3389/fpsyg.2015.00684

77. Wolpe N, Ingram JN, Tsvetanov KA, Geerligs L, Kievit RA, Henson RN, et al. Ageing increases reliance on sensorimotor prediction through structural and functional differences in frontostriatal circuits. Nat Commun [Internet]. 2016 Oct 3;7(1):13034. Available from: http://dx.doi.org/10.1038/ncomms13034

78. Folstein MF, Folstein SE, McHugh PR. Mini-Mental State Examination [Internet]. PsycTESTS Dataset. American Psychological Association (APA); 2014. Available from: http://dx.doi.org/10.1037/t07757-000

79. Henninger F, Shevchenko Y, Mertens U, Kieslich PJ, Hilbig BE. lab.js: A free, open, online experiment builder [Internet]. Zenodo; 2024. Available from: http://dx.doi.org/10.5281/ZENODO.597045

80. Shevchenko Y. Open Lab: A web application for running and sharing online experiments. Behav Res Methods [Internet]. 2022 Dec;54(6):3118–25. Available from: http://dx.doi.org/10.3758/s13428-021-01776-2

81. Foster ED, Deardorff A. Open Science Framework (OSF). J Med Libr Assoc [Internet]. 2017 Apr 4;105(2). Available from: http://dx.doi.org/10.5195/jmla.2017.88

82. Shannon CE. A mathematical theory of communication. Bell Syst Tech J [Internet]. 1948 Jul;27(3):379–423. Available from: http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x

83. Slaats S, Martin AE. What's surprising about surprisal [Internet]. PsyArXiv. 2023. Available from: http://dx.doi.org/10.31234/osf.io/7pvau

84. Schweter S. dbmdz/german-gpt2 [Internet]. Hugging Face. 2021 [cited 2023 May 16]. Available from: https://huggingface.co/dbmdz/german-gpt2

85. Van Rossum G, Drake FL Jr. Python 3 Reference Manual. Createspace; 2009. 244 p.

86. Little TD. Longitudinal Structural Equation Modeling (Methodology in the Social Sciences). The GuilfordPress; 2013.

87. Berger A, Kiefer M. Comparison of different response time outlier exclusion methods: A simulation study. Front Psychol [Internet]. 2021 Jun 14;12:675558. Available from: http://dx.doi.org/10.3389/fpsyg.2021.675558

88. Cousineau D, Chartier S. Outliers detection and treatment: a review. Int J Psychol Res (Medellin) [Internet]. 2010 Jun 30;3(1):58–67. Available from: http://dx.doi.org/10.21500/20112084.844

89. Tune S, Alavash M, Fiedler L, Obleser J. Neural attentional-filter mechanisms of listening success in middle-aged and older individuals. Nat Commun [Internet]. 2021 Jul 26;12(1):4533. Available from: http://dx.doi.org/10.1038/s41467-021-24771-9

90. Bell A, Fairbrother M, Jones K. Fixed and random effects models: making an informed choice. Qual Quant [Internet]. 2019 Mar;53(2):1051–74. Available from: http://dx.doi.org/10.1007/s11135-018-0802-x

91. Speer R, Chin J, Lin A, Jewett S, Nathan L. LuminosoInsight/wordfreq: v2.2 [Internet]. Zenodo; 2018. Available from: http://dx.doi.org/10.5281/ZENODO.1443582

92. Luke SG. Evaluating significance in linear mixed-effects models in R. Behav Res Methods [Internet]. 2017 Aug;49(4):1494–502. Available from: http://dx.doi.org/10.3758/s13428-016-0809-y

93. Satterthwaite FE. An approximate distribution of estimates of variance components. Biom Bull [Internet]. 1946 Dec;2(6):110. Available from: http://dx.doi.org/10.2307/3002019

94. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol [Internet]. 1995 Jan 1;57(1):289–300. Available from: http://dx.doi.org/10.1111/j.2517-6161.1995.tb02031.x

95. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2023. Available from: https://www.R-project.org/

96. Simpson GL. Gratia: An R package for exploring generalized additive models [Internet]. arXiv [stat.CO]. 2024. Available from: http://arxiv.org/abs/2406.19082

97. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models Usinglme4. J Stat Softw [Internet]. 2015;67(1). Available from: http://dx.doi.org/10.18637/jss.v067.i01

98. Kuznetsova A, Brockhoff PB, Christensen RHB. LmerTest package: Tests in linear mixed effects models. J Stat Softw [Internet]. 2017;82(13). Available from: http://dx.doi.org/10.18637/jss.v082.i13

99. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J R Stat Soc Series B Stat Methodol [Internet]. 2011 Jan 1;73(1):3–36. Available from: http://dx.doi.org/10.1111/j.1467-9868.2010.00749.x

100. Makowski D, Ben-Shachar MS, Patil I, Lüdecke D. Estimation of Model-Based Predictions, Contrasts and Means [Internet]. CRAN. 2020. Available from: https://github.com/easystats/modelbased

101. Long JA. interactions: Comprehensive, User-Friendly Toolkit for Probing Interactions [Internet]. 2024. Available from: http://dx.doi.org/10.32614/CRAN.package.interactions

102. Baddeley A, Emslie H, Nimmo-Smith I. The Spot-the-Word test: a robust estimate of verbal intelligence based on lexical decision. Br J Clin Psychol [Internet]. 1993 Feb;32(1):55–65. Available from: http://dx.doi.org/10.1111/j.2044-8260.1993.tb01027.x

103. Miwa K, Baayen H. Nonlinearities in bilingual visual word recognition: An introduction to generalized additive modeling. Biling (Camb Engl) [Internet]. 2021 Nov;24(5):825–32. Available from: http://dx.doi.org/10.1017/s1366728921000079

104. Wood SN. Generalized additive models [Internet]. Chapman and Hall/CRC; 2017. Available from: http://dx.doi.org/10.1201/9781315370279

105. Bamberger R, Vanecek E. Lesen - Verstehen - Lernen - Schreiben: Die Schwierigkeitsstufen von Texten in deutscher Sprache. 1984.

List of Legends for the Supplementary Materials

Figure S1. Comparison of age distribution between samples.

Figure S2. Task performance and reading time by age and cognitive load condition. Task performance (d-primes) in conditions with an n-back task (a). Accuracy in the comprehension question performance task (b). Reading Time by age and condition (c). Reading time by condition (d). Solid line: M, shaded area: 95% CI, point: mean reading time for one participant in the respective condition.

Figure S3. Individual predicted reading time. Reading time by age (a) and condition (b).

Figure S4. Comparison of factor smooths for different levels of cognitive load from the three-way interaction of age, surprisal, and cognitive load. The difference smooths show slightly stronger effects of high surprisal in young than older adults for the 1-back relative to the 2-back condition, and stronger effects of high surprisal in older adults for the Reading only relative to the n-back condition.

Figure S5. Comparison of the results of the LMM and GAM control analyses. Panel **c** illustrates the interaction between cognitive load and surprisal for a representative younger and older participant, estimated using the LMM (left) and the GAM (right). For a complementary visualisation of the three-way interaction between age, cognitive load, and surprisal, see Fig. S4. For a visualisation of the main effects of age, surprisal and cognitive load on reading time, please see Fig. 3c.

Figure S6. Three-way interaction of age, surprisal, and cognitive load in the replication sample.

Table S1. Results from models for task performance measures (N = 175).

Table S2. Results from models for reading time for full original sample (N = 175).

Table S3. Results from models for reading time for original online sample and online replication sample (N = 80 and N = 96, respectively).

Table S4. Results from models for control analysis (1-back vs. 2-back) of reading time for full original sample (N = 175).

Table S5. Results from GAM for control analysis of reading time for full original sample (N = 175).