UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Haeuser, K. I., Kray, J., & Borovsky, A. (2022). Hedging Bets in Linguistic Prediction: Younger and Older Adults Vary in the Breadth of Predictive Processing. *Collabra: Psychology*, *8*(1). https://doi.org/10.1525/collabra.36945

Developmental Psychology

Hedging Bets in Linguistic Prediction: Younger and Older Adults Vary in the Breadth of Predictive Processing

Katja I. Haeuser[1] [a], Jutta Kray[1], Arielle Borovsky[2]

[1] Department of Psychology, Saarland University, Germany; Collaborative Research Center Information Density and Linguistic Encoding (SFB 1102), Saarland University, Germany, [2] College of Health and Human Sciences, Purdue University, Purdue University, Lafayette, IN, USA

Keywords: sentence comprehension, reading, aging, prediction

https://doi.org/10.1525/collabra.36945

Collabra: Psychology

Vol. 8, Issue 1, 2022

Language processing is predictive in nature, but it is unknown whether language users generate multiple predictions about upcoming content simultaneously or whether spreading activation from one pre-activated word facilitates other words downstream. Simultaneously, developmental accounts of predictive processing simultaneously highlight potential tension among spreading activation vs. multiple activation accounts. We used self-paced reading to investigate if younger and older readers of German generate (multiple) graded predictions about the grammatical gender of nouns. Gradedness in predictions was operationalized as the difference in cloze probability between the most likely and second-most likely continuation that could complete a sentence. Sentences with a greater probabilistic difference were considered as imbalanced and more biased towards one gender. Sentences with lower probabilistic differences were considered to be more balanced towards multiple genders.
Both young and older adults engaged in predictive processing. However, only younger adults activated multiple predictions, with slower reading times (RTs) when gender representations were balanced, but facilitation when one gender was more likely than others. In contrast, older adults' RTs did not pattern with imbalance but merely with predictability, showing that, while able to generate predictions based on context, older adults did not predict multiple gender continuations. Hence, our findings suggest that (younger) language users generate graded predictions about upcoming content, by weighing possible sentence continuations according to their difference in cloze probability. Compared to younger adults, older adults' predictions are reduced in scope. The results provide novel theoretical insights into the developmental mechanisms involved in predictive processing.

Introduction

Language comprehension is fast. In normal, everyday language use, humans process two or three words per second. How do people manage this feat? Prediction is one cognitive mechanism that supports this process. Language users rapidly integrate prior information to anticipate or predict upcoming linguistic structures based on prior context. Consider the example *Gary believes a husband should not cheat on his* ... Here, several sentence completions are possible. Most people will strongly expect a completion like *wife*, while other continuations (*spouse, taxes*) – even though they might be plausible – are less likely. Importantly, as language users accrue experience and knowledge, the weightings of these expectations may change. In the current study, we explore whether and how older and younger adults generate multiple, graded expectations for upcoming words in language comprehension.

Specifically, we aim to evaluate two contrasting theoretical accounts regarding the plurality of linguistic prediction. "All or nothing" accounts of prediction argue that the parser entertains predominantly one option, the one with the highest cloze probability in prior ratings, and once that is disconfirmed, re-analyzes and starts anew (cf. Kuperberg & Jaeger, 2016). In contrast, graded accounts of prediction posit that the parser entertains multiple options at any given time and maintains these predictions (e.g., Levy, 2008, 2013), perhaps ordered by the strength of their cloze probability, until new information comes in that discredits some predictions and confirms others. In psycholinguistic

a Correspondence concerning this article should be addressed to Katja I. Haeuser, Department of Psychology, Campus A 1.3., room 2.13, Saarland University, 66123 Saarbruecken, Germany.
Email: khaeuser@coli.uni-saarland.de

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

research, it can be challenging to properly adjudicate between these accounts. For example, graded activation effects might also be explained by assuming that spreading activation from an initial single prediction spreads to other (semantically) related options later during processing (cf. evidence from Federmeier & Kutas, 1999).

A secondary goal of the current study is to evaluate whether and how predictive processes change with age. While many electrophysiological studies of aging demonstrate that older adults are less likely to use sentence context predictively (for review, see Wlotko et al., 2010), a range of behavioral and eye-tracking studies have attested to increased, not reduced, context use with age.

We seek to gain insight into these cognitive and developmental mechanisms in the current study by investigating whether younger and older adults of German create graded predictions about upcoming information during sentence reading. To anticipate our results, we find that both age groups use sentence context to predict upcoming words. However, only the predictions of younger adults are graded in nature. We begin by reviewing the evidence for graded prediction in younger and older adults.

Prediction in younger adults

A number of studies have demonstrated that readers or listeners show processing difficulty when encountering prenominal modifiers (e.g., articles or adjectives) whose gender marking does not agree with the gender of the noun that is predictable based on context. Many languages use a gender system to sort nouns into grammatical classes. Articles or adjectives that precede a noun must be gender-marked according to the head noun (e.g., Spanish, "la $_{\text{feminine}}$ roja $_{\text{feminine}}$ canasta $_{\text{feminine}}$"; German "die $_{\text{feminine}}$ rote $_{\text{feminine}}$ Krone $_{\text{feminine}}$", the red crown). Therefore, the gender marking of a definite article can be used as an early, pre-nominal cue that indicates whether the predictable noun will appear or not. Unlike nouns though, gender-marked articles carry little semantic information ("some (singular) thing"; Urbach et al., 2020), so any facilitation effects that occur at the level of the article (e.g., a speed-up in reading times, or a reduced N400 ERP component signaling facilitated processing) arguably constitute strong evidence for prediction as opposed to incremental integration.

In a seminal study Wicha and colleagues (2004) presented sentence frames that created a strong bias for a particular noun and its gender (e.g., "canasta $_{\text{feminine}}$", which requires the definite article "la $_{\text{feminine}}$"), but instead were followed by a prenominal gender-marked article from a different grammatical class (e.g., "el $_{\text{masculine}}$" …). Around 500 ms after presentation of the gender marked article, the EEG record in younger adults (age range = 18 – 31 years) showed a small positive deflection for unpredictable gender-marked articles compared to predictable ones. The authors argued that younger language users use gender information conveyed by gender-marked articles when processing a sentence in real time in order to build sentence meaning.

Converging evidence for pre-nominal prediction in younger adults has been presented in other languages as well, for example Dutch. Van Berkum and colleagues (2005) found that gender-marked adjective inflections (e.g., Dutch "grote" $_{\text{common gender}}$, big) that were inconsistent with the gender of the predicted noun (e.g., "schilderij" $_{\text{neuter}}$ (book case), requiring "groot" $_{\text{neuter}}$, not "grote" $_{\text{common}}$) showed a small positive increase in ERPs of younger adults (age range = 18 – 28) as early as 50ms after inflection onset (for reviews, see Kochari & Flecken, 2019; Nicenboim et al., 2020; but see Nieuwland, 2019, for critical discussion).

However, these experimental findings have not remained unchallenged. For Dutch, prediction effects on gender-marked adjective inflections failed to replicate (Kochari & Flecken, 2019; Nieuwland et al., 2020; young adult age ranges in these papers = 18 – 35 and 18 – 40 years, respectively). For example, Nieuwland and colleagues used Bayes factor analysis to replicate the findings reported by Van Berkum et al. (2005). Unlike the original study, they found that prediction-inconsistent adjective inflections elicited a negativity, but the Bayes factor analysis for this effect yielded neither strong evidence for nor against the null hypothesis. The authors tentatively concluded that it remains to be shown whether Dutch listeners consistently use adjectival inflections to inform their noun predictions. In a Bayesian meta-analysis that pooled the results from several ERP studies on prenominal features, Nicenboim and colleagues (2020) did find evidence for a prediction effect on prenominal articles. But since this effect was very small and only surfaced when multiple studies were combined, the authors concluded that prenominal prediction effects are probably difficult to detect reliably.

An additional question is whether the prediction effects reviewed above were graded in nature. Many prior studies are mute with respect to the gradedness of linguistic predictions, because these studies frequently probed one, and only one, highly probable sentence continuation (i.e., normally the one with the highest cloze probability in prior cloze ratings). Perhaps the strongest evidence attesting to a graded effect for linguistic predictions comes from an EEG study using English indefinite articles. DeLong and colleagues (2005) used sentence frames such as *Hannah wanted to live in a small town, but her husband preferred to live closer to…* which ended in a more or less predictable noun and its indefinite article (*a city*, *an airport*). Like in earlier studies, the indefinite article could be used as an early cue to foreshadow the anticipated noun. Crucially, the design of this study allowed for an investigation of graded prediction because the sentence endings that were presented to participants varied on a continuous spectrum from relatively low to high cloze. According to the results, the size of the N400 at the level of the indefinite article in younger adults (age range = 18 – 37 years) varied inversely with article cloze probability, that is, as article predictability increased gradually, the N400 became systematically smaller. Based on the observed correlation between N400 amplitude and article cloze probability, DeLong et al. (2005) argued that pre-activation is not "all or nothing" but occurs in a graded, probabilistic fashion: The strength of a word's pre-activation is proportional to its cloze probability.

However, subsequent studies obtained only mixed evidence for graded predictability effects in young adult samples (Ito et al., 2017; Martin et al., 2013; Nieuwland et al., 2018; age ranges = 18 – 29, 21 – 27, 18 – 35 years, respectively; but see Urbach et al., 2020; Yan et al., 2017). One of

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

the conclusions drawn in Nieuwland et al. (2018) is that the phonological dependency of indefinite articles in English might not readily lend itself to investigation of predictability effects, because *a* and *an* only need to align with the subsequent word of the sentence, which does not necessarily have to be a noun (e.g., when an adjective follows, as in *an old city*, *a new airport*).

Hence, it seems that there is not only rather mixed evidence with respect to the reliability of pre-nominal effects of prediction in general, but also to the gradedness of linguistic predictions in particular. The rather mixed pattern of findings so far suggests that linguistic pre-activation does not occur at the level of granularity and detail that is often assumed (for discussion, see Huettig & Guerra, 2019; Huettig & Mani, 2016). Of note, only a fraction of prior studies can unequivocally speak to graded effects of prediction (DeLong et al., 2005; Ito et al., 2017; Nieuwland et al., 2018), and the ones that can, obtained conflicting results.

Prediction and aging

Few studies from the cognitive aging literature unambiguously evaluate linguistic prediction, since most prior aging studies measured predictive effects at the moment that a predicted word is presented, which confounds prediction with semantic integration processes (Otten & Van Berkum, 2008; Pickering & Gambi, 2018). Accumulating evidence suggests there are differences in how older and younger adults use contextual information to support predictive processing. However, the precise nature of these developmental differences is debated.

Some researchers posit that younger adults are less sensitive than older adults to contextual cues during predictive processing. For example, the ERPs of older adults show both delayed and reduced facilitation effects for highly predictable nouns during sentence processing (Federmeier et al., 2003; Payne & Federmeier, 2018; Wlotko et al., 2010, 2012). Similarly, older adults do not show facilitated processing for words that are unpredictable but semantically related to a highly predictable word (e.g., there is no facilitation for the word *pines* when context biases the word *palms*), an effect that is readily observed in younger adults (Federmeier et al., 2002). Therefore, when older adults generate predictions during language processing, they do so with less detail and scope as younger adults (see Häuser et al., 2019; for converging evidence).

Other accounts point to evidence for increased, not reduced, effects of contextual facilitation in aging. Eye-tracking studies have shown that older adults gain more facilitation during sentence reading when words are more predictable based on context (Choi et al., 2017; Kliegl et al., 2004), and that they are more likely to skip words during first-pass reading and backtrack for repairs (e.g., DeDe, 2014; Rayner et al., 2006). Similarly, speech recognition studies suggest that elderly participants benefit more from high levels of contextual constraint than younger adults when listening to distorted, dubbed-over or truncated speech (Pichora-Fuller, 2008; Pichora-Fuller et al., 1995; Tun & Wingfield, 1994; Wingfield et al., 1985, 1991; Wingfield & Stine-Morrow, 2000). Together, these findings suggest older adults leverage top-down, context-driven processing and/or engage in late-stage re-analysis to minimize the impact of an impaired or unexpected bottom-up signal.

Perhaps more on point given our research question are two studies that investigated pre-nominal effects of prediction in older adults by using in(definite) articles. For example, using the *a/an* design of their original 2005 study, DeLong and colleagues (2012) found that in older adults, the N400 component elicited by phonologically aligned articles did not readily pattern with article cloze probability as is the case in younger adults, which suggests that older adults do not predict phonological features of nouns in a graded manner as younger adults do. In another study that used gender-marked articles from Dutch, Huettig and Janse (2016) found that old age actually increased the number of fixations on correct targets in a cross-modal visual word paradigm when working memory was controlled for, which suggests any effects of reduced linguistic prediction in aging, if they emerge, can be primarily attributed to age-related impairments in executive functions.

Altogether then, studies investigating linguistic prediction in older adults highlight two potential routes for how changes in sensitivity to context might alter graded linguistic prediction. Studies that probed prediction in the strict sense of the word (e.g., pre-nominally) failed to show conclusive findings. Preliminary evidence that requires further substantiation suggests that older adults may not show graded effects of prediction (e.g., DeLong et al., 2012; Federmeier et al., 2002).

The present study

In this study, we investigated pre-nominal prediction effects in a relatively large sample of German-speaking younger and older adults (N = 132). Using a moving-window self-paced reading (SPR) task, we presented German sentence stems such as "Als Paul endlich seinen Führerschein erhielt, fuhr er ständig mit …" (English: When Paul finally got his driver's license, he was always driving with …), that strongly biased a particular gender-marked noun and its corresponding article (e.g., "dem $_{\text{neuter}}$ Auto $_{\text{neuter}}$ von Freunden", the car of friends). Crucially, in half of the items, sentence stems were completed with gender-marked articles and nouns from a different grammatical class (e.g., "der $_{\text{feminine}}$ Gruppe $_{\text{feminine}}$ von Freunden", the group of friends). Hence, the prenominal article ("dem" vs "der") could serve as an early cue to indicate whether the sentence is progressing as expected.

Using SPR allowed us to investigate reading times at the article in a relatively isolated fashion, since SPR, unlike eye tracking, prevents all parafoveal preview effects that go beyond the length of the next word. Measuring reading times pre-nominally, and not the noun, allowed us to disentangle early effects of prediction (at the article) from late-stage effects of integration (at the noun). Crucially, by using gender-marked articles we were able to bypass some of the concerns associated with the "a/an" approach of measuring predictability effects, because a gender-marked article always needs to align with the head noun, irrespective of whether the noun follows directly or not.

In order to investigate graded effects of predictability, we took into account not only the most probable article-

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

noun combination that could finish a sentence (e.g., "dem Auto", the car, identified by means of prior cloze ratings), but also the second-most probable continuation (e.g., "dem Motorrad"; the motor bike). This design allowed us to not only investigate reading times (RTs) on predictable vs. unpredictable articles (henceforth: predictability), but in addition, the relative imbalance between people's most probable first choice continuation and their most probable second choice continuation (henceforth: imbalance). Hence, our measure of imbalance allowed for conclusions as to whether, and to what degree, younger and older adults initially predict multiple options during sentence processing, as opposed to only one.

Since we aimed to investigate the relative distances between the probability of each continuation, we computed a difference score between first- and second-choice completions (henceforth, "imbalance" value). Lower imbalance values (i.e. lower cloze differences) suggest a closer proximity in the likelihood of the first- and second-choice responses, and therefore, a greater balance between two relatively equi-probable sentence continuations. Higher imbalance values suggest a larger probabilistic difference between the first and second choice (i.e., more of an imbalance), and therefore, a stronger bias towards one specific continuation (as opposed to multiple).

Our predictions were as follows: If younger and older adults predict predominantly one sentence continuation (e.g., the one with the highest cloze probability), their RTs at the level of the pre-nominal article should show effects of article predictability, but not of cloze difference (imbalance). Such a finding would support "all-or-nothing" accounts of prediction.[1] If, on the other hand, article RTs were modulated by the difference in cloze probability between the first and second completion (i.e. imbalance), this would indicate that participants generate graded predictions about upcoming linguistic structures by considering multiple sentence continuations. Specifically, with respect to aging, and given conflicting findings from prior literature, we could expect a pattern of results where older adults show either reduced or more pronounced effects of linguistic prediction overall.

Method

Participants

Eighty-four younger (maximally 35 years of age) and 50 older (minimally 65 years of age) native speakers of German participated in the experiment. Younger adults were compensated with course credit for their time; older adults received financial compensation at a rate of € 10 per hour. All participants had normal or corrected-to-normal vision and reported no neuropsychiatric medication at the time of testing. Two older adults had to be excluded from the final analysis due to either MCI/beginning dementia at the time of testing (one subject; based on self-report) or abnormal performance in the AX-CPT cognitive control task (another subject). The final sample consisted of 132 participants: 84 younger adults (55 female, 29 male; *mean age* = 21 years, *SD* = 3) and 48 older adults (22 female, 26 male; *mean age* = 73 years, *SD* = 5). Informed written consent was obtained from all participants. All study procedures were in line with the Helsinki declaration on human subject testing.

To ensure that the younger and older adults in our sample were representative of their respective age groups, all participants completed a battery of standardized cognitive tests. These tests included a modified version of the AX-CPT task to assess context maintenance (for details, see Schmitt et al., 2014), the Reading Span task (adapted from Kane et al., 2004) to assess working memory, and a lexical-semantic knowledge task that tested knowledge of semantic associations and word synonyms (see Lorenz & Kray, 2019). Table 1 shows performance scores for younger and older adults in each of the three tasks. Note that, due to missing data, the number of participant data per task varied. Between-groups t-tests (see Table 1) indicated that the group of older adults performed consistently worse than the younger group in both tests that tap into processes of cognitive control or executive functions (i.e., the AX-CPT and Reading Span tasks).[2] In contrast, older and younger adults performed equally well in the lexical-semantic knowledge task. Taken together, this pattern of performance suggests an age-related decline in executive functions or cognitive control among the older adults, but spared lexical-semantic knowledge (see Braver & Barch, 2002; Burke & Shafto, 2008; Craik & Salthouse, 2000; Hasher & Zacks, 1988; Lindenberger, 2014).

Materials

Cloze ratings for an initial set of 73 items were obtained from 40 younger-adult native speakers of German (younger than 35 years; mostly Psychology students; age range = 18-27 years; 25 female, 15 male), and from 25 older-adult native German speakers (age range = 63 – 81years ; 10 female, 15 male), who did not participate in the main experiment. Participants were presented with sentence frames (e.g., "Nachdem Paul seinen Führerschein erhalten hatte, fuhr er ständig mit ..."; English approximation: "When Paul finally got his driver's license, he was always driving around with ...") that were truncated before the definite article and asked to generate a definite article and noun that best completed the sentences (e.g., "dem $_{\text{neuter}}$ Auto"; English: the car). Crucially, unlike many cloze rating tasks that only request a single response, participants were additionally asked to generate a second-best sentence completion (ar-

1 Note that our definition of gradedness somewhat differs from the one used in e.g., DeLong et al. (2005), where gradedness was defined as the correlation between predictability and ERP N400 component. Here, we define gradedness by means of the difference in cloze probability between first and second most likely continuation.

2 For each group comparison we assumed an alpha threshold of .05. There was no correction of multiple comparisons because each task was included to measure a different cognitive construct (e.g., context maintenance, working memory etc.).

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

Table 1. Results of the Cognitive Test Battery in Younger and Older Adults

	Younger Adults	Older Adults			
n	83	44	*t*(125) =	*p*	
AXCPT (context maintenance and updating):					
RTs on context-dependent (hard) trials (ms)	711 (238)	1217 (416)	-8.73	< .001	***
RTs on context-independent (easy) trials (ms)	559 (110)	943 (356)	-9.07	< .001	***
RT Cost (ms)	152 (156)	274 (214)	-3.69	< .001	***
n	82	46	*t*(126) =		
Reading Span (WM capacity)	0.86 (0.12)	0.72 (0.18)	5.38	< .001	***
n	83	48	*t*(129) =		
Verbal Knowledge (lexical-semantic knowledge)	0.78 (0.10)	0.78 (0.10)	-.03	= 0.98	

Note. Values for the Reading Span and verbal knowledge tasks indicate proportions of correct responses. *n* indicates the number of subjects that completed each cognitive test. Significance levels: * $p < .05$, ** $p < .05$, *** $p < .001$.

ticle and noun) that indicated how the sentence could be completed otherwise of the first completion (e.g., "dem Motorrad $_{neuter}$"; English: the motor bike). It was the combination of these first- and second-best completions in the cloze ratings that allowed us to investigate the effects of imbalance on reading times (see below). We computed the cloze probabilities for responses that were produced with the highest frequency in the first- and second-best guess ratings. These calculations were done separately for first and second responses, and they were done separately for articles and nouns.[3] In other words, for each set of responses (first and second completions separately) we identified the one completion that was produced with the highest frequency for articles and nouns, which resulted in four cloze probability values per item (two cloze values for the most frequent article in first and second responses, and two cloze values for the most frequent noun).

Crucially, our research questions are tied to two variables that were derived from the cloze ratings: predictability (referring to the cloze probability of the most predictable first-completion gender-marked article and noun), and imbalance (defined as the difference in cloze probability between the first- and second-completion article/noun). Example stimuli are presented in Table 2. Figure 1 shows a schematic display of the stimulus design.

Predictability

We selected an initial set of 48 stimuli that consistently elicited relatively high-cloze nouns and articles in the first completions in younger and older adults (i.e., were highly constraining towards a narrow set of gender-marked articles and head nouns); these 48 stimuli were included in the main experiment.

Only 48 items out of initially 73 were selected because many items did not consistently elicit definite gender-marked articles in the cloze ratings, even when participants produced the corresponding head nouns quite frequently and with considerable agreement. One example is the item "Im Sommerurlaub auf Mallorca schlecken die Kinder (das $_{\emptyset}$) Eis" (English approximation: "During their summer vacation in Mallorca, the kids are eating (the $_{\emptyset}$) ice cream"[4]) that did not make it into the experiment. Here, the head noun "Eis" had a cloze probability of 0.97 in the cloze ratings from younger adults, in other words 97% of younger participants produced that noun with the given context. However, the gender-marked article das only had a cloze probability of .05, because some participants only produced indefinite articles (e.g., "ein Eis", "an einem Eis"; an ice cream, at an ice cream), others produced possessive pronouns (e.g., their ice cream), some produced no articles/possessive pronouns at all, and so on. We note that this approach differs from some of the earliest studies on article predictability (see e.g., Wicha et al., 2004), which determined predictability on gender-marked articles by means of the predictability of the head noun only, and not the combination of the definite article and the head noun as is done here. The current norming approach highlights that high-cloze nouns do not necessarily elicit one type of gender-marked article only.

Cloze probabilities of the 48 final selected articles ranged from .43 to 1 in younger adults (M = .81, SD = .14) and from .32 to .96 in older adults (M = .75, SD = .17). For predictable nouns, cloze probabilities ranged from .30 to 1 in younger adults (M = .78, SD = .16), and .32 to 1 in older adults (M = .76, SD = .19).

The experimenters then chose unpredictable, low-cloze article-noun continuations for each sentence stems (see

3 Note that we also explored other options on how to calculate cloze ratings from first and second responses. One of these options, for example, was to collapse first and second responses into one column, but we ultimately deemed this approach inappropriate as it considerably reduced the cloze probability of the first response, and therefore was not true to the relatively constraining nature of our sentences that clearly bias one or two dominant responses. We return to this point in the GD

4 Note that, unlike in English, German mass nouns such as *ice cream* that are preceded by an (in)definite article are not ungrammatical (e.g., it is not ungrammatical to say "eat an ice cream"), so we had no reason to assume that participants would not produce gender-marked articles here.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

Table 2. Overview of Experimental Items in German (English Approximations in Parentheses)

		Younger Adults		Older Adults		
Context		First	Second	First	Second	(Unpredictable)
Nachdem Paul seinen Führerschein erhalten hatte, fuhr er ständig mit ... *(When Paul got his driver's license, he was always driving around with ...)*	Article	dem	dem	dem	dem	der
	Noun	Auto (car)	Motorrad (motor bike)	Auto (car)	Motorrad (motor bike)	Gruppe (group)
In der Nachmittagshitze war der Wein warm geworden, also stellte Johanna ihn in ... *(In the heat of the afternoon, the wine had become warm, and so Johanna put it it in ...)*	Article	den	den	den	den	die
	Noun	Kühlschrank (fridge)	Keller (basement)	Kühlschrank (fridge)	Schatten (shade)	Badewanne (bath tub)
Nach der Gartenparty spülen die Meiers in der Küche *(After the barbeque, the Smiths were in the kitchen, cleaning ...)*	Article	das	die	das	die	den
	Noun	Geschirr (dishes)	Teller (plates)	Geschirr (dishes)	Gläser (glasses)	Kasten (box)

Note. Definite articles always translate to English "the".

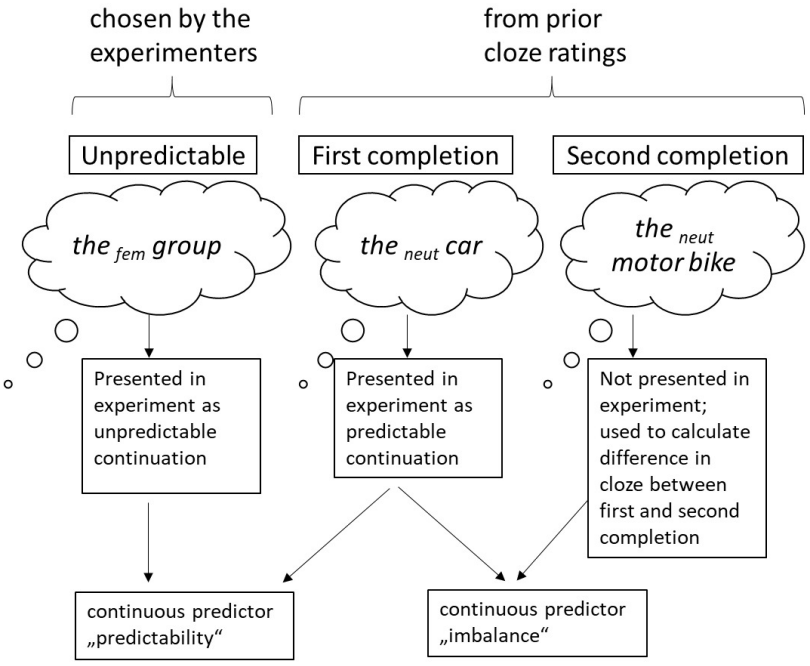

Figure 1. Stimulus Design

Note. Item used for illustration is "When Paul got his drivers' license, he was always driving around with …". Subscripts indicate the noun's grammatical gender, fem=feminine, neut=neuter.

Table 2 & Figure 1) making sure that a) unpredictable nouns had a different grammatical gender than the first-guess predictable nouns, b) unpredictable nouns were never produced as first- or second-guess completions in the cloze ratings, and c) the unpredictable nouns matched with the predictable nouns in frequency. These frequency estimates were based on the Zipf scale from the SUBTLEX DE data base (see Brysbaert et al., 2011). For example, the unpredictable article-noun combination for the item „Nachdem Paul seinen Führerschein erhalten hatte, fuhr er ständig mit

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

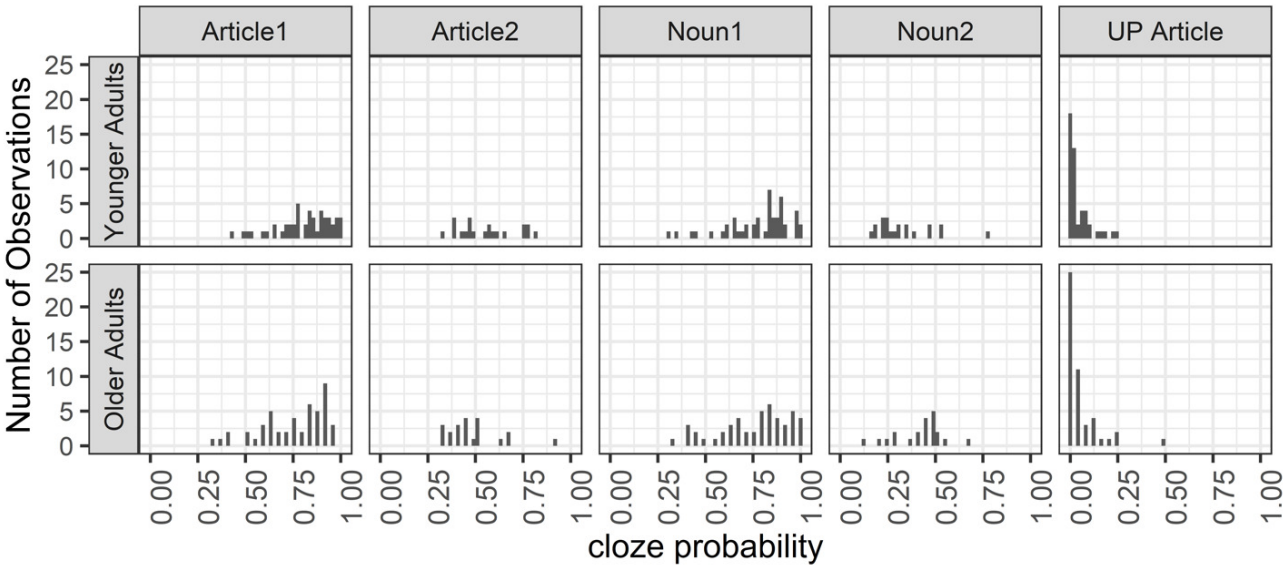

Figure 2. Histograms of Cloze Probabilities for Articles and Nouns in Younger and Older Adults

Note. "Article1" and "Article2" ("Noun1" and "Noun2") refer to the highest-cloze articles (nouns) selected from the first and second cloze rating responses. "UP Article" refers to unpredictable gender-marked articles. Cloze probabilities of unpredictable nouns are not shown here because they were zero throughout.

..." (English: When Paul finally got his driver's license, he was always driving around with ...; predictable continuation: "dem $_{\text{dative neuter}}$ Auto $_{\text{neuter}}$"; English: the car) was "der $_{\text{dative feminine}}$ Gruppe" (von Freunden) (English: the group of friends). By definition then, unpredictable nouns had zero cloze probabilities, but the cloze probabilities of the unpredictable articles were not always zero, since they depended on the fraction of participants who produced a given unpredictable article in the cloze ratings. The cloze values for unpredictable gender-marked articles ranged from 0 to .25 in younger adults (M = .04, SD = .05), and from 0 to .48 in older adults (M = .05, SD = .08). Figure 2 shows histograms of the distribution of cloze probabilities for all first- and second-response gender-marked articles and nouns used in the experiment, split out by age group.

Predictable and unpredictable nouns were matched in frequency, $t(93)$[5] = 1.25, p = .21, d = .27, but unpredictable nouns were slightly longer than predictable nouns (7 vs 6 characters, respectively), a significant difference, $t(94)$ = -2.50, p = .01, d = 0.51. Out of the 48 final experimental items, 46 items had the critical noun and article in the accusative case, whereas two had their critical articles and nouns in the dative case.[6] The final items were relatively balanced with respect to the frequency of occurrence of their definite article forms (see Table 3). Note that we included length and frequency as control variables to all statistical models examining reading times.

Younger and older adults' first completions in the cloze ratings largely agreed, both with respect to the definite article and the noun. For one item, the older adults' cloze ratings favored a different head noun than the cloze ratings from the younger adults. However, since that noun had the same grammatical gender (and therefore, was identical with the younger adults' cloze ratings with respect to the gender-marked article), this item was not removed from the analyses.

Table 3. Frequency of occurrence of definite article forms in the experiment

	das	*dem*	*den*	*der*	*die*
predictable	14	2	15	0	17
unpredictable	10	0	17	2	19

Note. In German "die" is ambiguous as it can refer to feminine singular nouns but also to plural nouns across gender.

Imbalance

Some of the 48 experimental items did not work for a systematic investigation of item imbalance (which we defined as the difference in cloze probability between the first and the second completions from the cloze ratings), because their second completions did not allow for such an analysis. For example, there were items which yielded lexically identical or near-synonym first and second completions (e.g., bus-bus; stove-stove top, goal-goal line). Other items yielded second completions whose highest-cloze ar-

5 Note the *df* for the t-test is 93 here (as opposed to 94 [(48*2)-2]), because one noun was not referenced in the SUBTLEX-DE data base.

6 Note that we initially ran all models presented below using case marking of the critical noun and article as an additional control variable. However, since case marking did not account for systematic variance in any of these models (presumably because of the low variability in this predictor), we dropped this control variable in the final models.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

ticles were indefinite or a possessive pronoun (e.g., "ein Bier", "ihre Mutter"; a beer, her mother). Yet, other items yielded no particular gender-marked article in the second completions (henceforth, "zero" response), i.e. participants left the second completion blank (since this happened frequently when first-guess articles and nouns had cloze probabilities larger than .8, so we assume it was difficult for participant to come up with a good second completion when the first completion was highly dominant). Because of these problems, some of the 48 experimental items were excluded. The final item pool that allowed for a systematic investigation of the second completions consisted of 32 items in younger adults, and 28 items for the older adults. However, these item sets were not mutually exhaustive, as there were items in the younger-adult subset that were not specified for older adults, and vice versa. As 21 items were consistent across both the younger and the older adult group, in order to make sure that the number of experimental items used for both age groups was identical when analyzing effects of imbalance, we subset to these 21 items throughout. We deemed this approach is more sound statistically, since it allowed us to properly estimate random effects where the factor age group was fully nested within items (Barr et al., 2013). After identifying all 21 items, we then proceeded to compute imbalance values for each item in each age group, by subtracting the cloze probability of the most frequent second completion from the cloze probability of the most frequent first completion. For example, in the older adults' cloze ratings, the first-completions article in the drivers' license item had a cloze probability of .76 ("dem [Auto]"), whereas the second-completion article ("dem [Motorrad]", the motor bike) had a cloze probability of .68. Therefore, the imbalance value (difference in cloze) for that item in that group was .08. In the cloze ratings from the younger adults, who (for this particular item) produced the same first and second-completion article as the older adults did, the difference in cloze was .09, corresponding to cloze values of .84 (first completion) and .75 (second completion). Larger imbalance values indicate items whose sentence context strongly favors one particular completion, as opposed to multiple completions that are equi-probable and balanced.

Imbalance values of the 21 final selected items ranged from -.09 to .63 (M = .22, SD = .18) and -.04 to .60 (M = .23, SD = .18) in younger and older adults, respectively, a non-significant difference, $t(40)$ = .28, p = .80). For nouns, imbalance values ranged from .06 to .73 in younger (M = .41, SD = .20), and from -.04 to .72 in older adults (M = .31, SD = .19), again a non-significant difference, $t(40)$ = -1.66, p = .11. Figure 3 shows histograms of the distribution of imbalance values for gender-marked articles and nouns in both age groups.

After identifying all 21 items that worked for investigations of article predictability and imbalance, two final steps remained for stimulus preparation. First, to avoid that the possibility that sentence-final effects would influence RTs on the head noun (see Just et al., 1982; Mitchell & Green, 1978, for evidence from self-paced reading; but see Stowe et al., 2018, for conflicting view), we added several words that represented a plausible continuation to the sentence (e.g., … "fuhr er ständig mit dem Auto $_{\text{predictable}}$ / der Gruppe $_{\text{unpredictable}}$ von Freunden auf den Landstraßen herum"; English: "When Peter finally got his driver's license, he was always driving around with the car $_{\text{predictable}}$ / the group $_{\text{unpredictable}}$ of friends on the roads").

Second, to account for spill-over effects on reading times for words following the definite article, the experimenters inserted three additional words (e.g., adverbs, adjectives) between the definite article and the noun, e.g. … "mit dem alten aber zuverlässigen Auto von Freunden, mit der alten aber zuverlässigen Gruppe von Freunden" (English: … with the old but reliable car from friends, with the old but reliable group of friends). All spill-over words were chosen so as to maintain existing predictions for the noun, while at the same time resulting in a plausible reading of the unpredictable version of the sentence. Note that the spill-over words were identical over predictable and unpredictable versions for each sentence, except for the fact that they sometimes differed in length, depending on whether or not an adjective were gender-marked.

Finally, all 48 experimental items (these included the critical 21 items that we would later use to estimate effects of imbalance) were evenly distributed on two experimental lists ($n_{\text{expected items}} = 48$, $n_{\text{unexpected items}} = 48$) so that each participant viewed only one experimental version of each item during testing (i.e. Latin square design). We added 36 predictable sentences from the Potsdam sentence corpus as fillers (adjusted for length, where necessary) to ensure that participants continued to make predictions during reading, despite occasionally encountering unpredictable sentence continuations (e.g., Brothers et al., 2017). In total, there were 48 experimental + 36 filler = 84 sentences per list. Out of the 48 experimental sentences on each list, 21 items allowed for an investigation of item imbalance. Comprehension questions (simple yes/no questions) were created for 25% of all sentences to make sure that participants read the sentences for content. The experimental items and fillers were randomly distributed on each list, with two constraints: 1) No more than four unexpected items in a row, and 2) no more than four items with comprehension questions in a row. Finally, in order to prevent trial-order effects, the experimenters created a reversed version of each list, yielding a total of four experimental lists.

Procedure

The experimental session consisted of the self-paced reading task (~ 20-25 min), followed by the cognitive test battery (~ 30 min). To avoid effects of fatigue, the administration order of the cognitive tests was counterbalanced. However, the self-paced reading task was always administered before the cognitive tests.

In the self-paced reading task, participants read sentences on a screen word-by-word. Each trial started with the presentation of the first word of the sentence, next to a number of underscores, separated by spaces, indicating the number of words to follow. By pushing the space bar with their dominant hand, participants proceeded to the next word, and the letters of the previous word were replaced with underscores (non-cumulative "moving window" format; Jegerski & VanPatten, 2014). Participants were instructed to read the sentences as fast as possible, and to an-

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

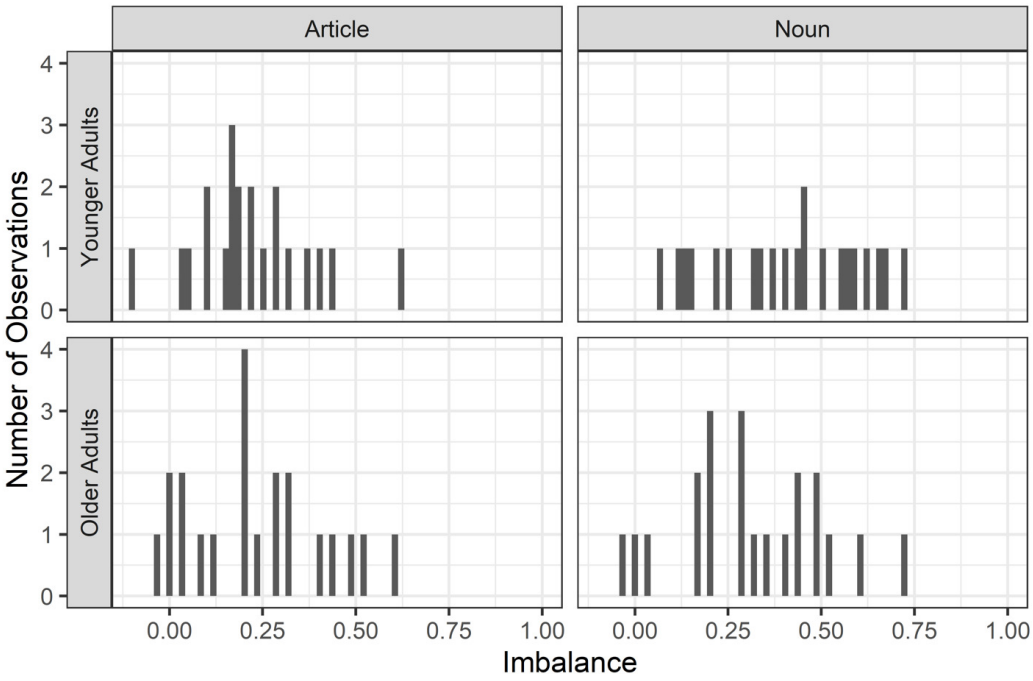

Figure 3. Histograms for Imbalance Values for Gender-marked Articles and Nouns in both Age Groups

swer all true/false comprehension questions as accurately as possible by pushing the "J" (Yes, correct) and "N" (No, incorrect) bars on the keyboard. Trials were separated by a 500 ms fixation cross.

All experimental tasks were presented on a Fujitsu Siemens P-19-2 monitor with a screen resolution of 1280 x 1024 pixels, using a Courier New 18pt font on a white background. All tasks were controlled using E-Prime 2.0 software (Psychology Software Tools, Pittsburgh, PA).

Approach to analysis

We present our data analysis in two sections to allow for the fullest number of observation possible when addressing each research question. The first section, which explores the effects of predictability, includes the full set of 48 experimental items, and examines the effects of predictability alone, not taking into account imbalance. The second section (Effects of imbalance) examines the effects of item imbalance (defined as the difference in cloze probability between the first and second completions in the cloze ratings), and includes only the 21 items that made it possible to measure this effect (i.e., a subset of the 48 original items). The data and analysis script can be found on this paper's project page using the link, https://osf.io/kh7tn/.

Prior to analysis, reading times (RTs) for each region were trimmed minimally by identifying all data points that fell 2 SD below or above the minimum or maximum of an individuals' mean response time per region-condition, and replacing these values with the minimum/maximum for that subject-condition in that region. Altogether, the replacement procedure affected less than 2% of all data points in younger adults, and less than 3% of all data points in older adults.[7]

We focus on three regions of interest, i.e. the gender-marked article (e.g., "dem/der", the), the spill-over region (e.g., "alten aber zuverlässigen", old but reliable), and the noun (e.g., "Auto/Gruppe", car/group). We included the spill-over region following an influential SPR-study by Van Berkum and colleagues (2005), even though we acknowledge that more recent studies consistently demonstrated predictability effects directly at prenominal articles (Fleur et al., 2020; Nicenboim et al., 2020). Therefore, our key measures for gender prediction were both the article and the spill-over region. Table-wise summaries of models fit for RTs of single words in the spill-over region, as well as their partial effects plots, are presented in Supplement 1. We fit separate models for each region of interest. In each model, the dependent variable was reading times (in ms), log-transformed to correct for skewness (Gelman & Hill, 2007). The predictor variables in each model differed per section and are therefore described in each section separately. We ran linear mixed effects models (LMEMs) as implemented in the lme4 library (Bates et al., 2015; version 1.1-19) in R (R Core Team, 2015; version 3.5.2).

All models were fit with the scaled length of the appropriate region as a continuous control variable (except for models on the article region since the article was equally

7 Note that we re-ran all final models reported below using log-transforms of raw, untrimmed, RT values for each region of interest. The effects stayed the same.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

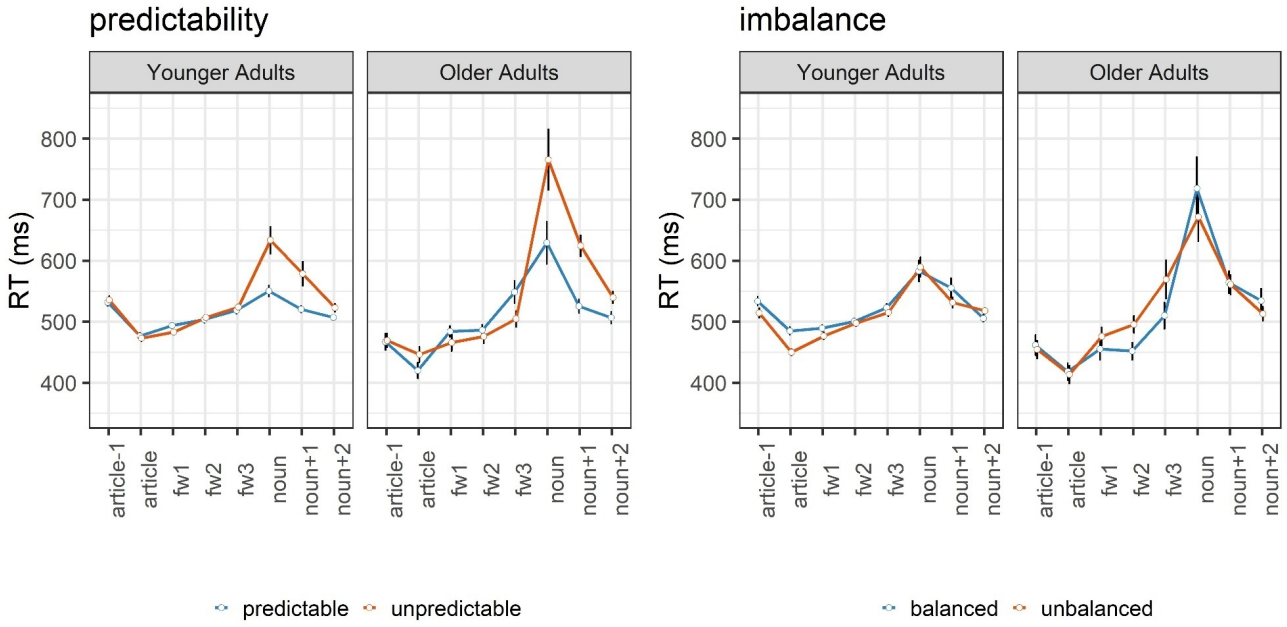

Figure 4. Reading Times on the Critical Region in Younger and Older Adults Illustrating the Effects of Predictability and Imbalance

Note. Error bars represent SE, adjusted for within-subject designs . fw1-3 = spill-over words 1-3 (e.g., "the old but reliable car"). "Unbalanced" means that items are strongly biased towards a specific gender-marked article or noun. Figure 4 may not directly reflect results obtained in our statistical analysis, which takes into account variance associated with items and control predictors.

long (three characters) across trials and conditions). Each model also contained an additional continuous control variable for trial number (scaled, too), in order to account for effects of fatigue or speed-up in RTs in the course of the experiment, and frequency, corresponding to the Zipf scale values from the SUBTLEX-DE data base (there was no frequency predictor for the spill-over region because that region comprised three words).

To protect against anti-conservative model estimates, all models were initially fit with random intercepts for subjects and items, and random slope adjustments for all corresponding within-subject and within-item effects warranted by the design, including their interactions (i.e., a fully maximal random-effects structure; see Barr et al., 2013). In the case of non-converging models, each model was simplified progressively using the *least variance* approach until convergence was achieved (see Barr et al., 2013; for guidelines). P-values were estimated using the Satterthwaite degrees of freedom method, as implemented in the R package *lmerTest* (Kuznetsova et al., 2017). For some models, we make use of model comparisons to check whether excluding non-significant and not trending-towards-significance interactions or model parameters improves model fit. For these comparisons, we use a likelihood ratio test and evaluate significance against the χ^2 distribution, taking as the degrees of freedom the difference in number of parameters between the two critical models.[8] To facilitate interpretation of the data presented below, Figure 4 shows raw, untransformed reading times for critical words and subsequent regions in younger and older adults, with the left panel illustrating the effects of predictability (split out in predictable/unpredictable), and the right panel illustrating the effect of imbalance (based on a median split of the imbalance values).

Results

Comprehension questions

Accuracy on the comprehension questions was near ceiling in both younger and older adults (M = .98, SD = .03, and M = .98, SD = .04, respectively), a non-significant difference, $t(130) = .36$, $p = .72$, $d = .06$. There were no significant differences within groups when accuracy for predictable and unpredictable items was compared: For younger adults, $t(83) = -.79$, $p = .44$, $d = .12$; for older adults, $t(47) = -.19$, $p = .85$, $d = .03$). Overall, these findings suggest that participants in both age groups were attentive during the experiment, and understood the sentences they were reading.

Effects of predictability

Predictor variables were predictability, reflecting the cloze probability of the predictable and unpredictable article and noun, and age group, including the interaction between these variables. Predictability was entered into the model as a scaled continuous variable. Age group was treatment/dummy coded, with younger adults set as the refer-

8 We emphasize that our results remained unchanged when models were not simplified and all critical predictors were left in the models.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

ence category. Hence, effects for age group represent simple, not main, effects as they compare one categorical level (here, older adults) against the baseline category level (here, younger adults). The formal lme4 specification of our models in this section is:

```
lmer(log(RT) ~ predictability * group + trial + length + frequency + (1 + predictability | subject) + (1 + predictability : group | item)).
```

There were simple effects of age group in all models, suggesting that older adults read considerably more slowly than younger adults. For the sake of brevity, and since our focus here is more on within-group differences between predictable and unpredictable items, these will be skipped in model reports. Partial effects plots for all models are presented in Figure 4. Table 4 shows LMER outputs for each model.

Article RTs. Recall that the article region was our primary measure of interest to measure predictability effects. The model for RTs on the gender-marked article showed a significant interaction between predictability and age group (b = -0.02, SE = 0.01, t = -3.41, p < .001). The model plot (see Figure 5, left panel) indicates that this interaction was driven by older adults, who showed facilitation when reading articles that were more predictable. In contrast, younger adults seemed to read articles equally fast, irrespective of whether or not they were high predictable.

Spill-over region. In the model for RTs on the spill-over region (Figure 5, middle panel), there was a hint of an interaction between predictability and age group that failed to reach statistical significance (b = 0.01, SE = 0.01, *t =*1.71, p = .09). The effect of predictability in this model was also not significant (b = 0.002, SE = 0.003, t = 0.81, p = .42).

Noun. The model for RTs on the noun showed a significant effect of predictability (b = -0.03, SE = 0.01, t = -4.28, p < .001), suggesting that as cloze probability increased, reading times went down. There was a hint of an interaction between predictability and age group that was not significant (b = -0.02, SE = 0.01, t = -1.69, p = .09), which suggested that numerically, there was a trend for older adults to gain more facilitation from high-predictability nouns (see Figure 5, right panel). Follow-up models, in which items were split by age group, confirmed this pattern, and showed that the model estimates for the effect of predictability were larger in older, compared to younger, adults (b = -0.05, and b = -0.03, respectively).

Summary of findings for predictability. There were two key findings. First, increasing article predictability facilitated reading times in older, but not younger adults. Noun predictability facilitated reading times in both younger and older adults.

Crucial to our research question is the finding that only older adults showed early (i.e., pre-nominal) expectation violation at the article. The question that emerges from these findings is whether younger adults did not use sentence context predictively. It is with this question in mind that we now turn to the analysis of the effects of imbalance.

Effects of imbalance

The models estimating the effects of imbalance take into account the difference in cloze probability between first and second completions from the cloze ratings. As noted above, since not all items yielded valid second completions from both age groups in the cloze ratings, the models in this section were run on a subset of all items – only those for which valid second completions were available (21 items in total). Again, the DV was log-transformed RTs. The predictor variables were age group (dummy-coded, with younger adults as the baseline category) and imbalance (defined as the difference in cloze probability between the first and the second completion), including the interaction between age group and imbalance. We also added predictability (i.e. the cloze probability of predictable and unpredictable articles and nouns) as control variable. The variables for predictability and imbalance were scaled (i.e., centered around their means) to reduce multicollinearity. As expected, our measures of predictability and imbalance were clearly correlated (Pearson's r = 0.68, 95% confidence interval = CI[0.51, 0.80]). However, this correlation is not, in principle, an obstacle to our approach. Variance Inflation Factors (VIFs) for predictability and imbalance were below 1.6 in all models, which is far below the values considered problematic when multicollinearity is an issue (e.g., Zuur et al., 2010). The formal specification of the models in this section is:

```
lmer(log(RT) ~ imbalance * group + trial + length + frequency + predictability + (1 + imbalance | subject) + (1 + imbalance : group | item)).
```

Again, simple effects of age group (indicating slower reading in older adults) will be skipped in model reports, since they are of no interest to our research questions. Partial effects plots for all models are presented in Figure 6. Table 5 shows LMER outputs for each model.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

Table 4. Effect sizes (b), Standard Errors (SE), and T-Values for Models Estimating the Effects of Predictability and Age Group on Log-Transformed RTs of the Critical Region

	Article				Spill-over region				Noun			
	b	*SE*	*t*	*p*	*b*	*SE*	*t*	*p*	*b*	*SE*	*t*	*p*
Fixed effects												
Predictability	0.004	0.004	0.84		0.002	0.003	0.81		-0.03	0.01	-4.28	***
Age group	0.34	0.05	7.20	***	0.37	0.05	7.09	***	0.44	0.07	6.44	***
Predictability: Age group	-0.02	0.01	-3.41	***	0.01	0.005	1.71		-0.02	0.01	-1.69	
Control predictors												
Trial number	-0.10	0.003	-36.08	***	-0.12	.002	-49.14	***	-0.14	0.004	-34.74	***
Length	-	-	-		0.01	0.01	2.02	*	0.02	0.01	3.02	**
Frequency	0.03	0.02	1.93		-	-	-		-0.01	0.01	-1.03	
Random effects	Variance				Variance				Variance			
Subject	0.07				0.08				0.14			
Subject | Predictability	NA				NA				0.002			
Item	0.004				0.002				0.004			
Item | Predictability	0.0002				NA				NA			
Item | Age group	0.001				0.001				0.003			
Item | Age group: Predictability	NA				NA				NA			

Note. Ø is used for predictors that were removed from the model because they did not contribute substantial variance. NA is used for predictors that were removed because of issues with convergence; this procedure followed the least-variance approach suggested by Barr and colleagues (2013). Significance levels: * $p < .05$, ** $p < .05$, *** $p < .001$.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

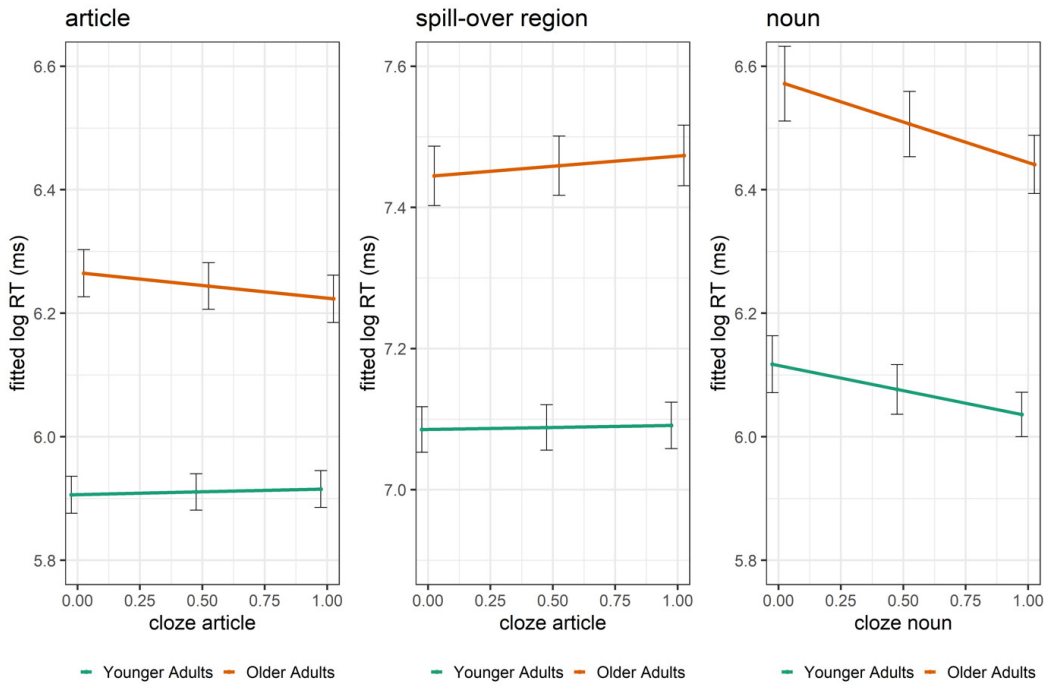

Figure 5. Partial Effects Plots Showing the Effects of Predictability and Age Group on Log-Transformed RTs

Note. Note the larger y-scale in the plot for the spill-over region. Error bars indicate standard error of the mean. Models showed an interaction of age and predictability at the article, a simple effect of age group at the spill-over region, and a main effect of predictability at the noun.

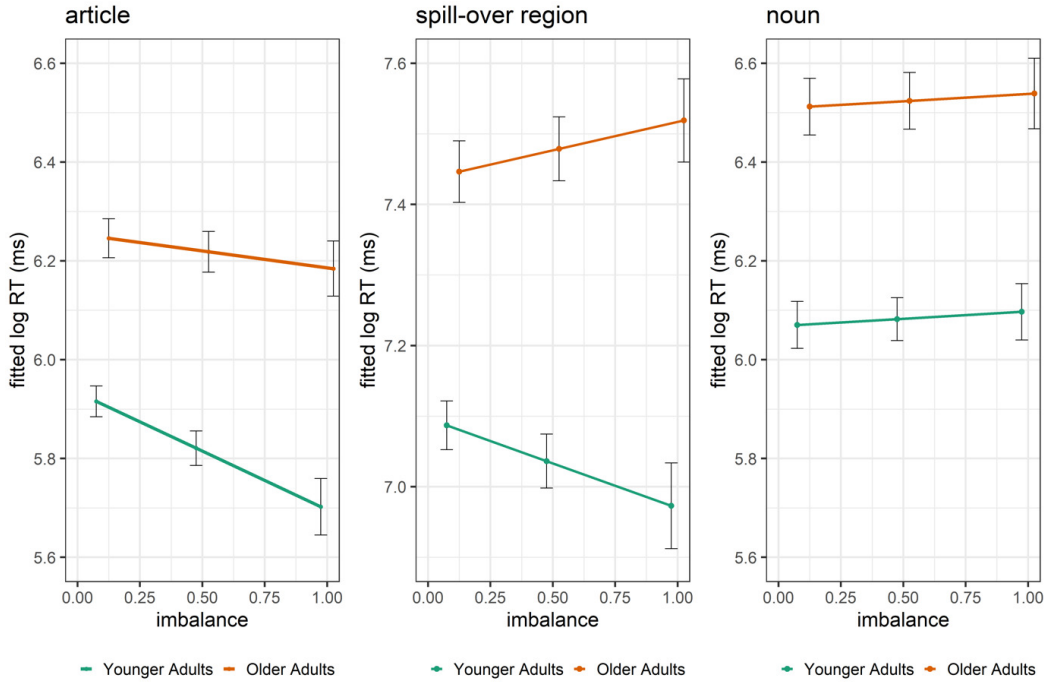

Figure 6. Partial Effects Plots Showing the Effects of Imbalance and Age Group on Log-Transformed RTs

Note. Error bars indicate standard error of the mean. Models showed an interaction between age group and imbalance at the article and at the spill-over region. At the noun, no effects emerged except for a simple effect of age group.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

Table 5. Effect sizes (b), Standard Errors (SE), and T-Values for Models Estimating the Effects of Imbalance and Age Group on Log-Transformed RTs of the Critical Region

	Article				Spill-over region				Noun			
	b	*SE*	*t*	*p*	*b*	*SE*	*t*	*p*	*b*	*SE*	*t*	*p*
Fixed effects												
Imbalance	-0.04	0.01	-3.85	***	-0.02	0.01	-2.68	**		Ø		
Age group	0.35	0.05	7.33	***	0.39	0.05	7.31	***	0.44	0.07	6.55	***
Imbalance: Age group	0.03	0.01	2.76	**	0.03	0.01	4.24	***		Ø		
Control predictors												
Predictability		Ø			0.01	0.004	2.43	*	-0.03	0.01	-4.25	***
Trial number	-0.10	.004	-24.55	***	-0.12	0.004	-33.74	***	-0.14	.01	-23.12	***
Length	-	-	-			Ø			0.03	0.01	2.58	*
Frequency	-0.01	0.02	-0.52		-0.05	0.02	-3.12	**	-0.01	0.02	-0.68	
Random effects	Variance				Variance				Variance			
Subject	.07				0.08				0.13			
Subject | Imbalance	NA				NA				NA			
Item	.002				.002				.003			
Item | Imbalance	NA				NA				NA			
Item | Age group	0.001				NA				.003			
Item | Age group: Imbalance	NA				NA				NA			

Note. NA is used for predictors that had to be removed during model fitting because of issues with convergence; this procedure followed the least-variance approach suggested by Barr and colleagues (2013). Ø is used for predictors that were removed from the final model because they did not contribute substantial variance. Significance levels: * $p < .05$, ** $p < .05$, *** $p < .001$.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

Article RTs. VIFs computed to check for multicollinearity were 1.6 for imbalance and 1.00 for predictability. Model comparisons showed that the scaled control variable for predictability did not explain substantial variance in the data, so it was removed from the model (χ^2 (1) = 2.24, p > .10). The final model showed a significant interaction between imbalance and age group (b = 0.03, SE = 0.01, t = 2.76, p = .01).[9] This interaction was mostly driven by younger adults, as the partial effects plot for this model suggests (see Figure 6, left panel). Specifically, younger adults showed facilitated reading when imbalance was high, in other words, when the sentence context strongly favored one particular completion (as opposed to multiple completions that are equi-probable). Instead, there was slowing when multiple completions were equi-probable and balanced. Older adults, in contrast, showed no difference in reading times depending on item imbalance.

Spill-over region. We found a significant interaction between imbalance and age group (b = 0.03, SE = 0.01, t = 3.50, p < .01). Figure 6 (middle panel) suggests that imbalance had differential effects on reading times in younger vs older adults: Whereas younger adults showed facilitated reading of the spill-over region when imbalance was high, older adults instead showed slowing.

Noun. In the model for the noun (Figure 6, right panel), imbalance did not explain a significant amount of variance, neither as a main effect nor in its interaction with age group. Therefore, both predictor terms were removed from the model (interaction: χ^2 (1) = 0.76, p = .38; main effect: χ^2 (1) = 0.11, p = 0.74). The final model showed no further effects of interest.

Summary of findings for imbalance. Two findings are crucial in the models estimating the effects of imbalance (i.e., cloze difference between first and second completion) on reading times. First, larger imbalance values facilitated reading times for younger (and not older) adults. In other words, younger adults showed faster reading times when predictions were biased towards one probable completion (as opposed to when there were multiple probable completions with comparable cloze probabilities). Second, this early effect of facilitation for more imbalanced items continued onto RTs of the spill over region in younger adults. In older adults, surprisingly, RTs on the spill-over region showed the reversed pattern, in that RTs actually increased with higher imbalance, in other words, when completions were biased towards one response.[10]

Power

There was no a priori power calculation that was well suited given our analysis method, i.e. running linear mixed effects models on trial-by-trial data. However, we did run post-hoc power simulations using the *simr* package in R (Green & MacLeod, 2016) that estimated whether we were sufficiently powered to obtain effects substantially *smaller* than the ones reported in the results. The rationale behind these simulations was that they would give us an idea regarding the stability of our results and yet avoid reporting power estimates for the actually obtained effects (Lakens, 2021).

In running the power simulations, we started with the observed effect sizes for the crucial interaction terms (for age group and predictability: b = - 0.02; for age group and imbalance: b = 0.03), and worked our way down to smaller effect sizes. Using power = 80% as a cutoff, we identified the smallest effect sizes for both interaction terms we were sufficiently powered to find (100 simulations for each new level of b, alpha= .05). Accordingly, we were sufficiently powered to detect an interaction effect as small as b = -0.015 for age group and predictability, and as small as b = 0.022 for age group and imbalance. Note that these values are below (but not dramatically below) the crucial effect sizes reported in the results (b = -0.02 and b = 0.03, respectively). Altogether, we take this to indicate that our experiment was reasonably powered to find effects even smaller than the ones reported.

Discussion

We investigated whether younger and older adults of German generate graded predictions about upcoming sentence content when reading sentences for comprehension. We presented pre-nominal gender-marked articles that were either consistent or inconsistent with the noun that was predictable based on context. In order to investigate gradedness of predictions, we measured the probabilistic difference between the first-best and the second-best article-noun choices that could plausibly complete each sentence (assessed by means of prior cloze ratings in separate groups of younger and older adults, where people were asked to provide two continuations for sentence fragments). We argued that lower probabilistic differences between the two completions were a measure for the imbalance of an item, as they indicated that two sentence continuations were roughly equally probable and balanced. In contrast, larger difference values would indicate a greater imbalance between possible sentence continuations, and therefore, a stronger bias towards a specific sentence completion. We hypothesized that effects of imbalance on reading times would constitute evidence in favor of graded (as opposed to "all or nothing") accounts of prediction since they would demonstrate that language users not only generate multiple predictions about upcoming content when reading a sentence, but that they are sensitive to probabilistic differences between these predictions. With respect to aging, and taking into account conflicting findings from the literature, we hypothesized that older adults would either show reduced or increased effects of prediction, compared to younger adults.

9 The size of the interaction was unchanged (if at all, numerically it seemed to increase) when predictability was not removed from the model, b = 0.03, SE = 0.01, t = 2.83, p = .009.

10 Upon reviewer request, we re-ran models for imbalance that excluded zero and below-zero imbalance items. These analyses maintained 18 out of 21 items. The results replicated all effects reported in our main analysis.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

Our findings indicate that both age groups used context predictively to anticipate grammatical gender of nouns. Hence, older adults, as much as their younger counterparts, predict upcoming linguistic structures. However, only younger adults showed graded effects in generating predictions, in other words, only younger adults took into account the probabilistic difference between competing sentence completions. We discuss these findings in greater detail below.

(Gender) prediction in younger adults

According to our results, younger adults' reading of gender-marked articles was facilitated when sentences were more unbalanced, in other words, more biased towards one particular gender-marked noun and its corresponding article. In addition, there was slowing when several sentence completions were balanced, in other words, when there was no particular bias for one completion over others. Crucially, this early, pre-nominal facilitation continued during the next three words of the sentence (i.e. the spill over region; old but reliable (car)), where processing was still facilitated in younger adults when expectations were biased towards a particular gender-marked noun.

These findings lend support to graded accounts of prediction, because they demonstrate that younger adults use sentence context to predict multiple potential continuations rather than only one. In conditions when several completions are somewhat probable and balanced, there is competition between simultaneously activated meanings, which leads to slowing or processing difficulty. In conditions when the sentence context clearly favors one completion (and one grammatical gender) over others, there is facilitation.

Our findings support information-theoretic accounts that describe language processing as a continuous process in belief updating. In its initial state (here, at the beginning of a sentence), the language parser might not have a strong bias (or belief) about possible outcomes, but the more information accrues, the more it will shift the probability space towards one or two possible continuations (Levy, 2008, 2013). This way, the parser computes multiple beliefs in parallel, each with some degree of probability, and updates their beliefs continuously by discrediting some parses and updating others. Rather than computing an estimate of absolute likelihood for multiple outcomes, the parser assigns a relative weighting, by taking into account the relative distance in their probability. Towards the end of a highly constraining sentence, the probability space may have shifted in such a way that the remaining parses are highly imbalanced, where one particular continuation "weighs" heavier than all the others. In such a situation, processing is facilitated. In other cases (e.g., in sentences where multiple continuations are plausible), the probability space might be distributed such that relatively equal weight is assigned to the remaining possible continuations, in other words, there is no clear and absolute "winner". In these latter situations, processing is more difficult.

Prior studies on prediction of grammatical gender have shown conflicting findings regarding whether people routinely use gender-marked articles (e.g., Fleur et al., 2020; Guerra et al., 2018; Huettig & Janse, 2016; Nicenboim et al., 2020; Wicha et al., 2004) or adjective inflections (Kochari & Flecken, 2019; Nieuwland et al., 2020; Van Berkum et al., 2005) to inform their predictions about upcoming nouns. Even the subset of studies that obtained evidence in favor of gender prediction, showed vast differences regarding timing and polarity of the effects (e.g., Van Berkum et al., 2005; Wicha et al., 2004), and more recently, replication attempts of landmark studies on gender prediction failed (e.g., Kochari & Flecken, 2019; Nieuwland et al., 2020). Potential reasons that could have led to the conflicting results (see Nicenboim et al., 2020, and Nieuwland et al., 2020) are, for example, small sample sizes of earlier studies that are more likely to give rise to noisy estimates, or problematic choice of analysis time windows that are based on visual inspection of raw ERP data. Taken together though, these earlier studies also indicate that gender prediction may be more complex and multi-faceted than initially assumed. For example, listeners show different ERP effects depending on whether a sentence pragmatically licenses the use of a definite gender-marked article (over, e.g., an indefinite article; see Fleur et al., 2020). This suggests that language users are not only sensitive to gender information conveyed by prenominal articles, but that they also take into account pragmatic cues that may be transported in these articles.

Prediction in older adults

The results obtained in the present study suggest that older adults are able to use sentence context and generate predictions during reading. This conclusion is supported by our findings on RTs of the definite article, where older adults showed slowing when reading gender-marked articles that did not match with the gender of the most predictable noun. The conclusion is also supported by older adults' reading behavior on the spill-over region, where older adults showed slowing in the predictable condition and slowing when items were imbalanced (i.e. strongly biased towards a particular gender-marked article). Interestingly, it is possible that the two effects (slow-down at the article and spill-over region) may be qualitatively different from one another. While the RT effect at the gender-marked article suggests an effect of gender prediction, our findings at the spill-over region could indicate processing difficulty when strong expectations for a particular sentence constituent (i.e., the head noun) were consistently violated through insertion of pre-nominal modifiers. This interpretation, if correct, implies that older adults were not only sensitive to *morphological* prediction violation on the article, but that they were also sensitive (maybe more so than younger adults) to *syntactic* expectation violation. Future reserach could look into substantiating this hypothesis. Regardless, our older-adult findings clearly suggest that older adults generate predictions about upcoming linguistic structures, and that they incur a processing cost when these predictions are not fulfilled, as indicated by prolonged reading times. Hence, older adults' sentence processing is predictive in nature.

However, the predictions older adults generate may be narrower in scope when compared to those of younger adults. The fact that older adults did not show reading fa-

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

cilitation for imbalanced items at the article (like younger adults did) may indicate that they are less likely to gain facilitation when multiple predictions are biased towards one prominent alternative. Hence, older adults may not generate *graded* predictions.

This aspect of our findings dovetails with other studies, where older adults are sometimes reported to not show facilitation for nouns that are unpredictable in a context, but share semantic features with the most predictable noun (i.e., older adults do not show facilitation for the word "pines" when the sentence context biased the word "palms"; Federmeier et al., 2002; Wlotko et al., 2012). One of the conclusions drawn in these earlier studies was that older adults do not generate graded predictions about upcoming semantic features during sentence processing, in that they might only predict the most likely continuation (i.e. „palms") that could finish a sentence, but not others (i.e. „pines"). The data presented here support this general conclusion, by suggesting that expectations for morpho-syntactic features of nouns are also reduced in older adults.

One possible explanation for this finding could be that generating multiple predictions is costly because the parser needs to continuously maintain and update possible continuations (Amer et al., 2022; Pickering & Gambi, 2018; Ryskin et al., 2020). Therefore, age-related impairments in working memory prevent older adults from generating multiple, graded predictions. Of note, exploratory analyses that were conducted after data analysis was completed, did not generally support this idea: No reliable effects emerged when we ran additional models on article RTs that included the data from two tasks that assess working memory skills, including the reaction time cost score from the AXCPT task, and the number of correct responses from the Reading Span task. However, we would like to refrain from over-interpreting this result because the old-adult sample in our study (n = 50) was presumably too small to allow for a reliable investigation on individual differences.[11]

Finally, a somewhat unexpected finding in the old-adults data is their reading slow-down in the spill-over region when items were more biased, but we believe that this effect is primarily driven by the same effect the older adults showed for predictability: Recall that our measure of imbalance ultimately reflects a difference score between the predictabilities of people's first and second choice in the cloze ratings. As argued above, we believe that the reading slow-down at the spill-over region reflects syntactic surprisal about the unexpected adjectives when a noun was relatively more predictable.

Imbalance and entropy

Our measure of imbalance is conceptually similar, but not identical to, entropy, a prominent measure of information theory besides cloze probability. Entropy quantifies the degree of uncertainty about what is being communicated as a sentence unfolds. High entropy values refer to greater uncertainty; an entropy value of zero refers to maximal certainty.[12] Conceptually then, imbalance and entropy are related. However, one major difference between the two is that entropy is normally quantified by only taking account responses in a one-shot cloze task (where entropy is then calculated as the sum of the log base two cloze probabilities of all discrete responses). Our measure of imbalance goes beyond these first responses, by taking into account people's second-best intuitions about how a sentence could be continued (also see Limitations and future directions). Higher entropy values correspond to a situation where a large proportion of participants in a cloze task provide distinct first-guess responses about how a sentence could be completed. The imbalance measure in our study is different, because it is derived from a situation when entropy about a first-guess continuation was generally low (i.e., participants in our cloze task generally agreed upon the first best continuation of a given item), but participants had diverging intuitions about the probability space for a second completion. According to our definition of imbalance, when there was high agreement about the second intuition among participants, an item would be balanced. If there was little agreement, the probability space in the second completion was more diversified, which rendered the item imbalanced towards the first response. In sum, imbalance and entropy are conceptually related, but they measure distinct constructs.

Limitations and future directions

A possible concern about the present study could be that our measures of imbalance and predictability are correlated, and that, because of this correlation, it may be difficult to disentangle their separate contributions towards language processing. However, we do not think that the correlation between imbalance and predictability is a fundamental obstacle to our approach. We checked for multi-collinearity in our analyses by computing variance inflation factors (VIFs) for the critical predictors. VIFs were below 2 in all models, which is well below the recommended limit of 10 when multicollinearity is an issue (Cohen et al., 2003). Above and beyond this, we believe that, since imbalance and predictability are separable constructs on a theoretical level, it is all the more important to investigate their (potentially distinct) contributions to language processing. Recent studies examining similarly correlated variables have yielded novel insights into the real-time mechanics of language processing (Lowder et al., 2018; Nieuwland, 2019).

A definite limitation of the present study is the inclusion of the gender-marked article "die" (i.e., "the") in the unpredictable condition, as it is ambiguous between feminine

11 Assuming a small-to-medium effect size (r = .3), an alpha-level of .05 and power of 80% yields a sample size of n = 67 subjects that would be necessary to reliably interpret results from a simple linear regression, where article reading times are predicted by an individual difference variable.

12 In the psycholinguistic literature, low entropy normally goes along with high constraint, because in highly constraining contexts the cloze probability mass is concentrated in one (or few) responses (see Nicenboim et al., 2020).

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

singular nouns (e.g., feminine "die Sonne", the sun) and plural nouns across grammatical gender (e.g., feminine: "die Lampen", the lamps; neuter: "die Autos", the cars; masculine: "die Tische", the desks). Hence, when reading slow-down occurs at unpredictable "die", it is difficult to reliably estimate what this slowing means, over and above the notion that a different gender-marked article was apparently favored, given the context. On the one hand, if participants read unpredictable "die" as the feminine singular determiner, slowing at this wordcould indicate (gender) surprisal, or even revision of the initial noun prediction towards a feminine noun (Fleur et al., 2020). If readers interpret "die" as plural, slowing could indicate a mixture of cognitive processes: It could indicate a similar revision of the noun prediction where people update their expectations from a singular noun to the same noun in its plural form. But it could also signal pragmatic surprisal about the unexpectedly marked plural form (see Fleur et al., 2020, for evidence suggesting that Dutch definite articles elicit surprisal when the discourse only licenses use of the indefinite form). While it is presumably safe to say that, across these different possibilities, slowing at "die" signals surprisal, it is difficult to ascertain the precise cognitive cause of this surprisal.

Another limitation of the present study is that the nature of our cloze task forces a certain commitment regarding the cognitive processes inherent in cloze probability ratings, while diverging accounts are also feasible.[13] In the cloze test of the present study, we asked people to provide a first-choice continuation that came to mind upon reading the context, and additionally, provide a second-choice continuation how the sentence could be completed otherwise (paraphrasing from German, "wie der Satz noch weitergehen könnte", "how the sentence could be continued alternatively"). Based on the cloze responses we obtained, we infer that many (albeit not all; see Materials) participants took this instruction to mean that the second choice should be an alternative continuation (different in form and meaning) in case the first response was not an option. In other words, the second response was contingent on the first response. This contingency may not readily reflect what all researchers believe to be the cognitive processes underlying cloze ratings. For example, Staub and colleagues (2015) argue that responses in a cloze task represent a race for activation, such that multiple responses are issued in parallel and the one with the highest cloze probability is simply the one that reaches activation thresholds fastest in most people. Obviously, such a race situation is not well captured in our cloze task because in the race model, it is feasible that semantically overlapping or completely synonymous responses are racing for competition. In fact, Staub and colleagues (2015) specifically discuss the possibility that competing responses may be those that are closest to the sentence context by means of semantic association. In this light, a methodological issue that warrants further research is how to optimally choose first and second-best completions based on the cloze ratings, if not by means of first and second responses. One option, for example, is to have people provide only one continuation, and compute first- and second-best completions based on whatever continuation was provided with the highest and second-highest frequency in these one-shot ratings. Of note, we explored this approach during analysis of our cloze ratings, by collapsing first and second-best completions into one column. We ended up dismissing it because it inevitably lowered the cloze probability of the first completion and thereby blurred the fact that most of our sentences were, after all, biased towards one or at most two particular completions. However, we emphasize that this approach may very well be feasible for one-shot cloze ratings.

A surprising finding in this study that warrants further investigation is the lack of a basic predictability effect in younger adults. If younger adults showed fully graded prediction of grammatical gender, there should have been an imbalance effect *on top* of a basic predictability effect in this age group. We can only speculate as to why this was not the case, but one potential explanation for this surprising finding could lie in the distributions of cloze probabilities obtained for the predictable and unpredictable (first-choice) articles in our cloze norms. Compared to the norms obtained for older adults, which showed relatively larger variability with respect to predictability (SD = .17 and SD = .1, for predictable and unpredictable articles), the variability in the young-adults norms for predictability was somewhat smaller (SD = .14 and SD = 0.05). This may have prevented predictability effects from becoming more influential in the young adults sample.

Ultimately, one important next step to settle these remaining questions will be replication. Our measure of imbalance is novel, and therefore, requires substantiation from additional studies. Replication is also important because several studies have concluded that prenominal prediction effects in language processing, if they exist at all, are small in size and difficult to detect reliably. Even though we have no reason to assume the present study was underpowered (see Power), we note that recent sample size estimates from pre-registered ERP studies in Dutch ranged from 80 to 189 subjects (Fleur et al., 2020; Nieuwland et al., 2020). While it is true that sample size estimates for ERPs studies may not directly apply to behavioral paradigms such as self-paced reading, we have no reason to assume that sample size estimates for this latter type of study should be any lower. If anything, they should presumably be higher, since ERP studies likely over-estimate predictability effects during reading (because of their artificially slow word presentation rate; see Huettig & Guerra, 2019).

13 It is a different question altogether to what extent responses in the cloze task capture processes going on during real-time language comprehension.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

Conclusions

We investigated if groups of younger and older adults generate graded predictions about upcoming content when reading sentences for comprehension. Gradedness was operationalized as the scaled difference in cloze probability between the most likely and second-most likely sentence completion that could finish the sentence. Sentences with a greater probabilistic difference between the two completions were considered as imbalanced and thought to create a greater bias towards one particular completion (as opposed to many).

Results showed evidence for predictive processing in both age groups. However, only younger adults seemed to generate graded predictions. In particular, younger adults demonstrated slowed reading comprehension when possible completions were more balanced. Instead, there was facilitation when predictions were imbalanced (i.e., more biased). In contrast, older adults' RTs patterned mostly with predictability, showing that, while able to generate predictions based on context, older adults may not predict multiple continuations.

Our findings for younger adults add to a growing body of research that defines language processing as a continuous process of probabilistic updating, in which multiple expected sentence continuations are maintained and updated in the probability space at any point in time. Crucially, our results indicate that these continuations are weighted by the strength of their probability, so that highly likely completions might receive relatively greater weight in the probability space and less likely completions receive relatively less weight. Older adults, while able to use sentence context predictively, do not predict multiple sentence continuations like younger adults.

Acknowledgements

The authors would like to thank Luzi Warnatsch and Marlow Springmeier for help with the data collection. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

Contributions

Conception and Design: KIH and AB. Analysis and Interpretation: KIH and AB. Writing: KIH, AB and JK.

Data Accessibility Statement

All data files and analysis scripts can be found on this paper's project page using the link https://osf.io/kh7tn/

Competing interests

The authors declare no conflict of interest.

Submitted: November 11, 2021 PDT, Accepted: June 30, 2022 PDT

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CCBY-4.0). View this license's legal deed at http://creativecommons.org/licenses/by/4.0 and legal code at http://creativecommons.org/licenses/by/4.0/legalcode for more information.

References

Amer, T., Wynn, J. S., & Hasher, L. (2022). Cluttered memory representations shape cognition in old age. *Trends in Cognitive Sciences*, *26*(3), 255–267. https://doi.org/10.1016/j.tics.2021.12.002

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Braver, T. S., & Barch, D. M. (2002). A theory of cognitive control, aging cognition, and neuromodulation. *Neuroscience & Biobehavioral Reviews*, *26*(7), 809–817. https://doi.org/10.1016/s0149-7634(02)00067-2

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *Journal of Memory and Language*, *93*, 203–216. https://doi.org/10.1016/j.jml.2016.10.002

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, *58*(5), 412–424. https://doi.org/10.1027/1618-3169/a000123

Burke, D. M., & Shafto, M. A. (2008). Language and aging. In F. I. M. Craik & T. A. Salthouse (Eds.), *The handbook of aging and cognition* (3rd ed., pp. 373–443). Psychology Press.

Choi, W., Lowder, M. W., Ferreira, F., Swaab, T. Y., & Henderson, J. M. (2017). Effects of word predictability and preview lexicality on eye movements during reading: A comparison between young and older adults. *Psychology and Aging*, *32*(3), 232–242. https://doi.org/10.1037/pag0000160

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression*. Lawrence Erlbaum Associates.

Craik, F. I. M., & Salthouse, T. A. (Eds.). (2000). *The handbook of aging and cognition* (2nd ed.). Erlbaum.

DeDe, G. (2014). Sentence comprehension in older adults: Evidence for risky processing strategies. *Experimental Aging Research*, *40*(4), 436–454. https://doi.org/10.1080/0361073x.2014.926775

DeLong, K. A., Groppe, D. M., Urbach, T. P., & Kutas, M. (2012). Thinking ahead or not? Natural aging and anticipation during reading. *Brain and Language*, *121*(3), 226–239. https://doi.org/10.1016/j.bandl.2012.02.006

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117–1121. https://doi.org/10.1038/nn1504

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*(4), 469–495. https://doi.org/10.1006/jmla.1999.2660

Federmeier, K. D., McLennan, D. B., De Ochoa, E., & Kutas, M. (2002). The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: An ERP study. *Psychophysiology*, *39*(2), 133–146. https://doi.org/10.1111/1469-8986.3920133

Federmeier, K. D., Van Petten, C., Schwartz, T. J., & Kutas, M. (2003). Sounds, words, sentences: Age-related changes across levels of language processing. *Psychology and Aging*, *18*(4), 858–872. https://doi.org/10.1037/0882-7974.18.4.858

Fleur, D. S., Flecken, M., Rommers, J., & Nieuwland, M. S. (2020). Definitely saw it coming? The dual nature of the pre-nominal prediction effect. *Cognition*, *204*, 104335. https://doi.org/10.1016/j.cognition.2020.104335

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multi-level hierarchical models* (Vol. 1). Cambridge University Press.

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210x.12504

Guerra, E., Nicenboim, B., & Helo, A. V. (2018). *A crack in the crystal ball: Evidence against pre-activation of gender features in sentence comprehension* [Poster]. Architectures and Mechanisms for Language Processing (AMLaP), Berlin, Germany.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22, pp. 193–225). Academic Press.

Häuser, K. I., Demberg, V., & Kray, J. (2019). Effects of aging and dual-task demands on the comprehension of less expected sentence continuations: Evidence from pupillometry. *Frontiers in Psychology*, *10*, 709. https://doi.org/10.3389/fpsyg.2019.00709

Huettig, F., & Guerra, E. (2019). Effects of speech rate, preview time of visual context, and participant instructions reveal strong limits on prediction in language processing. *Brain Research*, *1706*, 196–208. https://doi.org/10.1016/j.brainres.2018.11.013

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

Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience*, *31*(1), 19–31. https://doi.org/10.1080/23273798.2015.1072223

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

Ito, A., Martin, A. E., & Nieuwland, M. S. (2017). How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Language, Cognition and Neuroscience*, *32*(8), 954–965. https://doi.org/10.1080/23273798.2016.1242761

Jegerski, J., & VanPatten, B. (Eds.). (2014). *Research methods in second language psycholinguistics*. Routledge.

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228–238. https://doi.org/10.1037/0096-3445.111.2.228

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189–217. https://doi.org/10.1037/0096-3445.133.2.189

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*(1–2), 262–284. https://doi.org/10.1080/09541440340000213

Kochari, A. R., & Flecken, M. (2019). Lexical prediction in language comprehension: A replication study of grammatical gender effects in Dutch. *Language, Cognition and Neuroscience*, *34*(2), 239–253. https://doi.org/10.1080/23273798.2018.1524500

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

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lakens, D. (2021). Sample Size Justification. *PsyArXiv*. https://doi.org/10.31234/osf.io/9d3yf

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

Levy, R. (2013). Memory and surprisal in human sentence comprehension. In R. P. G. van Gompel (Ed.), *Sentence processing* (pp. 78–114). Psychology Press.

Lindenberger, U. (2014). Human cognitive aging: Corriger la fortune? *Science*, *346*(6209), 572–578.

Lorenz, C., & Kray, J. (2019). Are mid-adolescents prone to risky decisions? The influence of task setting and individual differences in temperament. *Frontiers in Psychology*, *10*, 1497. https://doi.org/10.3389/fpsyg.2019.01497

Lowder, M. W., Choi, W., Ferreira, F., & Henderson, J. M. (2018). Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive Science*, *42*, 1166–1183. https://doi.org/10.1111/cogs.12597

Martin, C. D., Thierry, G., Kuipers, J.-R., Boutonnet, B., Foucart, A., & Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. *Journal of Memory and Language*, *69*(4), 574–588. https://doi.org/10.1016/j.jml.2013.08.001

Mitchell, D. C., & Green, D. W. (1978). The effects of context and content on immediate processing in reading. *Quarterly Journal of Experimental Psychology*, *30*(4), 609–636. https://doi.org/10.1080/14640747808400689

Nicenboim, B., Vasishth, S., & Rösler, F. (2020). Are words pre-activated probabilistically during sentence comprehension? Evidence from new data and a Bayesian random-effects meta-analysis using publicly available data. *Neuropsychologia*, *142*, 107427. https://doi.org/10.1016/j.neuropsychologia.2020.107427

Nieuwland, M. S. (2019). Do 'early' brain responses reveal word form prediction during language comprehension? A critical review. *Neuroscience & Biobehavioral Reviews*, *96*, 367–400. https://doi.org/10.1016/j.neubiorev.2018.11.019

Nieuwland, M. S., Arkhipova, Y., & Rodríguez-Gómez, P. (2020). Anticipating words during spoken discourse comprehension: A large-scale, pre-registered replication study using brain potentials. *Cortex*, *133*, 1–36. https://doi.org/10.1016/j.cortex.2020.09.007

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D. J., Rousselet, G. A., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., … Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife*, *7*, 33468. https://doi.org/10.7554/elife.33468.024

Otten, M., & Van Berkum, J. J. A. (2008). Discourse-based word anticipation during language processing: Prediction or priming? *Discourse Processes*, *45*(6), 464–496. https://doi.org/10.1080/01638530802356463

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

Pichora-Fuller, M. K. (2008). Use of supportive context by younger and older adult listeners: Balancing bottom-up and top-down information processing. *International Journal of Audiology*, *47*(sup2), S72–S82. https://doi.org/10.1080/14992020802307404

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America*, *97*(1), 593–608. https://doi.org/10.1121/1.412282

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, *144*(10), 1002–1044. https://doi.org/10.1037/bul0000158

R Core Team. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C., & Pollatsek, A. (2006). The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and Aging*, *21*(3), 448–465. https://doi.org/10.1037/0882-7974.21.3.448

Ryskin, R., Levy, R. P., & Fedorenko, E. (2020). Do domain-general executive resources play a role in linguistic prediction? Re-evaluation of the evidence and a path forward. *Neuropsychologia*, *136*, 107258. https://doi.org/10.1016/j.neuropsychologia.2019.107258

Schmitt, H., Ferdinand, N. K., & Kray, J. (2014). Age-differential effects on updating cue information: Evidence from event-related potentials. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(3), 1115–1131. https://doi.org/10.3758/s13415-014-0268-9

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, *82*, 1–17. https://doi.org/10.1016/j.jml.2015.02.004

Stowe, L. A., Kaan, E., Sabourin, L., & Taylor, R. C. (2018). The sentence wrap-up dogma. *Cognition*, *176*, 232–247. https://doi.org/10.1016/j.cognition.2018.03.011

Tun, P. A., & Wingfield, A. (1994). Speech recall under heavy load conditions: Age, predictability, and limits on dual-task interference. *Aging, Neuropsychology, and Cognition*, *1*(1), 29–44. https://doi.org/10.1080/09289919408251448

Urbach, T. P., DeLong, K. A., Chan, W.-H., & Kutas, M. (2020). An exploratory data analysis of word form prediction during word-by-word reading. *Proceedings of the National Academy of Sciences*, *117*(34), 20483–20494. https://doi.org/10.1073/pnas.1922028117

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443–467.

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: An event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience*, *16*(7), 1272–1288. https://doi.org/10.1162/0898929041920487

Wingfield, A., Aberdeen, J. S., & Stine, E. A. L. (1991). Word onset gating and linguistic context in spoken word recognition by young and elderly adults. *Journal of Gerontology*, *46*(3), 127–129. https://doi.org/10.1093/geronj/46.3.p127

Wingfield, A., Poon, L. W., Lombardi, L., & Lowe, D. (1985). Speed of processing in normal aging: Effects of speech rate, linguistic structure, and processing time. *Journal of Gerontology*, *40*(5), 579–585. https://doi.org/10.1093/geronj/40.5.579

Wingfield, A., & Stine-Morrow, E. A. L. (2000). Language and speech. In F. I. M. Craik & T. A. Salthouse (Eds.), *The handbook of aging and cognition* (2nd ed., pp. 359–416). Erlbaum.

Wlotko, E. W., Federmeier, K. D., & Kutas, M. (2012). To predict or not to predict: Age-related differences in the use of sentential context. *Psychology and Aging*, *27*(4), 975–988. https://doi.org/10.1037/a0029206

Wlotko, E. W., Lee, C.-L., & Federmeier, K. D. (2010). Language of the aging brain: Event-related potential studies of comprehension in older adults. *Language and Linguistics Compass*, *4*(8), 623–638. https://doi.org/10.1111/j.1749-818x.2010.00224.x

Yan, S., Kuperberg, G. R., & Jaeger, T. F. (2017). Prediction (or not) during language processing. A commentary on Nieuwland et al. (2017) and DeLong et al. (2005). *BioRxiv*. https://doi.org/10.1101/143750

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1), 3–14. https://doi.org/10.1111/j.2041-210x.2009.00001.x

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022

Supplementary Materials

Peer Review History

Download: https://collabra.scholasticahq.com/article/36945-hedging-bets-in-linguistic-prediction-younger-and-older-adults-vary-in-the-breadth-of-predictive-processing/attachment/94032.docx?auth_token=WhkJP5fieupX8qD-a5xV

Supplement

Download: https://collabra.scholasticahq.com/article/36945-hedging-bets-in-linguistic-prediction-younger-and-older-adults-vary-in-the-breadth-of-predictive-processing/attachment/94033.docx?auth_token=WhkJP5fieupX8qD-a5xV

Downloaded from http://online.ucpress.edu/collabra/article-pdf/8/1/36945/727807/collabra_2022_8_1_36945.pdf by University of Saarland user on 22 November 2022