The role of language-specific and domain-general working memory resources in predictive language processing

Tovah Irwin (irwint@g.ucla.edu)
Department of Psychology, University of California, Los Angeles
Los Angeles, CA 90095 United States

Idan Blank (iblank@psych.ucla.edu)
Department of Psychology, University of California, Los Angeles
Los Angeles, CA 90095 United States

Abstract

A core aspect of language comprehension is predictive processing, which supports real-time inferences under uncertainty. Real-time prediction is constrained by the mind's limited working memory (WM) resources, which are required for maintaining the context that supports prediction, and for reallocating degrees of belief across inferred interpretations as new pieces of information (e.g., words) are perceived. What is the nature of this WM resource? Is it language-specific or shared between language and other cognitive domains? Do both domain-general and domain-specific resources support prediction? Here, we study this question using an individual differences approach. We collected self-paced reading times of naturalistic paragraphs in English and measured, for each participant, their domain-general WM (backwards digit span) and linguistic WM (reading span). We quantified predictive processing as the relationship between surprisal and reading time. We found that surprisal influenced reading times more strongly in participants with (1) *stronger* domain-general WM, but (2) *weaker* linguistic WM, although the latter relationship was less reliable. Our results indicate that domain-general WM could contribute to predictive processing during language comprehension. We discuss several theoretical interpretations of this finding, we well as potential reasons for the discrepancy between our results and past studies.

Keywords: Language; Comprehension; working memory; executive functions; predictive processing

Introduction

Language comprehension is incremental: we do not wait for the end of an utterance to start extracting its meaning. While an utterance unfolds, the mind must deal with uncertainty due to incomplete input, and thus makes predictive inferences about potential meanings prior to receiving all the relevant evidence (Kuperberg & Jaeger, 2016; Ryskin & Nieuwland, 2023). With every new piece of information (e.g., each additional word), the mind reallocates probabilities between potential meanings, i.e., their degrees of activation change. This costly reallocation is a major determinant of processing difficulty (Hale, 2001; Levy, 2008): it registers behaviorally, as slowed-down reading (Shain, 2024; Smith & Levy, 2013) and neurally, as increased brain activity (Michaelov et al., 2024; Shain et al., 2020). Predictive processing is robustly evident across a variety of languages (Wilcox et al., 2023).

Generating predictions requires working memory resources: the mind must maintain the prior context (Futrell et al., 2020), as well as parallel potential interpretations, and continuously update their probabilities. Yet the exact nature of these WM resources is under-specified in theories of predictive processing. This is the focus of the current study.

Traditional models of WM resources posit a WM resource for all phonological material, i.e., the phonological loop, which is distinct from the visuospatial sketchpad for visual processing (Baddeley & Hitch, 1974), and from the episodic buffer (Baddeley, 2000). The phonological loop is assumed to process both linguistic representations (e.g., sentences, which have structure and compositional meaning) and other, non-linguistic but phonologically-encoded representations (e.g., lists of unconnected words or digits, which have neither structure nor compositional meaning). However, later models have introduced a distinction between this domain-general phonological (or "verbal") WM and a separate, language-specific WM system (Martin, 2021). This linguistic (or, in some accounts, "semantic") resource is recruited during comprehension and production, but not during processing of other phonologically-encoded material.

Evidence for this distinction comes from studies finding that reading comprehension correlates more strongly with performance on tasks of WM that include sentence materials ("reading span", see below), and less strongly with tasks of WM for lists of digits or unrelated words (digit / word span) (Daneman & Carpenter, 1980; Daneman & Merikle, 1996). Moreover, brain damage can cause selective deficits on tasks that tap phonological vs. semantic demand (Caplan & Waters, 1999; Martin, 2021; Martin & Schnur, 2019). Finally, neuroimaging studies have demonstrated a dissociation between the two types of WM. On the one hand, WM tasks involving digits recruit a set of domain-general brain regions that are also recruited by a variety of other cognitively demanding stimuli and tasks (Fedorenko et al., 2013); in contrast, WM load during story listening (as measured by the length of syntactic dependencies) recruits a distinct set of language-specific brain regions (Shain et al., 2022). Therefore, language-specific vs. general WM resources are implemented in dissociated neural mechanisms (Fedorenko & Shain, 2021).

Language comprehension (as measured using both scores on standardized tests and performance on experimental tasks) correlates with (Daneman & Merikle, 1996). However, these effects are inconsistent (e.g., Roberts & Gibson, 2002; Van

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

Dyke et al., 2014), and the contributions of the two WM resources specifically to predictive processing remain unclear. Some studies have found that when reading isolated sentences, individuals with higher (vs. lower) linguistic WM, as measured by reading span, show increased sensitivity to word predictability in their reading times (Farmer et al., 2017), eye-movements towards semantically predictable visual referents in a visual word paradigm (Li & Qu, 2023), and EEG/ERP markers of prediction and integration (Ding et al., 2023; Van Patten et al., 1997; Yang et al., 2020). However, other studies did not find such effects (Cutter et al., 2023; Nieuwland & Van Berkum, 2006) (perhaps because their tasks did not sufficiently tax predictive mechanisms). Similarly, whereas some studies have found that predictive processing benefits from higher non-linguistic WM (Huettig & Janse, 2016; Koch et al., 2023; Nicenboim et al., 2015), these effects might not be robust (Hintz et al., 2024). Moreover, fMRI studies suggest that linguistic prediction recruits language-specific brain regions, rather than regions that implement domain-general WM (Shain et al., 2022). Finally, studies implicating either kind of WM in predictive processing have been challenged by alternative explanations (MacDonald & Christiansen, 2002; Ryskin et al., 2020).

Given these inconsistent findings, here we test which WM resources—language-specific and/or domain-general—support predictive language processing. We use an individual differences approach, asking whether people vary in their predictive processing as a function of their linguistic and/or domain-general WM. We measured each type of WM using tasks that are sufficiently reliable in capturing individual differences: reading span (Daneman & Carpenter, 1980; Waters, 1996) for linguistic/semantic WM, and backwards digit span (Wechsler, 2012) for domain-general/phonological WM. We measured predictive processing during real-time comprehension via self-paced reading of real, multi-sentence paragraphs from the internet, which are more naturalistic than materials in most previous studies. Specifically, we quantified how each participant's reading times were influenced by the unpredictability of each word given its context, as quantified by surprisal (see below). We predicted that, if a certain WM resource supports predictive language processing, we would find an interaction between span scores on the corresponding task and the effect of surprisal: namely, participants with higher span would exhibit reading times that are more strongly modulated by surprisal.

Methods

All tasks were administered online, using PsychoJS (Peirce et al., 2019) hosted on Pavlovia.org, in a 120min session that included forced 1 min breaks every 5 paragraphs and optional breaks between tasks as needed. For all tasks, text was presented in white, monospace font on a dark grey background. Participants were instructed to complete the tasks in a distraction-free environment.

Participants

We recruited English speaking undergraduate students at the University of California, Los Angeles (*N*=78). Following exclusions, our final sample had 62 participants (51 women, mean age=20.7). To avoid potential differences in domain-general WM resources (and, more generally, executive functions) between monolingual and bilingual populations (Lehtonen et al., 2018), we screened participants using the Language Experience and Proficiency Questionnaire (Marian et al., 2007), excluding those who reported second language exposure before the age of 7. All participants received course credit for participation.

Self-Paced Reading Task

Materials. Self-paced reading materials were created from paragraphs from English Wikipedia articles on a wide range of topics. Based on subjective judgment, articles were selected to be (1) engaging, and (2) devoid of highly technical concepts or terminology. Paragraphs were edited by a native English speaker to ensure that the majority of the text was words (not, e.g., numbers, links, or citations). The resulting stimulus set included 20 paragraphs, with an average of 200 words each, for a total of 4,001 tokens per participant.

Procedure. Paragraphs were presented one sentence at a time. When a sentence first appeared, all words (including spaces) were masked with the character "#". When the participant pressed the spacebar, the first word was revealed; upon each successive key press, the previously presented word was masked again, and the next word was revealed. The times between key-presses served as the dependent variable.

Participants were told to read carefully, and were informed that they would be answering comprehension questions. Each paragraph was followed by a four-alternative, forced-choice comprehension question. If a participant did not correctly answer a question, all self-paced reading times (SPRTs) for the corresponding paragraph were excluded. We also excluded 11 participants whose accuracy across paragraphs was lower than 50% (chance accuracy = 25%).

Individual Differences Measures

Both WM span tasks required storage of items as well as manipulation of information in WM. They differed in terms of the stimuli presented and their required manipulations.

Domain-general Working Memory: Backwards Digit Span. Each trial began with a fixation cross (1000ms), after which participants were presented with a list of digits, each for 1s (ISI = 1ms). They were instructed to retain the digits and, at the end of the list, type them backwards. Participants were given three practice trials. Test trials started at a span of 3. If participants successfully completed 3 trials with a given span, the span was increased by 1; the task terminated when they performed incorrectly on 2 out of 3 trials in a given span. A participant's digit span was calculated as the maximum span recalled correctly for 2 trials during the task.

Linguistic Working Memory: Reading Span. We administered a modified version of the reading span task,

following Waters (1996). On each trial, participants were presented with a series of 2-6 sentences, one by one. After each sentence, participants made a binary plausibility judgment (e.g., plausible: "The message was delivered by the helper"; implausible: "The helper was delivered by the message"). In addition to these judgements, participants were also asked to retain the final word of each sentence in the trial and, at the end of the trial, to list the sentence-final words they recalled in any order. Across trials, each participant saw a total of 100 sentences, with 5 trials of each span . The order of span lengths was randomized. A participant's reading span was calculated as the total number of words recalled across all 100 trials (Friedman & Miyake, 2005). We note that this task is not a "pure" measure of linguistic WM (similarly to any other task); it includes artificial demands, and such demands are known to recruit cognitive systems that do not support language comprehension in more natural circumstances (e.g., Diachek et al., 2020). Nonetheless, compared to backwards digit span, reading span appears to recruit linguistic computations more heavily.

Analysis

Quantifying Linguistic (Un)predictability via Surprisal. A theoretical measure of processing difficulty associated with updating probabilistic inferences (predictions) during online language comprehension is surprisal—the negative log probability of a word given its preceding context (Hale, 2001; Levy, 2008). We computed surprisal for each word in our paragraphs, given the entire preceding context of its paragraph, based on GPT2-small (Radford et al., 2019). Surprisal metrics from GPT2 and similar models correlate with human language processing difficulty (including during naturalistic reading) as reflected in both behavioral (Oh & Schuler, 2023; Shain, 2024; Wilcox et al., 2023) and neural (Michaelov et al., 2024) measures.

Control Variables. We included three control metrics known to influence reading time: (1) word length in characters; (2) word position in a sentence; and (3) word frequency, from the SUBTLEXus corpus (Brysbaert & New, 2009), which was log-transformed following White et al., (2018) (words with 0 frequency were smoothed to 1).

Reliability of Predictive Processing. Our measure of a participant's predictive processing was the relationship between surprisal and SPRTs. We tested the reliability of individual differences in this measure, which is an upper-bound for correlating it with other individual differences measures (the span tasks). To this end, we split the paragraphs in two (odd vs. even) and, for each participant and data split, estimated a fixed-effects model that predicted SPRTs from surprisal of the current word *w* (surprisal_w), surprisal of the previous word *w-1* (surprisal_{w-1}), and the control variables. Each data split thus yielded one beta coefficient per participants for the effect of surprisal_w and one coefficient for the effect of surprisal_{w-1}.

We correlated the beta coefficients from the two data splits across participants, which showed moderate reliability for of surprisal_w (r=0.58), but sufficiently high reliability for surprisal_{w-1} (r=0.80). For each participant, we also computed the absolute difference between their two beta coefficients of surprisal_w from the two data splits, and removed 5 participants for whom this difference was more than 2.5 SDs away from the average difference in the sample.

Statistical Model. We excluded the first word in each sentence and all SPRTSs that were ±2.5 SD away from each participant's mean. We then analyzed SPRTs in a linear, mixed effects model, using the lme4 (Bates et al., 2015) package in R. We included fixed effects of (1) surprisal_w , (2) surprisal_{w-1}, to quantify spillover effects, (3) backwards digit span (BDS; *z*-scored across participants), (4) reading span (RS; *z*-scored), and (5) control variables. To test our main question about the relationship between WM resources and predictive processing, we included interaction terms between BDS and surprisal_w, between RS and surprisal_w, and between each span and surprisal_{w-1}. We included both surprisal measures because effects of surprisal on SPRTs are often found on subsequent words (Smith & Levy, 2013), and because our participant-wise estimate of predictive processing was more reliable at word *w-1*. If a certain WM resource supports prediction, then participants with higher span would show a stronger surprisal-SPRT relationship, i.e., a stronger effect of contextual predictability on reading behavior (see also Discussion). Random intercepts were included by participant and by paragraph (Baayen et al., 2008). For a word *w*, the overall formula is therefore:

$$
\begin{aligned}
SPRT \sim\ & Surprisal_w + Surprisal_{w-1} + BDS + RS + \\
& \boldsymbol{BDS \times Surprisal_w} + \boldsymbol{RS \times Surprisal_w} + \\
& \boldsymbol{BDS \times Surprisal_{w-1}} + \boldsymbol{RS \times Surprisal_{w-1}} + \\
& LogFrequency + WordLength + WordPosition + \\
& (1 \mid Participant) + (1 \mid Paragraph)
\end{aligned}
$$

Effects of WM on Offline Comprehension. We attempted to replicate findings that reading span scores correlated with reading comprehension measures (Daneman & Carpenter, 1980; Waters, 1996). Whereas SPRTs were analyzed only for paragraphs whose comprehension questions were answered correctly, we also analyzed individual differences in comprehension accuracy across all paragraphs. To this end, we regressed participants' accuracy on the comprehension questions (% correct) against both their BDS and RS.

Results

The main results are presented in **Table 1**. As a validation of our SPRT measure, we replicated the well-known slowdowns for longer words and for less frequent words (Kapteijns & Hintz, 2021; Rayner, 1998; Shain, 2024). We also found a significant effect of surprisal of the current word, and a stronger spillover effect from the surprisal of the previous word, as has been previously reported (Shain et al., 2024; Smith & Levy, 2013). The difference between these two surprisal effects was not significant ($t_{(60)}$= -1.027, p= 0.5446).

Our measures of domain-general WM (backwards digit

span) and linguistic WM (reading span, which had a split-half correlation of r=0.82 in our sample) showed a weak-to-moderate correlation of r=0.3 (p=0.0085)[1] across participants. To account for this collinearity, fixed effects and interactions were tested by ablative model comparisons in likelihood ratio tests. Neither span reliably affected SPRTs overall, likely because any such effects were absorbed into the by-participant random intercept. However, our main interest was in their influence on predictive processing, i.e., on the relationship between surprisal and SPRTs. At the critical word, we found an interaction between backwards digit span and surprisal (β=0.88, $\chi^2_{(1)}$= 59.56, p<10^{-13}), such that participants with *higher* span showed *stronger* effects of surprisal on SPRTs (**Fig. 1A**). We also found an interaction of reading span and surprisal, but in an unexpected direction (β=-0.33, $\chi^2_{(1)}$ = 8.851, p=0.0029): participants with *higher* span showed *weaker* effects of surprisal on SPRTs (**Fig. 1B**). Because difference in significance does not indicate a significant difference, we directly compared the difference in absolute magnitude between these two interaction effects, which was significant (z=4.05, p=0.0001). The interaction between BDS and surprisal remained significant in the spillover region (β=0.37, $\chi^2_{(1)}$=10.93, p=0.0009), but the one between RS and surprisal did not (β=0.02, $\chi^2_{(1)}$=0.0287, p=0.86) (**Fig. 1C-D**). These two interaction terms significantly differed from one another (z=2.65, p=0.016).

Table 1: Linear, mixed-effects model predicting SPRTs

	Estimate	SE	p
(Intercept)	270.50	15.250	$<10^{-15}$
Word length	1.59	0.195	$<10^{-15}$
Word position	0.0009	0.034	0.975
Log frequency	-1.12	0.158	$<10^{-11}$
Surprisal_{w}	0.81	0.127	$<10^{-9}$
Surprisal_{w-1}	3.27	0.106	$<10^{-15}$
BDS	8.92	12.250	0.4694
RS	12.65	12.250	0.3059
$\text{BDS}\times\text{Surprisal}_{w}$	0.88	0.115	$<10^{-13}$
$\text{RS}\times\text{Surprisal}_{w}$	-0.33	0.112	0.0029
$\text{BDS}\times\text{Surprisal}_{w-1}$	0.37	0.113	0.0009
$\text{RS}\times\text{Surprisal}_{w-1}$	0.02	0.110	0.8655

The analysis of offline comprehension accuracy revealed that participants with *higher* backwards digit span showed higher accuracy (β=0.03, SE=0.01, p=0.02, one-tailed test), but reading span did not influence accuracy (β=0.02, SE=0.013, p=0.14). However, the difference between these two effects was not significant ($t_{(58)}$=0.526, p=0.601).

Discussion

This study asked what types of WM resources support predictive processing during incremental comprehension. We measured self-paced reading times of natural texts, quantified prediction as the relationship between word surprisal (extracted from GPT2) and SPRTs, and tested how this relationship was influenced by individual variation in two WM resources: a more domain-general/phonological WM, estimated using a backwards digit span task, and a more linguistic/semantic WM, estimated with a reading span task. We found that individuals with *higher* domain-general WM capacity showed *stronger* effects of surprisal on SPRTs, i.e., exhibit "better" prediction and/or rely more strongly on predictive processes. We also found a puzzling pattern, whereby participants with *weaker* linguistic WM capacity showed *stronger* effects of surprisal on SPRTs. However, this effect occurred only at the critical word—where overall signatures of predictive processing were descriptively weaker, and not at the spillover word—where predictive processing was more strongly evident. Finally, we found that domain-general but not linguistic WM capacity significantly predicted offline comprehension accuracy.

Our findings support the hypothesis that the main WM resources supporting linguistic prediction are domain-general in nature, as some accounts of language processing posit (either explicitly or implicitly) (Abney & Johnson, 1991; Huettig & Mani, 2016; Pickering & Gambi, 2018; Rasmussen & Schuler, 2018; Resnik, 1992; Smith & Levy, 2013; van Schijndel et al., 2013). They are consistent with past reports that individuals with higher domain-general ("phonological" / "verbal") WM show stronger signatures of predictive processing (Huettig & Janse, 2016; Koch et al., 2023; Nicenboim et al., 2015). Nonetheless, our results are inconsistent with more recent claims that predictive processing is a "canonical" cognitive computation that, for many domains, is carried out by specialized systems (Bastos et al., 2012; Bubic et al., 2010; Keller & Mrsic-Flogel, 2018; Singer et al., 2018; Wacongne et al., 2011)—in our case, a language-specific system with its own pool of WM resources (Fedorenko & Shain, 2021). Indeed, in contrast to our findings, several prior studies investigating reading of isolated sentences found stronger signatures of predictive processing in individuals with higher linguistic WM as measured by reading span (Ding et al., 2023; Farmer et al., 2017; Li & Qu, 2023; Ness & Meltzer-Asscher, 2018; Van Patten et al., 1997; Yang et al., 2020) although this pattern was not observed by other studies (Cutter et al., 2023; Nieuwland & Van Berkum, 2006). Moreover, neuroimaging studies have suggested that the domain-general brain system that is recruited during tasks that tax WM for digit strings (Fedorenko et al., 2013) does not support language prediction: its activity does not closely track the unfolding linguistic input during story listening (Blank & Fedorenko, 2017), and does not correlate with SPRTs (Wehbe et al., 2021) or with GPT2 surprisal during listening (Shain et al., 2022). Instead, these effects all arise in a language-specific

[1] Reading span showed a weak correlation with performance on the Simon task, which indexes inhibitory control r=0.22, p=0.044), although this test does not survive correction for multiple comparisons. Backwards digit span was not significantly correlated with performance on this task (r=0.09, p=0.24).

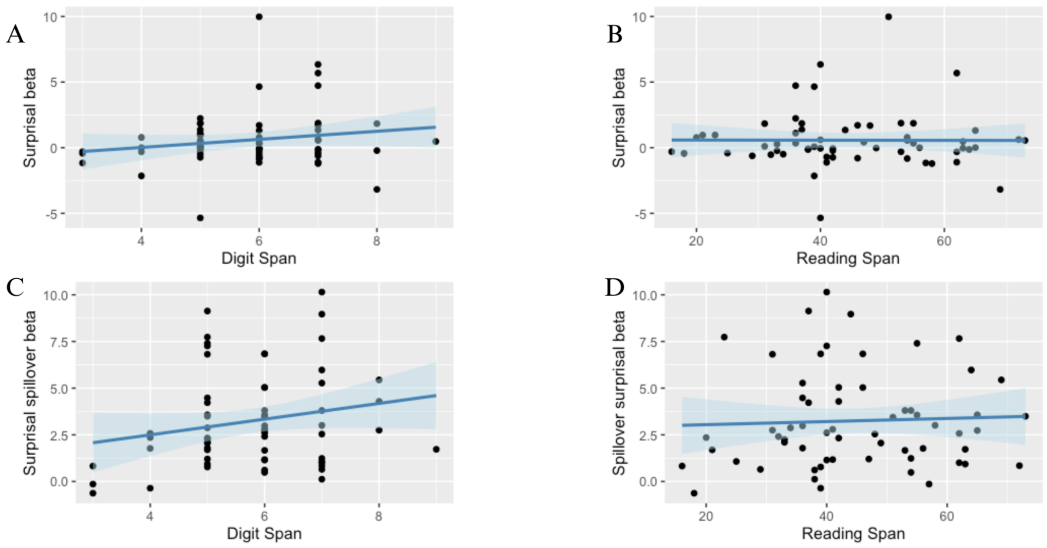

Figure 1: Interaction between measures of WM and predictive processing. In each panel, the *x*-axis shows participant scores on either the backwards digit span (panels A, C) or reading span (panels B, D). The *y*-axis shows each participant's beta coefficient of the relationship between surprisal of the current word (panels A, B) or previous word (panels C, D) and SPRTs. Each participant's coefficient was estimated in a separate model regressing their SPRTs against surprisal and the control variables as fixed effects.

brain system, which also registers WM costs during incremental building of linguistic structure, and thus putatively implements linguistic WM (Shain et al., 2022).

Across individuals in our study, domain-general WM capacity co-varied with the relationship between surprisal and SPRTs. The strength of this relationship is typically taken to index predictive processing. What does it mean for some individuals to show "stronger" predictive processing? Several interpretations are possible. First, the incremental probabilistic inferences of these individuals might consider (i.e., activate) a wider array of potential sentence interpretations (e.g., possible but very unlikely ones). Thus, upon encountering a new word, more potential interpretations would have to be pruned (under some formulations of this idea). The greater updating of the probability distribution over interpretations—which is the cognitive mechanism underlying linguistic prediction (Levy, 2008)—would incur a higher processing cost, leading to a steeper slope for the relationship between surprisal and SPRTs. Under this view, higher WM capacity allows one to entertain a larger number of potential interpretations.

Second, when combining incoming bottom-up linguistic signals with top-down predictions, individuals who are "stronger predictors" may weigh the latter more heavily than "weaker predictors". In other words, predictions would more strongly "bias" linguistic representations by being more informative, i.e., with lower entropy (unlike "weaker" predictors, who assign more uniform probabilities to interpretations, resulting in higher entropy). Whereas this is a sensible cause of individual differences, the ability of domain-general mechanisms to "bias" representations in other cognitive systems is typically thought of as dependent on cognitive control, not WM capacity (but see Engle, 2002).

The third interpretation stems from the fact that predictions must derive from a mental model of the statistical properties of linguistic strings. Because we quantify predictive processing as the relationship between SPRTs and surprisal extracted from GPT2-small, individuals who show stronger signatures of predictive processing might have a mental model that is more aligned with the statistics learned by GPT2-small. Thus, individuals differ not in how "strongly" they rely on predictions, but in their mental model (e.g., Ryskin et al., 2020). To the extent that GPT2-small captures rich and detailed information about the statistics of language, individuals whose mental model is more aligned with our surprisal measure make predictions that are more accurate. Those individuals would have likely had more exposure to the type of statistics relevant for this experiment, i.e., written texts. However, reading experience has been linked not to performance on backwards digit span, but to performance on reading span (MacDonald & Christiansen, 2002), and we did not find that higher reading span modulated the surprisal-SPRT relationship in the expected direction.

Perhaps differences in (domain-general) WM capacity are not related to the amount of reading experience, but to the ability to extract language statistics and integrate them into one's mental model. For example, a larger WM buffer may be able to store a less noisy context, a longer context, and/or context that combines more information sources (e.g. Boudewyn et al., 2013; Caplan & Waters, 1999; Nicenboim et al., 2015, 2016; Waters & Caplan, 1996), and thus compute statistics about the distributional patterns of language that are more accurate. Moreover, maintaining longer or more complex contexts aides not only learning but also real-time

predictions, because "learning" and "processing" are intertwined.

In addition to a positive relationship between backwards digit span and surprisal in SPRTs, our results suggested a negative relationship between reading span and surprisal effects. There is at least one account by which higher WM capacity could lead to a weaker relationship between surprisal and SPRTs. Here, WM capacity (perhaps with processing speed) influences the efficiency with which processing resources are re-allocated between potential interpretations when a new word is encountered. If two individuals make the same predictions but differ in WM capacity, the one with the higher capacity would be able to shift cognitive resources more quickly, leading to a lower processing cost per unit of surprisal (i.e., a smaller beta coefficient relating surprisal and SPRTs). The reading span task does require the ability to control interference between reading a new sentence to judge its plausibility and rehearsing a list of sentence-final words, and such linguistic control may be important for recovering from cases of high surprisal (for a similar explanation, see Nicenboim et al., 2015). However, this is unlikely. It would not be parsimonious to conclude that one WM resource (domain-general) supports certain aspects of the prediction process (e.g., generating or maintaining predictions) whereas another WM resource (language-specific) supports a different aspect (e.g., updating).

We emphasize that the negative relationship between reading span and surprisal effects on SPRTs might not reflect a true effect. This relationship did not emerge in the spillover region, where estimates of individual differences in predictive processing were most reliable.

Our findings that domain-general WM supports predictive processing, but linguistic WM does not, contradicts strong evidence that a language-specific cognitive system—but not domain-general ones—stores linguistic knowledge and processes linguistic input. One way to reconcile our findings with those theories is to challenge the construct validity of the span tasks used here.

First, both backwards digit span and reading span tasks might capture not only phonological and semantic WM respectively, but also parts of an individual's capacity that apply across cognitive systems, such as neural efficiency, or attentional control (Draheim et al., 2022). Such properties would affect language-specific mechanisms, among other systems, and could support linguistic prediction. Second, reading span might not capture linguistic WM but, rather, reading experience (MacDonald & Christiansen, 2002). Whereas reading experience could support predictive processing, our study might not be ideal for revealing it, because (1) individuals in our homogeneous sample (college students, mostly in a Psychology Department) might not sufficiently differ from one another in their experience, and (2) effects of reading experience on predictive processing might be easier to detect in stimuli that are explicitly manipulated to be highly unpredictable (very surprising words or structures are, by definition, rare in real texts such as ours). Alternatively, if the negative relationship we found between reading span and prediction is real, perhaps reading experience improves visual word recognition (e.g., Gordon et al., 2020), leading to *weaker* reliance on predictive processing. Indeed, surprisal effects on reading times could reflect top-down support for visual word recognition (Carpenter & Williams, 1995; Norris, 2006; Norris, 2009). To address concerns about construct validity, a larger battery of tasks is needed to measure each type of WM resource, with composite scores that are more likely to reflect the shared mechanisms across each set of tasks (e.g., James et al., 2018).

Third, whereas surprisal estimates from GPT2 correlated with human reading times, GPT2 does not suffer from the same memory constraints when performing predictions. It can store a maximum context of 1024 tokens—much longer than the longest tokenized paragraph in our self-paced reading stimuli. Therefore, GPT2's estimates of a word's surprise in context might leave out aspects of a word's (un)predictability that depend on decaying activation of context in WM (Futrell et al., 2020). Instead, our surprisal estimates may capture predictions that are WM-independent, like those based in common sense or local word co-occurrences. Future studies could estimate surprisal using models that explicitly implement a WM-like mechanism.

An alternate (but, in our opinion, unlikely) interpretation of the interaction between reading span and surprisal is that linguistic WM capacity supports the incremental building of an accurate sentence representation, i.e., a process of "integration" (Futrell et al., 2020). Perhaps individuals who are better at bottom-up representation building do not need to rely as strongly on top-down predictive processing. However, these two processes are highly related, and might not "trade-off" against each other, because syntactic dependencies correspond to words that are highly predictive of one another (Futrell et al., 2019); therefore, better integration of words into a sentence can support more accurate predictions.

In sum, our findings provide evidence that domain-general working memory (WM), as indexed by backwards digit span, may play an important role in supporting predictive processing during incremental language comprehension. At the same time, our results appear inconsistent with the notion that a linguistic WM resource supports predictive processing, although future work should improve the measurement of this domain-specific capacity (whatever abilities are measured by the reading span task do not appear to correlate with predictive processing as quantified here). By characterizing the cognitive mechanisms that support linguistic prediction, our work contributes to broader theories of how memory and expectation interact in real-time language processing.

References

Abney, S. P., & Johnson, M. (1991). Memory requirements and local ambiguities of parsing strategies. *Journal of Psycholinguistic Research*, *20*(3), 233–250. https://doi.org/10.1007/BF01067217

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory*

and Language, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A. D., & Hitch, G. (1974). Working Memory. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 8, pp. 47–89). Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

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

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

Blank, I. A., & Fedorenko, E. (2017). Domain-General Brain Regions Do Not Track Linguistic Input as Closely as Language-Selective Regions. *Journal of Neuroscience*, *37*(41), 9999–10011. https://doi.org/10.1523/JNEUROSCI.3642-16.2017

Boudewyn, M. A., Long, D. L., & Swaab, T. Y. (2013). Effects of Working Memory Span on Processing of Lexical Associations and Congruence in Spoken Discourse. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00060

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

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

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences*, *22*(01). https://doi.org/10.1017/S0140525X99001788

Carpenter, R. H. S., & Williams, M. L. L. (1995). Neural computation of log likelihood in control of saccadic eye movements. *Nature*, *377*(6544), 59–62. https://doi.org/10.1038/377059a0

Cutter, M. G., Paterson, K. B., & Filik, R. (2023). Syntactic prediction during self-paced reading is age invariant. *British Journal of Psychology*, *114*(1), 39–53. https://doi.org/10.1111/bjop.12594

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*(4), 422–433. https://doi.org/10.3758/BF03214546

Diachek, E., Blank, I., Siegelman, M., Affourtit, J., & Fedorenko, E. (2020). The Domain-General Multiple Demand (MD) Network Does Not Support Core Aspects of Language Comprehension: A Large-Scale fMRI Investigation. *Journal of Neuroscience*, *40*(23), 4536–4550. https://doi.org/10.1523/JNEUROSCI.2036-19.2020

Ding, J., Zhang, Y., Liang, P., & Li, X. (2023). Modulation of working memory capacity on predictive processing during language comprehension. *Language, Cognition and Neuroscience*, *38*(8), 1133–1152. https://doi.org/10.1080/23273798.2023.2212819

Draheim, C., Pak, R., Draheim, A. A., & Engle, R. W. (2022). The role of attention control in complex real-world tasks. *Psychonomic Bulletin & Review*, *29*(4), 1143–1197. https://doi.org/10.3758/s13423-021-02052-2

Engle, R. W. (2002). Working Memory Capacity as Executive Attention. *Current Directions in Psychological Science*, *11*(1), 19–23. https://doi.org/10.1111/1467-8721.00160

Farmer, T. A., Fine, A. B., Misyak, J. B., & Christiansen, M. H. (2017). Reading Span Task Performance, Linguistic Experience, and the Processing of Unexpected Syntactic Events. *Quarterly Journal of Experimental Psychology*, *70*(3), 413–433. https://doi.org/10.1080/17470218.2015.1131310

Fedorenko, E., Duncan, J., & Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, *110*(41), 16616–16621. https://doi.org/10.1073/pnas.1315235110

Fedorenko, E., & Shain, C. (2021). Similarity of computations across domains does not imply shared implementation: The case of language comprehension. *Current Directions in Psychological Science*, *30*(6), 526. https://doi.org/10.1177/09637214211046955

Friedman, N. P., & Miyake, A. (2005). Comparison of four scoring methods for the reading span test. *Behavior Research Methods*, *37*(4), 581–590. https://doi.org/10.3758/BF03192728

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Cognitive Science*, *44*(3), e12814. https://doi.org/10.1111/cogs.12814

Futrell, R., Qian, P., Gibson, E., Fedorenko, E., & Blank, I. (2019). Syntactic dependencies correspond to word pairs with high mutual information. In K. Gerdes & S. Kahane (Eds.), *Proceedings of the Fifth International Conference on Dependency*

Linguistics (Depling, SyntaxFest 2019) (pp. 3–13). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-7703

Gordon, P. C., Moore, M., Choi, W., Hoedemaker, R. S., & Lowder, M. W. (2020). Individual differences in reading: Separable effects of reading experience and processing skill. *Memory & Cognition*, *48*(4), 553–565. https://doi.org/10.3758/s13421-019-00989-3

Hale, J. (2001). A Probabilistic Earley Parser as a Psycholinguistic Model. *Second Meeting of the North American Chapter of the Association for Computational Linguistics*. NAACL 2001. https://aclanthology.org/N01-1021/

Hintz, F., Voeten, C. C., Dobó, D., Lukics, K. S., & Lukács, Á. (2024). The role of general cognitive skills in integrating visual and linguistic information during sentence comprehension: Individual differences across the lifespan. *Scientific Reports*, *14*(1), 17797. https://doi.org/10.1038/s41598-024-68674-3

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

James, A. N., Fraundorf, S. H., Lee, E.-K., & Watson, D. G. (2018). Individual differences in syntactic processing: Is there evidence for reader-text interactions? *Journal of Memory and Language*, *102*, 155–181. https://doi.org/10.1016/j.jml.2018.05.006

Kapteijns, B., & Hintz, F. (2021). Comparing predictors of sentence self-paced reading times: Syntactic complexity versus transitional probability metrics. *PLOS ONE*, *16*(7), e0254546. https://doi.org/10.1371/journal.pone.0254546

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive Processing: A Canonical Cortical Computation. *Neuron*, *100*(2), 424–435. https://doi.org/10.1016/j.neuron.2018.10.003

Koch, E. M., Bulté, B., Housen, A., & Godfroid, A. (2023). The predictive processing of number information in subregular verb morphology in a first and second language. *Applied Psycholinguistics*, *44*(5), 750–783. https://doi.org/10.1017/S014271642300022X

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

Lehtonen, M., Soveri, A., Laine, A., Järvenpää, J., de Bruin, A., & Antfolk, J. (2018). Is bilingualism associated with enhanced executive functioning in adults? A meta-analytic review. *Psychological Bulletin*, *144*(4), 394–425. https://doi.org/10.1037/bul0000142

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

Li, X., & Qu, Q. (2023). Verbal working memory capacity modulates semantic and phonological prediction in spoken comprehension. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-023-02348-5

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, *109*(1), 35–54; discussion 55-74. https://doi.org/10.1037/0033-295x.109.1.35

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing Language Profiles in Bilinguals and Multilinguals. *Journal of Speech, Language, and Hearing Research*, *50*(4), 940–967. https://doi.org/10.1044/1092-4388(2007/067)

Martin, R. C. (2021). The Critical Role of Semantic Working Memory in Language Comprehension and Production. *Current Directions in Psychological Science*, *30*(4), 283–291. https://doi.org/10.1177/0963721421995178

Martin, R. C., & Schnur, T. T. (2019). Independent contributions of semantic and phonological working memory to spontaneous speech in acute stroke. *Cortex*, *112*, 58–68. https://doi.org/10.1016/j.cortex.2018.11.017

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2024). Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects. *Neurobiology of Language*, *5*(1), 107–135. https://doi.org/10.1162/nol_a_00105

Ness, T., & Meltzer-Asscher, A. (2018). Predictive Pre-updating and Working Memory Capacity: Evidence from Event-related Potentials. *Journal of Cognitive Neuroscience*, *30*(12), 1916–1938. https://doi.org/10.1162/jocn_a_01322

Nicenboim, B., Logačev, P., Gattei, C., & Vasishth, S. (2016). When High-Capacity Readers Slow Down and Low-Capacity Readers Speed Up: Working Memory and Locality Effects. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00280

Nicenboim, B., Vasishth, S., Gattei, C., Sigman, M., & Kliegl, R. (2015). Working memory differences in long-distance dependency resolution. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00312

Nieuwland, M. S., & Van Berkum, J. J. A. (2006). Individual differences and contextual bias in pronoun

resolution: Evidence from ERPs. *Brain Research*, *1118*(1), 155–167. https://doi.org/10.1016/j.brainres.2006.08.022

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*(2), 327–357. https://doi.org/10.1037/0033-295X.113.2.327

Norris, D. (2009). Putting It All Together: A Unified Account of Word Recognition and Reaction-Time Distributions. *Putting It All Together : A Unified Account of Word Recognition and Reaction-Time Distributions*, *116*(1), 207–219.

Oh, B.-D., & Schuler, W. (2023). Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of the Association for Computational Linguistics*, *11*, 336–350. https://doi.org/10.1162/tacl_a_00548

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

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

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (n.d.). *Language Models are Unsupervised Multitask Learners*.

Rasmussen, N. E., & Schuler, W. (2018). Left-Corner Parsing With Distributed Associative Memory Produces Surprisal and Locality Effects. *Cognitive Science*, *42*(S4), 1009–1042. https://doi.org/10.1111/cogs.12511

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Resnik, P. (1992). Left-corner parsing and psychological plausibility. *Proceedings of the 14th Conference on Computational Linguistics - Volume 1*, 191–197. https://doi.org/10.3115/992066.992098

Roberts, R., & Gibson, E. (2002). Individual Differences in Sentence Memory. *Journal of Psycholinguistic Research*, *31*(6), 573–598. https://doi.org/10.1023/A:1021213004302

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

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

Shain, C. (2024). Word Frequency and Predictability Dissociate in Naturalistic Reading. *Open Mind*, *8*, 177–201. https://doi.org/10.1162/opmi_a_00119

Shain, C., Blank, I. A., Fedorenko, E., Gibson, E., & Schuler, W. (2022). Robust Effects of Working Memory Demand during Naturalistic Language Comprehension in Language-Selective Cortex. *Journal of Neuroscience*, *42*(39), 7412–7430. https://doi.org/10.1523/JNEUROSCI.1894-21.2022

Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., & Fedorenko, E. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, *138*, 107307. https://doi.org/10.1016/j.neuropsychologia.2019.107307

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, *121*(10), e2307876121. https://doi.org/10.1073/pnas.2307876121

Singer, Y., Teramoto, Y., Willmore, B. D., Schnupp, J. W., King, A. J., & Harper, N. S. (2018). Sensory cortex is optimized for prediction of future input. *eLife*, *7*, e31557. https://doi.org/10.7554/eLife.31557

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

Van Dyke, J. A., Johns, C. L., & Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition*, *131*(3), 373–403. https://doi.org/10.1016/j.cognition.2014.01.007

Van Patten, C., Weckerly, J., McIsaac, H. K., & Kutas, M. (1997). Working memory capacity dissociates lexical and sentential context effects. *Psychological Science*, *8*(3), 238–242. https://doi.org/10.1111/j.1467-9280.1997.tb00418.x

van Schijndel, M., Exley, A., & Schuler, W. (2013). A Model of Language Processing as Hierarchic Sequential Prediction. *Topics in Cognitive Science*, *5*(3), 522–540. https://doi.org/10.1111/tops.12034

Wacongne, C., Labyt, E., van Wassenhove, V., Bekinschtein, T., Naccache, L., & Dehaene, S. (2011). Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences*, *108*(51), 20754–20759. https://doi.org/10.1073/pnas.1117807108

Waters, G. S., & Caplan, D. (1996). The Measurement of Verbal Working Memory Capacity and Its Relation to Reading Comprehension. *The Quarterly Journal*

of Experimental Psychology Section A, *49*(1), 51–79. https://doi.org/10.1080/713755607

Wechsler, D. (2012). *Wechsler Adult Intelligence Scale—Fourth Edition* [Dataset]. https://doi.org/10.1037/t15169-000

Wehbe, L., Blank, I. A., Shain, C., Futrell, R., Levy, R., von der Malsburg, T., Smith, N., Gibson, E., & Fedorenko, E. (2021). Incremental Language Comprehension Difficulty Predicts Activity in the Language Network but Not the Multiple Demand Network. *Cerebral Cortex*, *31*(9), 4006–4023. https://doi.org/10.1093/cercor/bhab065

White, S. J., Drieghe, D., Liversedge, S. P., & Staub, A. (2018). The word frequency effect during sentence reading: A linear or nonlinear effect of log frequency? *Quarterly Journal of Experimental Psychology*, *71*(1), 46–55. https://doi.org/10.1080/17470218.2016.1240813

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

Yang, X., Zhang, X., Zhang, Y., Zhang, Q., & Li, X. (2020). How working memory capacity modulates the time course of semantic integration at sentence and discourse level. *Neuropsychologia*, *140*, 107383. https://doi.org/10.1016/j.neuropsychologia.2020.107383