# Is Prediction-by-Production Winner-Takes-All? Evidence from Reading Times and Brain Potentials

**Miriam Schulz (mschulz@lst.uni-saarland.de), Masato Nakamura, Francesca Delogu & Matthew W. Crocker**
Department of Language Science and Technology, Saarland University

**Abstract**

Comprehenders routinely generate predictions during language comprehension in real time, yet the mechanisms underlying prediction generation remain debated. The prediction-by-production hypothesis proposes that comprehenders use their own production system to anticipate upcoming linguistic input. While an initial body of work supports production-based accounts of prediction, it remains unclear whether increased engagement of the production system benefits all predictable words – i.e., both high- and low-cloze continuations – or only the most expected ones in a "winner-takes-all" fashion. In the present work, we investigate this question through a German self-paced reading and an event-related potentials (ERP) experiment employing a novel within-participants task paradigm manipulating engagement of the production system during comprehension. Reading times and ERPs show that increased engagement of the production system facilitiates processing for both high- and low-cloze words, suggesting that prediction-by-production supports parallel pre-activation of multiple plausible continuations, with ERP results showing strongest facilitation for highly predictable words.

**Keywords:** comprehension; prediction-by-production; self-paced reading; event-related potentials

## Introduction

While there is broad consensus regarding the central importance of prediction in language comprehension, the mechanisms underlying the generation of predictions remain less well understood. One recent proposal – the *prediction-by-production* hypothesis – argues that comprehenders use their own production system to internally simulate how an utterance might continue, based on their partial understanding of the speaker's intended message (Pickering & Gambi, 2018; Pickering & Garrod, 2013). Empirical support for this account comes from studies showing that increased engagement of the comprehender's own production system results in the generation of stronger predictions during comprehension (Hintz et al., 2016; Ito et al., 2017; Lelonkiewicz et al., 2021) and to enhanced memory for predicted information (Rommers et al., 2020), while suppressing the availability of the production system impairs prediction generation (Martin et al., 2018).

A number of challenges and open questions remain, however, concerning the strength, task-dependence, and replicability of prediction-by-production effects, as well as the compatibility of the predictions made by production-based accounts with recent empirical findings from the prediction error literature (see e.g., Brothers et al., 2023). In this work, we present behavioural and electrophysiological evidence from two experiments using a novel task paradigm to shed light on these open questions: in a self-paced reading study (SPR, Experiment 1) and an event-related potentials study (ERP, Experiment 2), participants read target sentences while engaged either in a standard comprehension-based task or in a task which strengthened the engagement of their production system during reading.

In previous studies that varied tasks to modulate the engagement of the production system, such manipulations were often implemented between participants (Martin et al., 2018) or even between experiments (Hintz et al., 2016), and in some cases effects were not stable across individual experiments and tasks (cf. e.g. Lelonkiewicz et al., 2021). More recently, critics have argued that a "winner-takes-all" prediction-by-production mechanism should entail not only facilitation for predicted continuations but also *inhibition* of competing alternatives, assuming competitive dynamics in language production (Brothers et al., 2023; Staub, 2025). Consider the following sentence: *The day was breezy so the boy went outside to fly his kite*. If comprehenders use their own production system to pre-activate the word *kite*, this should lead to inhibition of other plausible continuations, such as *airplane* (see also Ness & Meltzer-Asscher, 2021). Consequently, production-based accounts would predict increased processing costs for any continuation other than the specifically predicted word. However, there is growing evidence against such prediction error effects, even in highly predictive contexts (see e.g., Brothers et al., 2023; Frisson et al., 2017; Wong et al., 2024), a pattern that has been taken to challenge the prediction-by-production hypothesis. An alternative possibility, however, is that prediction-by-production does not commit to a single *winner-takes-all* candidate but instead promotes multiple candidates with varying degrees of predictability in parallel.

## Research aims and hypotheses

In the present work, we address these open questions with two experiments that employ a novel task paradigm: we contrast a standard comprehension task *(Reading-for-Comprehension)* with a task encouraging participants to engage in covert production *(Reading-for-Production)* during comprehension of target nouns manipulated for expectancy. Our goal was to examine whether we find evidence for prediction-by-production effects using this within-participants task manipulation in both a self-paced reading (SPR) and an event-related potentials



In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* 

**Context sentence** (displayed in full): Bobby will einen Zettel am Kühlschrank befestigen.
*Bobby wants to attach a note to the refrigerator.*

**Target sentence** (word by word):

| Condition | Start | Pre-critical | Target | Spill 1 | Spill 2 | Production/End | Cloze | Plausibility |
|---|---|---|---|---|---|---|---|---|
| HC: High cloze | Er nimmt einen<br>*He takes a* | kleinen<br>*small* | **Magneten**<br>***magnet*** | aus<br>*out of* | der<br>*the* | \| Abstellkammer.<br>\| *storage room.* | 0.78 (0.12) | 6.76 (0.27) |
| LC: Low cloze | Er nimmt einen<br>*He takes a* | kleinen<br>*small* | **Hocker**<br>***stool*** | aus<br>*out of* | der<br>*the* | \| Abstellkammer.<br>\| *storage room.* | 0.06 (0.02) | 5.14 (1.15) |
| IM: Implausible | Er nimmt einen<br>*He takes a* | kleinen<br>*small* | **Drucker**<br>***printer*** | aus<br>*out of* | der<br>*the* | \| Abstellkammer.<br>\| *storage room.* | 0.00 (0.00) | 1.56 (0.61) |

Figure 1: Example item (materials partially adapted from Nicenboim et al. (2020) and Stone et al. (2023)) together with the mean target cloze probability (Nicenboim et al., 2020) and mean plausibility rating (Likert scale: 1–7) for each condition.

(ERP) experiment. Importantly, while previous work suggests that prediction-by-production facilitates processing for highly predictable with respect to less predictable words, we aimed to directly investigate whether increased engagement of the production system further leads to facilitative or inhibitory effects for less predictable but still plausible words – such as *airplane* in the example above – by comparing them with implausible targets that were fully unpredictable given the preceding context (*The day was breezy so the boy went outside to fly his* ***guitar***). This results in a 3*x*2 design combining three levels of expectancy with two tasks manipulations, as illustrated in Figure 1.

Expectation-based accounts of language processing propose that reading times (RTs) reflect processing difficulty as a function of contextual (un)expectedness (Levy, 2008; Smith & Levy, 2013), while in the electrophysiological domain, contextual expectancy has been shown to be indexed by the N400 and late positivities (Kutas & Federmeier, 2011; Van Petten & Luka, 2012). If the strength of prediction can be influenced by a task modulating the engagement of the production system during comprehension, we expect that our task manipulation leads to task-specific differences in these measures, as outlined below. Importantly, the prediction-by-production hypothesis assumes that comprehenders generate predictions by covertly engaging their production system, which selects upcoming words through competitive lexical selection. However, it is still unclear to which extent prediction-by-production allows for parallel activation of competing lexical candidates.

***High cloze versus low cloze words*** We predict that stronger engagement of the production system in the Reading-for-Production task will amplify facilitation for highly predictable targets. Specifically, we expect faster RTs and reduced N400 amplitudes for high cloze (HC) relative to low cloze (LC) words in the Reading-for-Production task compared with the Reading-for-Comprehension task. Additionally, given that late positivities have been shown to be elicited by unpredicted but plausible words (DeLong et al., 2014; Van Petten & Luka, 2012), a "winner-takes-all" account would predict LC vs. HC continuations to elicit more enhanced late positivities through prediction-by-production.

***Low cloze versus implausible words*** Furthermore, if prediction-by-production leads comprehenders to commit more strongly to a single continuation through competitive lexical selection ("winner-takes-all"), increased production engagement should strengthen expectations for the selected HC word while simultaneously lowering expectations for alternative plausible continuations. Under this account, low cloze (LC) words should therefore show reduced facilitation – in both RTs and the N400 – relative to implausible (IM) targets, together with an enhanced late positivity reflecting prediction error.

By contrast, if prediction-by-production supports the parallel prediction of multiple candidates, LC words should benefit from increased production engagement compared to implausible words. This graded account predicts faster RTs and reduced N400 amplitudes for LC vs. IM targets in the Reading-for-Production task.

## Methods and Materials

### Experimental materials

We adapted 120 German items selected from Nicenboim et al. (2020) and Stone et al. (2023). All items consisted of a context and a target sentence that were highly constraining for a specific target noun (the high cloze condition – HC). From the cloze data collected by Nicenboim et al. (2020), we then selected an alternative plausible target noun that was less expected but still attested in the cloze responses (the low cloze condition – LC). Finally, we added an additional zero cloze implausible target noun for each item (the implausible condition – IM). Plausibility of our final target nouns was confirmed through a norming test with 45 native German participants recruited via Prolific (Palan & Schitter, 2018). An example experimental item including mean cloze probability and plausibility rating per condition is shown in Figure 1.

In addition to the 120 experimental items, we constructed 60 filler items with short target sentences (three to seven words). These were included to prevent participants from anticipating the position of the production prompt in the cloze-style production task.

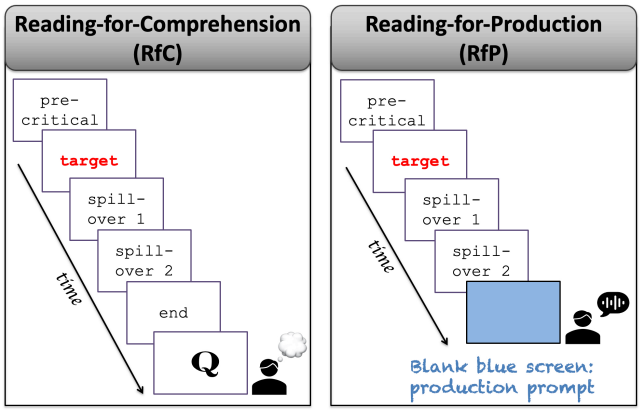


Figure 2: Task regimes. Left: Reading-for-Comprehension (RfC): each trial was followed by a comprehension question. Right: Reading-for-Production (RfP): the last one to two words of the target sentence were not shown, and participants had to quickly complete each sentence aloud with one to two words in a cloze-style sentence completion task *after* the target word.

### Task design

Participants in both experiments completed two tasks in a blocked design, with task order counterbalanced across participants.

**Reading-for-Comprehension** In a standard reading task (*Reading-for-Comprehension*, hereafter: RfC), participants first read the full context sentence. They then read the target sentence word-by-word and answered a comprehension question on every trial (see Figure 2, left).

**Reading-for-Production** In the *Reading-for-Production* task (hereafter: RfP), participants also first read the entire context sentence, and then the target sentence word-by-word, including the target noun and at least two spillover words following the target; however, the final one or two words of the target sentence – crucially, *after* the target and spillover regions had been read – were not shown. Instead, participants saw a blank blue screen (the "production prompt") and were instructed to finish each target sentence with one to two words spoken aloud as soon as the blue screen appeared (Figure 2, right).

## Experiment 1: Self-paced reading

### Participants

We collected data from 130 participants on Prolific; ten were excluded due to predefined exclusion criteria, resulting in a final sample of 120 native German participants (mean age 29 years, 75 female), paid £11 for their participation.

### Procedure

The experiment was implemented in PC Ibex (Zehr, 2018). Latin square lists were constructed such that each list only contained each item context and each target noun once, and task order was balanced across participants.

After task-specific instructions and practice trials, participants completed two 45-trial blocks in their respective first task. Following a short break, they received instructions and practice trials for their second task and completed two further 45-trial blocks.

In both tasks, each experimental trial began with the context sentence displayed in full in the middle of the screen. After participants had finished reading and pressed a key to continue, they proceeded through the target sentence word by word by pressing the space bar, each word displayed in the middle of the screen, preceded with a # to indicate where the words would appear.

In RfC, participants read the complete target sentence and then a comprehension question was presented on the screen. Participants had ten seconds to press a key to provide their answer (*yes* or *no*). In RfP, by contrast, the target sentences were only partially presented. The target word and at least two following spillover words were shown, whereas the final one to two words were removed and replaced by a blank blue screen that served as the production prompt (see Figure 2, right). Once the blue screen appeared, participants had a maximum of four seconds to finish each sentence with one to two words spoken aloud and to press a button when they had finished speaking. Their answers were recorded.

### Analysis and Results

**Behavioural results** Mean accuracy in the comprehension questions during RfC was 91% (HC: 92%, LC: 92% IM: 90%). In the RfP task, mean cloze probability among participants' answers was 0.50 (HC = LC = IM: 0.50). High accuracy on comprehension questions and high cloze probabilities across all conditions indicate that participants were able to perform both tasks successfully.

**Reading times** *Analysis.* We analyzed the RT results by fitting separate linear mixed-effects models as implemented in the *lme4* R package (Baayen et al., 2008; Bates, Mächler, et al., 2015) for each critical region (target, first and second spillover word, since effects in SPR often emerge in spillover regions; Mitchell, 1984). RTs were log-transformed prior to analyis. All models included Task (RfC vs. RfP), Condition (three-level factor), and their interaction as critical predictors, along with pre-critical reading time, task order, experiment half, trial position, word position, word length, SUBTLEX frequency (Brysbaert et al., 2011), as well as relevant interactions (including Task-by-trial position and Task-by-word position). Condition was effects coded to allow comparisons between HC vs. LC and LC vs. IM, and binary factors were effects coded as −0.5 vs. 0.5 (for Task: RfC = −0.5 and RfP = 0.5); continuous predictors were $z$-scored. After random effects selection (Bates, Kliegl, et al., 2015), models included random intercepts for subjects and items, item-level random slopes for Condition, and a by-subject random slope for Task in the first spillover region. Prior to analysis, outliers were

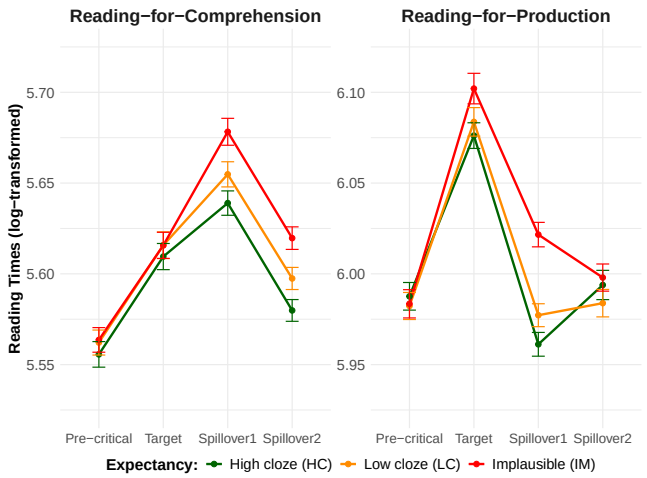


Figure 3: Log-transformed RTs for each Condition and Task (error bars indicate ±1 SE). Left: Reading-for-Comprehension (RfC); Right: Reading-for-Production (RfP).

removed in two steps by first removing extreme outliers using a hard threshold (RTs < 80 ms or > 5000 ms) and then removing any RTs further than 2.5 standard deviations from the mean RT on a by-subject by-task basis, resulting in the elimination of a total of 2.52% of critical trials.

*Results.* Log-transformed reading times are shown in Figure 3, and model results are reported in Table 1.

Across all critical regions, we observe a main effect of Task, with substantially longer RTs in RfP. In addition, we found main effects of Condition reflecting target expectancy: IM was read slower than LC in all critical regions, but while LC was read numerically slower than HC across all regions, this effect was only significant in the first spillover region.

Crucially, regarding the interaction terms, we find a significant Task x Condition interaction for LC vs. IM revealing even longer RTs for IM in RfP than in RfC in both the target and the first spillover region. However, we find no significant interaction of Task for HC vs. LC in these regions (though numerically, LC targets were read slower than HC targets in RfP). In the second spillover region, by contrast, we observe longer RTs in RfC instead of RfP for LC vs. HC.

### Discussion

The SPR data revealed an interaction of plausibility with task, with implausible continuations eliciting longer RTs than low cloze continuations in RfP in the target and first spillover region. This finding suggests that no additional processing costs are incurred for LC while engagement of comprehenders' production system was increased; instead, RTs for the LC target words were reduced in RfP compared to the IM continuations, in line with graded, parallel prediction through production instead of a "winner-takes-all" processing strategy. This is further supported by the absence of an additional facilitation for HC vs. LC continuations when participants were engaged in RfP.

It is worth noting, however, that the main effect for our HC vs. LC expectancy manipulation was rather small and significant only in the first spillover region, suggesting that the absence of an Expectancy x Task interaction might be due to a generally reduced main effect for LC vs. HC in our SPR data. This may in turn be due to a limited sensitivity to the HC-LC expectancy effect in this SPR study, despite robust plausibility effects. Importantly, RTs were overall substantially longer in RfP, likely because participants used the extra time to prepare for production. This may have shifted the time course of effects across tasks, potentially also explaining the reversed pattern between the target/first spillover and the second spillover region.

Table 1: Fixed-effect estimates from the SPR linear mixed-effects models for the target, first spillover and second spillover regions. Reported are main effects of Task and Condition and their interaction; additional predictors included in the models are not shown for reasons of space.

| Region | Predictor | Estimate | t-value | p-value |
|---|---|---|---|---|
| Target | LC – HC | 0.0033 | 0.594 | .553 |
| Target | IM – LC | 0.0139 | 2.434 | **.016** |
| Target | Task | 0.3101 | 66.310 | **<.001** |
| Target | Task x LC – HC | 0.0087 | 0.931 | .352 |
| Target | Task x IM – LC | 0.0195 | 2.082 | **.037** |
| Spill 1 | LC – HC | 0.0166 | 3.040 | **.003** |
| Spill 1 | IM – LC | 0.0351 | 5.858 | **<.001** |
| Spill 1 | Task | 0.2482 | 15.923 | **<.001** |
| Spill 1 | Task x LC – HC | 0.0014 | 0.171 | .864 |
| Spill 1 | Task x IM – LC | 0.0182 | 2.167 | **.030** |
| Spill 2 | LC – HC | 0.0029 | 0.599 | .550 |
| Spill 2 | IM – LC | 0.0178 | 3.389 | **<.001** |
| Spill 2 | Task | 0.2788 | 58.941 | **<.001** |
| Spill 2 | Task x LC – HC | −0.0029 | −2.818 | **.005** |
| Spill 2 | Task x IM – LC | −0.0100 | −1.052 | .293 |

## Experiment 2: Event-related potentials

### Participants

Forty native German right-handed participants were recruited for our EEG experiment, of which four had to be excluded due to predefined exclusion criteria. The final 36 participants (mean age 25 years, 30 female) all had normal or corrected to normal vision and did not report any reading or neurological disorders. Participants were paid 25€ for their participation.

### Procedure

We recorded participants' electroencephalogram (EEG) while they were seated in a sound-proof, electromagnetically shielded chamber. The experimental procedure was identical to that of the SPR experiment, except that EEG was recorded instead of RTs and target sentence words were pre-

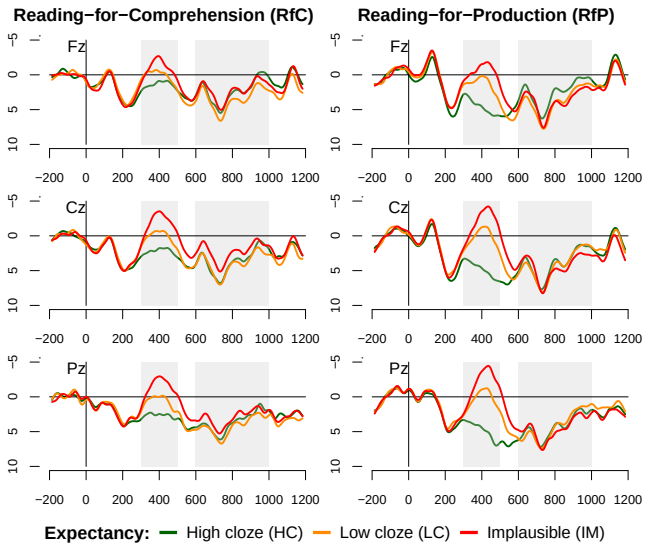


Figure 4: Grand-average ERP waveforms for each Condition and Task. Left: Reading-for-Comprehension (RfC); Right: Reading-for-Production (RfP).

sented in rapid serial visual presentation (RSVP) using E-prime 3 (Schneider et al., 2002). The sentence initial fixation marker # was displayed for 750 ms and each word was shown for 350 ms plus a 150 ms blank inter-stimulus interval.

EEG was recorded from 26 active Ag/AgCl scalp electrodes, using the standard 10 – 20 system, with FCz as online reference and AFz as ground. The data were digitized at a sampling rate of 1000 Hz. Eye movements were monitored using four electrodes placed around the eyes. Recordings were re-referenced offline to the average of the mastoids and band-pass filtered between 0.01 and 30 Hz. Horizontal eye movements were corrected using ICA in BrainVision Analyzer 2. Epochs were extracted from −200 ms to 1200 ms relative to target onset. Baseline correction was performed with respect to pre-stimulus activity ranging from −200 to 0 ms. Remaining artefacts were excluded using a semi-automatic procedure, leading to the exclusion of 2.25% of experimental trials.

## Analysis and Results

**Behavioural results** As in our SPR experiment, participants were able to successfully complete both tasks in all conditions, with an average comprehension question accuracy of 92% (HC: 93%, LC: 92% IM: 91%) in RfC, and a mean cloze probability between participants' answers of 0.53 (HC: 0.52, LC: 0.55, IM: 0.52) in RfP.

**ERPs** *Analysis.* Separate linear mixed effects models were fit to trial-level mean amplitudes for two regions of interest: the N400 time window (300 – 500 ms post stimulus onset) and a late post-N400 positivity (PNP) time window (600 – 1000 ms post stimulus onset). The same predictors and interactions were used as in the SPR analysis except for pre-critical RT, and instead, two three-level effects coded predictors were added for Region and Hemisphere as well as their interactions with each other and with Task and Condition. Models included random intercepts for subjects and items and, following random-effects selection (Bates, Kliegl, et al., 2015), by-subject random slopes for Condition in the N400 model and for Condition and Task in the post-N400 model.

Table 2: Fixed-effect estimates from the ERP linear mixed-effects models for the N400 and post-N400 positivity (PNP) time windows. Reported are main effects of Task and Condition and their interaction; additional predictors included in the models (e.g., Region, Hemisphere, and covariates) are not shown for reasons of space.

| Time window | Predictor | Estimate | t-value | p-value |
|---|---|---|---|---|
| N400 | LC – HC | −2.6638 | −7.912 | **<.001** |
| N400 | IM – LC | −1.9178 | −6.246 | **<.001** |
| N400 | Task | 0.9151 | 10.740 | **<.001** |
| N400 | Task x LC – HC | −2.0544 | −9.841 | **<.001** |
| N400 | Task x IM – LC | −0.5691 | −2.725 | **.006** |
| PNP | LC – HC | 0.7972 | 2.525 | **.017** |
| PNP | IM – LC | −0.6089 | −1.788 | .083 |
| PNP | Task | 0.8783 | 2.595 | **.014** |
| PNP | Task x LC – HC | 0.0148 | 0.063 | .950 |
| PNP | Task x IM – LC | 1.1725 | 5.010 | **<.001** |

*Results.* Figure 4 shows the ERP waveworms on the midline electrodes and Table 2 shows the model results for the predictors of interest.

In the N400 time window, we observe main effects of target expectancy with more negative amplitudes for both LC vs. HC and IM vs. LC. This effect further interacts with Task, such that the facilitation for both HC vs. LC and LC vs. IM was greater in RfP. We additionally observe a main effect of Task with more positive amplitudes overall in RfP.

In the post-N400 time window, we observe a late post-N400 positivity for LC vs. HC, which did not however interact with our task manipulation. For LC vs. IM, we found no main effect, but in interaction with task, we observe more positive amplitudes for IM vs. LC in RfP. As in the N400 time window, we again observe more positive amplitudes overall in RfP.

## Discussion

The ERP data show a facilitation for high cloze with respect to low cloze continuations, as indexed by reduced N400 amplitudes for HC vs. LC. Crucially, in contrast to the SPR data, this effect is further amplified in RfP.

In addition, our results show that modulating engagement of the production system further led to greater attenuation of N400 amplitude for low cloze with respect to implausible continuations in RfP. This finding is in line with our SPR results and provides further support for the hypothesis that

production facilitates prediction across the board, and not just for the most predictable words in highly constraining contexts.

In the post-N400 time window, a late positivity was observed for LC vs. HC, but was not modulated by task. That is, while the late positivity of LC compared to HC is consistent with prediction error or general message level updating effects in both task regimes (Van Petten & Luka, 2012), we find no evidence for increased prediction error costs in RfP (see also Lai et al., 2024). By contrast, we observe a late positivity for IM vs. LC in RfP only. We speculate that the absence of a P600 in RfC may reflect component overlap in this task (Delogu et al., 2019). We also note that overall, ERPs were more positive in RfP than in RfC.

In sum, our findings from the N400 and the post-N400 time window do not support a strict "winner-takes-all" processing strategy in Reading-for-Production, although highly predictable words reveal the strongest processing advantage.

## General Discussion and Conclusions

Using a novel within-participants task manipulation, we examined how engaging the production system during comprehension shapes predictive processing, and whether prediction-by-production entails a "winner-takes-all" mechanism or allows for graded, parallel prediction of multiple continuations.

In the SPR experiment, the facilitation for low cloze (LC) over implausible (IM) continuations was greater in the Reading-for-Production task in the target and first spillover region, indicating that increased production engagement does not incur additional processing costs for plausible but less predictable words. In addition, we observe no additional facilitation for HC vs. LC through prediction-by-production. Together, this pattern might be taken as evidence against a "winner-takes-all" account and instead support graded, parallel facilitation of multiple continuations in prediction-by-production. However, we argue that one should interpret these results with caution, due to the small main effect observed for HC vs. LC, the overall longer reading times in Reading-for-Production, and the possibility that the latter introduced an offset in the time course of effects between regions of interests in the two tasks.

By contrast, the ERP results provided clearer evidence for expectancy-driven facilitation and its modulation by task. The well-established attenuation of N400 amplitude for HC relative to LC continuations (Kutas & Federmeier, 2011; Van Petten & Luka, 2012) was amplified in Reading-for-Production, consistent with enhanced predictive processing under increased production engagement, in line with prior findings (Hintz et al., 2016; Ito et al., 2017; Lelonkiewicz et al., 2021; Martin et al., 2018). Crucially, LC words also elicited reduced N400 amplitudes compared to IM continuations in the Reading-for-Production task, providing electrophysiological evidence in line with our behavioural results that production-based prediction does not inhibit the processing of alternative plausible continuations.

In the post-N400 time window, our results additionally show a late positivity for LC vs. HC in both tasks (Van Petten & Luka, 2012), but no modulation of this effect by task, further supporting that prediction-by-production does not entail enhanced prediction error or reanalysis costs for less predictable but plausible words. By contrast, the expected P600 for implausible compared to low cloze but plausible continuations (Kuperberg et al., 2020) was observed only in Reading-for-Production; however, we interpret this interaction with caution, due to the possibility that component overlap may have masked the P600 effect in Reading-for-Comprehension (Delogu et al., 2019).

Taken together, these findings stand in contrast with recent critiques of the prediction-by-production hypothesis (Brothers et al., 2023; Staub, 2025). While Pickering and Gambi (2018) remain agnostic with respect to lexical inhibition through prediction-by-production, these critiques assume that engagement of the production mechanism inevitably causes inhibition and thus predicts enhanced expectation for a single highly expected continuation and inhibition for any other candidates ("winner-takes-all"). This would in turn conflict with growing evidence that prediction in comprehension is probabilistic and graded, even in highly constraining contexts (Frisson et al., 2017; Wong et al., 2024). Our findings, however, reveal that production-based prediction can be compatible with such graded expectations. Although highly predictable words showed the strongest facilitation in the N400 time window, we found no evidence for increased processing difficulty for low cloze but plausible continuations in Reading-for-Production: ERP results showed reduced N400 amplitudes relative to implausible continuations and no enhanced late positivity relative to high cloze continuations, and behavioural results showed faster reading times for low cloze compared to implausible continuations, but no additional advantage for high cloze over low cloze continuations. Our results thus show that engaging the production system benefits multiple plausible continuations, possibly with a greater benefit for highly expected words. This suggests an account in which the engagement of the production system benefits multiple expected continuations, without inhibiting those that are less likely.

To conclude, the present results contribute to ongoing debates about the mechanisms underlying predictive language processing by demonstrating that task demands influencing the engagement of the production system can modulate prediction, without necessarily narrowing it to a single outcome by activating competitive inhibition. By introducing a task manipulation that encourages covert production during comprehension, the current study offers a new methodological avenue to investigate the mechanisms underlying the generation of predictions in online language processing.

## Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — Project-ID232722074 — SFB/CRC 1102.

## References


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

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*. https://doi.org/10.48550/arXiv.1506.04967

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

Brothers, T., Morgan, E., Yacovone, A., & Kuperberg, G. (2023). Multiple predictions during language comprehension: Friends, foes, or indifferent companions? *Cognition*, *241*, 105602. https://doi.org/10.1016/j.cognition.2023.105602

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect. *Experimental psychology*. https://doi.org/10.1027/1618-3169/a000123

Delogu, F., Brouwer, H., & Crocker, M. W. (2019). Event-related potentials index lexical retrieval (n400) and integration (p600) during language comprehension. *Brain and cognition*, *135*, 103569. https://doi.org/10.1016/j.bandc.2019.05.007

DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late erp positivities during written sentence comprehension. *Neuropsychologia*, *61*, 150–162. https://doi.org/10.1016/j.neuropsychologia.2014.06.016

Frisson, S., Harvey, D. R., & Staub, A. (2017). No prediction error cost in reading: Evidence from eye movements. *Journal of Memory and Language*, *95*, 200–214. https://doi.org//10.1016/j.jml.2017.04.007

Hintz, F., Meyer, A. S., & Huettig, F. (2016). Encouraging prediction during production facilitates subsequent comprehension: Evidence from interleaved object naming in sentence context and sentence reading. https://doi.org/10.1080/17470218.2015.1131309

Ito, A., Dunn, M. S., & Pickering, M. (2017). Effects of language production on prediction: Word vs. picture visual world study. *Mental Architecture for Processing and Learning of Language (MAPLL)*.

Kuperberg, G. R., Brothers, T., & Wlotko, E. W. (2020). A tale of two positivities and the n400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of cognitive neuroscience*, *32*(1), 12–35. https://doi.org/10.1162/jocn_a_01465

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, *62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Lai, M. K., Payne, B. R., & Federmeier, K. D. (2024). Graded and ungraded expectation patterns: Prediction dynamics during active comprehension. *Psychophysiology*, *61*(1), e14424. https://doi.org/10.1111/psyp.14424

Lelonkiewicz, J. R., Rabagliati, H., & Pickering, M. J. (2021). The role of language production in making predictions during comprehension. *Quarterly Journal of Experimental Psychology*, *74*(12), 2193–2209. https://doi.org/10.1177/17470218211028438

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

Martin, C. D., Branzi, F. M., & Bar, M. (2018). Prediction is production: The missing link between language production and comprehension. *Scientific reports*, *8*(1), 1079. https://doi.org/10.1038/s41598-018-19499-4

Mitchell, D. C. (1984). An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading. In D. E. Kieras & M. A. Just (Eds.), *New methods in reading comprehension research*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9780429505379

Ness, T., & Meltzer-Asscher, A. (2021). Love thy neighbor: Facilitation and inhibition in the competition between parallel predictions. *Cognition*, *207*, 104509. https://doi.org/10.1016/j.cognition.2020.104509

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

Palan, S., & Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of behavioral and experimental finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

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

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and brain sciences*, *36*(4), 329–347. https://doi.org/10.1017/S0140525X12001495

Rommers, J., Dell, G. S., & Benjamin, A. S. (2020). Word predictability blurs the lines between production and comprehension: Evidence from the production effect in memory. *Cognition*, *198*, 104206. https://doi.org/10.1016/j.cognition.2020.104206

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-prime user's guide. pittsburgh: Psychology software tools.

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

Staub, A. (2025). Predictability in language comprehension: Prospects and problems for surprisal. *Annual Review of Linguistics*, *11*(1), 17–34. https://doi.org/10.1146/annurev-linguistics-011724-121517

Stone, K., Nicenboim, B., Vasishth, S., & Rösler, F. (2023). Understanding the effects of constraint and predictability in erp. *Neurobiology of Language*, *4*(2), 221–256. https://doi.org/10.1162/nol_a_00094

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and erp components. *International journal of psychophysiology*, *83*(2), 176–190. https://doi.org/10.1016/j.ijpsycho.2011.09.015

Wong, R., Veldre, A., & Andrews, S. (2024). Are there independent effects of constraint and predictability on eye movements during reading? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *50*(2), 331. https://doi.org/10.1037/xlm0001206

Zehr, J. (2018). Penncontroller for internet based experiments (ibex). https://doi.org/https://doi.org/10.17605/OSF.IO/MD832