Rommers, J., Dell, G. S., Benjamin, A. S. (in press). Word predictability blurs the lines between production and comprehension: Evidence from the production effect in memory. *Cognition.*

Word predictability blurs the lines between production and comprehension: Evidence from the production effect in memory

Joost Rommers[1], Gary S. Dell[2,3], Aaron S. Benjamin[2,3]

1. School of Psychology, University of Aberdeen, UK
2. Department of Psychology, University of Illinois, Urbana-Champaign, USA
3. Beckman Institute for Advanced Science and Technology, University of Illinois, Urbana-Champaign, USA

Running head: The production effect for predictable words

Please address correspondence to:

Joost Rommers
School of Psychology
William Guild Building
University of Aberdeen
Aberdeen, AB24 3FX
United Kingdom
Email: joost.rommers@abdn.ac.uk

Abstract

Predictions about likely upcoming input may promote rapid language processing, but the mechanisms by which such predictions are generated remain unclear. One hypothesis is that comprehenders use their production system to covertly produce what they would say if they were the speaker. If reading predictable words involves covert production, this act might have consequences for memory. The present study capitalized on the production effect, which is the observation that words read aloud are remembered better than words read silently. Participants read sentence-final predictable and unpredictable words aloud or silently, followed by a surprise recognition memory task. If reading predictable words involves covert production, the memory improvement from actually producing the words should be smaller for predictable words than for unpredictable words. This was confirmed in Experiment 1, which tested item memory using old/new judgments. Experiment 2 followed the same procedure, except that participants now made aloud/silent judgments probing their memory for prior acts of production. Here the hypothesis was that, relative to unpredictable words, it should be more difficult to remember whether predictable words had been read aloud or silently. Indeed, word predictability tended to make it harder to tell the difference, suggesting that predictability blurred the lines between production and comprehension. Taken together, the findings support the idea that reading predictable words can involve covert production and show that this act has consequences for what readers retain.

Keywords: *sentence comprehension; word production; recognition memory; production effect*

Introduction

There is an emerging consensus that language comprehension is guided not just by signal-driven processes, but also by predictions about likely upcoming input (Altmann & Mirković, 2009; Christiansen & Chater, 2016; Dell & Chang, 2013; Federmeier, 2007; Kamide, 2008; Kutas, DeLong, & Smith, 2011; Pickering & Garrod, 2013; see also Elman, 1990; Marslen-Wilson, 1973). Predictions are thought to facilitate stimulus processing (van Berkum, 2010) and promote learning (Chang, Dell, & Bock, 2006; Friston, 2005; Rao & Ballard, 1999). However, the mechanisms by which predictions are generated in the first place remain unclear.

One hypothesis is that comprehenders use their production system to covertly produce what they would say if they were the speaker (Chang et al., 2006; Federmeier, 2007; Pickering & Garrod, 2007; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005). Evidence for this idea thus far is primarily correlational in nature. For example, participants who are good at production tasks show stronger prediction-related effects as measured by event-related brain potentials (Federmeier, Kutas, & Schul, 2010; Federmeier, Mclennan, De Ochoa-Dewald, & Kutas, 2002; Rommers & Federmeier, 2018a) and eye movements (Hintz, Meyer, & Huettig, 2017; Mani & Huettig, 2012; Rommers, Meyer, & Huettig, 2015). Furthermore, prediction seems to rely on the left hemisphere of the brain, which is also the dominant hemisphere for language production (Federmeier & Kutas, 1999; Wlotko & Federmeier, 2007). Finally, prediction can influence articulation (Drake & Corley, 2015), and there is some evidence that articulatory suppression can reduce prediction-related effects seen in event-related potentials (Martin, Branzi, & Bar, 2018).

The present study took an experimental approach and investigated whether predictable words leave behind memory traces that are "production-like." The experiments reported here capitalized on the *production effect*, which is the observation that words read aloud are remembered much better than words read silently (Conway & Gathercole, 1987; Hopkins & Edwards, 1972; MacLeod, Gopie, Hourihan, Neary, & Ozubko, 2010). This effect is strongest for spoken production, but mouthing, writing, and whispering also improve memory (Forrin, MacLeod, & Ozubko, 2012; Gathercole & Conway, 1988; MacLeod et al., 2010), suggesting that the covert production that may underlie prediction could be measurable. Most studies of the production effect have employed word lists, but the effect also obtains with sentences and educational texts (Lindner, Drouin, Tanguay, Stamenova, & Davidson, 2015; Ozubko, Hourihan, & MacLeod, 2012).

Several theories have attempted to explain the origins of the production effect. According to the *distinctiveness* account (e.g., Conway & Gathercole, 1987), producing a word yields a distinctive acoustic and motoric memory that uniquely characterizes the episode of processing the word, characteristics that are not available for words read silently. In recognition tests, participants might then use a distinctiveness heuristic: if they remember having said a word aloud, it can be inferred that the word must have been studied (Dodson & Schacter, 2001). Although distinctiveness is still considered a major explanation of the production effect (e.g., MacLeod et al., 2010), it has been contrasted with a *memory strength* account in which production simply creates stronger memory traces than silent reading (e.g., Bodner & Taikh, 2012). The possibility that production improves memory simply by counteracting “lazy reading” has not been supported: production effects of equivalent magnitude have been obtained regardless of elaborative vs. non-elaborative encoding during study (Forrin, Jonker, & MacLeod, 2014; MacLeod et al., 2010). Although more studies are needed to fully clarify the mechanisms underlying the production effect, it is clear that the act of production increases what readers ultimately retain in memory.

Against this background, two experiments investigated whether producing words versus silently reading them are more similar in terms of their consequences for memory when the words are predictable than when they are unpredictable. In both experiments, participants read predictable and unpredictable sentence-final words aloud or silently, and their recognition memory for these words was assessed.

In Experiment 1, the memory test consisted of old/new judgments, and analyses focused on the production effect as an index of the difference between production and comprehension. We hypothesized that, if reading predictable words already involves covert production, then the memory improvement from actually producing the words should be smaller for predictable words than for unpredictable words.

In Experiment 2, the memory test consisted of aloud/silent judgments, thus directly assessing participants' memory for previous acts of production. Analyses focused on aloud/silent confusability as an index of the similarity of production and comprehension. Here, the hypothesis was that, relative to unpredictable words, any covert production should make it harder to remember whether the predictable words had been read aloud or silently. Thus, both experiments tested in different ways whether predictability makes production and comprehension more similar.

Experiment 1

Method

Participants. Seventy University of Illinois students (25 men, average age 19 years, range 18-25 years) took part in the experiment in exchange for course credit. Based on the G*Power software (Faul, Erdfelder, Lang, & Buchner, 2007), this sample size was expected to have over

80% power to detect a small effect.[1] The participants were native speakers of American English and had normal or corrected-to-normal vision and hearing. One additional participant was run but excluded for reading aloud on all silent trials.

Materials and Design. The stimuli consisted of 120 pairs of sentences. In each pair, one sentence was highly constraining and the sentence-final critical word was predictable (e.g., "John swept the floor with a broom."). The other sentence in a pair was weakly constraining, making the same critical word unpredictable (e.g., "He could not find the red broom."). Word predictability and sentential constraint were confirmed using a cloze probability test, in which a separate group of participants completed each sentence with the word they would generally expect to find completing the sentence fragment (for details, see Rommers & Federmeier, 2018b). The cloze probability of each critical word in its sentence, defined as the proportion of participants who completed the sentence with that word, was higher for predictable words (average ± SD: 0.87 ± 0.13, range 0.5-1) than for unpredictable words (0.01 ± 0.03, range 0-0.2). The predictable words were always the most frequently provided completion. The cloze probability of the most frequently provided completion in (that is, the constraint of) the weakly constraining sentence frames was low (0.19 ± 0.08, range 0-0.35), suggesting that the unpredictable words did not violate a strong, consistent expectation. Both strongly and weakly constraining sentences were on average 10 words long (range: 4-21 words).

Critical words were rotated through five lists such that they occurred only once on each list and, across lists, each critical word represented every condition. Each participant saw one of the lists. On every list, 24 critical words each represented the four conditions resulting from the 2

[1] The following settings yielded a required sample size of 62, which was rounded up for full counterbalancing: Statistical test = "ANOVA: Repeated measures, within factors", effect size f = 0.1, power = 0.8, number of measurements = 4, correlation among repeated measures = 0.78 (the latter was based on pilot data).

× 2 crossed manipulations of Predictability (High, Low) and Production (Aloud, Silent), and 24 words served as new words on the memory test. Seventy-two additional new words were added to each memory list to balance the number of old and new words; responses to these words were not analyzed. Lists were pseudo-randomized individually, separately for the reading and memory tasks, under the constraint that no more than three trials of the same condition occurred consecutively.

Procedure. First, in the reading phase of the experiment, participants read 96 sentences for comprehension. The sentences were presented word-by-word in the center of the screen, in Arial font size 32 on a black background. Each word was presented for 200 ms with a 300 ms inter-stimulus interval. The sentence completion was always printed in white, to keep the visual stimulus identical across conditions. Participants read the final word aloud or silently depending on the color of the words in the sentence frame preceding it, which was red or blue (counterbalanced between participants). A microphone recorded each response from -500 ms until 1500 ms relative to critical word onset. In order to encourage participants to pay attention, they were told that their comprehension would be tested later, but they were not specifically informed that there would be a memory test. The reading phase started with a practice block of four sentences. After every 12 trials, participants were reminded of the color associated with reading aloud and were given the opportunity to take a break before continuing.

After reading all of the sentences, participants performed a distraction task in which they completed as many basic math problems as they could within 30 seconds (adding up two numbers between 10 and 100). This was followed by a surprise memory test in which participants were presented with the 96 critical old words that they had read and the 96 new words (24 of which were critical words that were old words for other participants), and made old/new judgments on a four-point rating scale (“Sure New”, “Maybe New”, “Maybe Old”, “Sure Old”).

On each trial, a word was presented in the center of the screen in green Arial font size 32 on a black background, above the four response options (also in green) which corresponded to the keyboard buttons *v*, *b*, *n*, and *m*. There was no time limit for giving the response. Participants were encouraged to use the entire scale. When one of the buttons was pressed, the corresponding option on the screen lit up in white for 300 ms, followed by a 200 ms blank screen before the next trial started.

Analysis. For the memory responses, a standard measure of discriminability (d_a) was calculated from ROC curves in *z*-space based on unbiased slopes and intercepts (Stanislaw & Todorov, 1999, p. 147). Larger d_a values indicate better discriminability between old and new words. The index d_a is similar to *d'*, but does not assume equal variances of the underlying probability distributions (Swets, 1986; Wixted, 2007), and it has desirable metric properties which allow for interval interpretation (Matzen & Benjamin, 2009), thus avoiding the problem of removable interactions (Loftus, 1978). Discriminability was analyzed using Analyses of Variance (ANOVAs), supplemented with effect sizes (Cohen's d_z for within-subjects designs) and confidence intervals.

Results and Discussion

Responses on aloud trials during the sentence reading phase were considered correct if the sentence-final word was produced without disfluencies. Responses on silent trials were considered correct if no speech sound was made. Accuracy in reading aloud and silently on the corresponding trials was, on average across participants, 99% in every condition (range in individual conditions in individual participants: 83-100%) and was not further analyzed. Naming latencies from correct trials were analyzed after adding a constant of 160 ms to offset a few negative values and then log-transforming to reduce skewness. Compared with unpredictable words (615 ms, de-logged), predictable words elicited shorter naming latencies by 53 ms (95% CI [40, 67]), Cohen's $d_z = 0.90$, t (69) = 7.569, $p < .0001$. Thus, readers were sensitive to predictability.

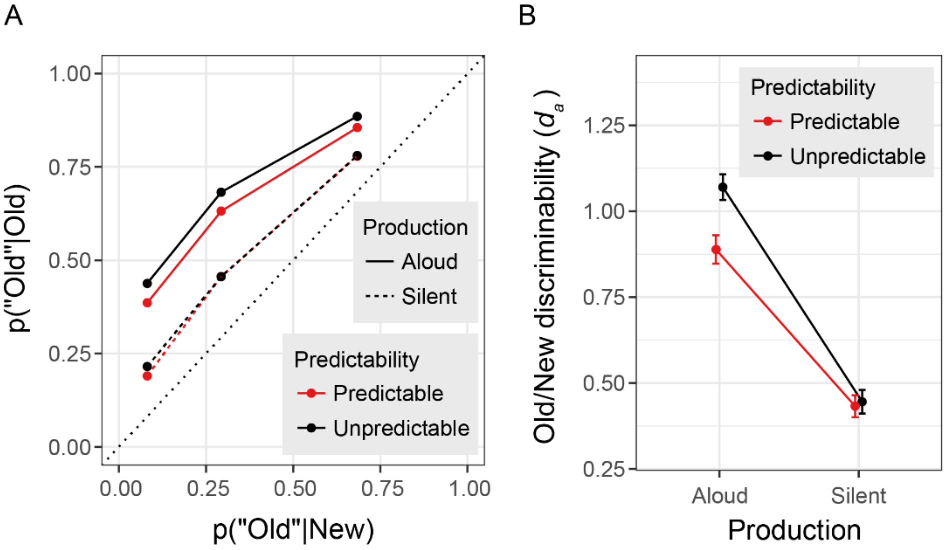

Figure 1. Grand average item memory results (Experiment 1). A) Receiver-operating characteristic (ROC) curves showing, at the different levels of confidence, cumulative hit rates for the four types of Old words as a function of the corresponding cumulative false alarm rates for New words. B) Old/new discriminability in each condition. Error bars represent unbiased within-participant SEM (Cousineau, 2005; Morey, 2008).

Our main interest was in memory performance for critical words that had elicited correct responses during the sentence reading phase. Overall, more "Sure Old" and "Maybe Old" responses were given to old words (55%) than to new words (29%), demonstrating that participants remembered the words (see Appendix for all proportions). Note that performance was not as accurate as in some other production effect experiments, which is to be expected after an incidental encoding task with a relatively large number of stimuli (including all sentence contexts, participants read up to 570 words). The full rating scale was analyzed using signal-detection theoretic measures. Figure 1A displays receiver-operating characteristic (ROC) curves showing hit rates in response to the four types of old words (cumulated across confidence levels, beginning at "Sure Old") as a function of the corresponding cumulative false alarm rates in response to the counterbalanced new words. Discriminability (d_a), related in Figure 1A to the distance from the dashed diagonal that represents chance performance (e.g., Benjamin & Diaz, 2008, p. 83), was calculated for each participant and condition. One participant did not follow the instructions and almost exclusively made "Sure Old" and "Sure New" judgments, which precluded the calculation of d_a, leaving 69 participants for the analysis. As shown in Figure 1B, discriminability was better for words read aloud than words read silently (the production effect), difference $d_a = 0.54$ (95% CI [0.46, 0.62], Cohen's $d_z = 1.68$), F (1,68) = 193.890, $p < .0001$. Unpredictable words were remembered better than predictable words, difference $d_a = 0.10$ (95% CI [0.03, 0.16], $d_z = 0.35$), F (1,68) = 8.684, $p = .004$, although this main effect was driven by the simple effect of predictability within the words read aloud. Critically, the memory improvement from reading aloud was smaller for predictable words (silent $d_a = 0.43$; aloud $d_a = 0.89$) than for unpredictable words (silent $d_a = 0.45$; aloud $d_a = 1.07$), interaction $d_a = 0.17$ (95% CI [0.02, 0.32], $d_z = 0.27$), F (1,68) = 5.025, $p = .028$. The Bayes Factor associated with the interaction was $BF_{+0} = 2.693$ (using a directional prior with a standard Cauchy scale $r = .707$). The interaction went in this predicted direction for 62% of the participants. In sum, production improved memory, but to a lesser extent for highly predictable words than for unpredictable words. This is consistent with the idea that reading predictable words involves covert production, such that adding actual production to the task had less of an effect on memory compared with unpredictable words.

The fact that predictability reduced naming latencies (as reported previously; e.g., Griffin & Bock, 1998; Stanovich & West, 1979) fits well with the idea that some of the production processes required for reading aloud were facilitated for predictable words, or had even been prepared in advance. Alternatively, one might argue that producing unpredictable words was difficult in some way, and that this difficulty improved memory (for review of related “desirable difficulties”, see Bjork, 1994). Recall that the design already avoided a particular source of difficulty, expectation violations, by embedding the unpredictable words in relatively neutral contexts which did not afford strong, consistent expectations. In order to further disentangle production difficulty and predictability, we examined whether better memory was associated with longer naming latencies, presumably reflecting difficult trials, when considering unpredictable words only (the processing of which cannot be facilitated by predictability). The naming latencies of subsequently “remembered” unpredictable words (judged “Sure Old” or “Maybe Old”; 615 ms) differed only numerically from the naming latencies of subsequently “forgotten” unpredictable words (judged “Sure New” or “Maybe New”; 625 ms), t (69) = -1.379, p = .172, and in the opposite direction as that view would predict (-11 ms, 95% CI [-26, 5], d_z = 0.16).[2] Thus, when decoupled from any facilitatory influences of predictability, there was no evidence that production difficulty per se improved subsequent memory.

It should be noted that the reduced production effect for predictable words seemed to arise from a memory disadvantage for the predictable words read aloud. One could argue that, if predictable words are covertly produced, there should have been a memory advantage for silently read predictable words. However, production is not the only factor that determines memory, and other factors work in the opposite direction: in particular, silently read predictable words are generally unsurprising and can therefore show a memory disadvantage relative to unpredictable words (e.g., Cairns, Cowart, & Jablon, 1981; Corley, MacGregor, & Donaldson, 2007; Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007; O’Brien & Myers, 1985; Perry &

[2] Differences between items could theoretically contaminate such a subsequent memory effect, but a linear mixed-effects regression model which simultaneously took into account items and participants as random effects (e.g., Baayen, Davidson, & Bates, 2008) confirmed the numerical difference in the opposite direction, β = -0.00675 (-5 ms), t = -0.5, χ^2 (1) = 0.303, p = 0.582.

Wingfield, 1994). For this reason we refrained from hypothesizing about pairwise differences between conditions, but focus instead on the interaction. This interaction took on exactly the hypothesized form.

In sum, the finding that word predictability reduced the production effect, a difference between production and silent reading, suggests that production and comprehension are more similar when prediction is made possible by the sentence context. This result is consistent with the idea that predictions are generated by the production system.

Experiment 2 provided a different assessment of production and comprehension similarity by changing the memory test to aloud/silent judgments (a source memory task), thus specifically probing participants' memories for previous acts of production. Participants were presented only with words that they had read previously and indicated for each word whether they had produced it or read it silently. Here we hypothesized that, if reading predictable words involves covert production, then for predictable words (compared with unpredictable words) it should be more difficult to remember whether they had been produced or read silently.[3] In other words, Experiment 2 tested whether predictability increases the confusability of comprehension and production.

[3] For similar logic in the investigation of top-down effects on speech perception, see Samuel (1981), and in memory for action events, see Goff and Roediger (1998).

Experiment 2

Methods

Participants. Seventy University of Illinois students (28 men, average age 19 years, range 18-28 years) who had not participated in Experiment 1 took part in Experiment 2 in exchange for course credit. The data were collected at the same time as those of Experiment 1, with half of the total number of participants being randomly assigned to each of the two experiments. The sample size was determined using the same power analysis (although a slightly larger number of items per condition should increase power relative to Experiment 1). Participants were native speakers of American English and had normal or corrected-to-normal vision and hearing. One additional participant was run but excluded for whispering on the silent trials.

Materials and Design. The stimuli were identical to Experiment 1. However, because only old words were presented in the memory test, there were only four conditions yielding four lists, and therefore each participant saw thirty items in each condition rather than 24.

Procedure. The procedure was identical to Experiment 1, except that during the memory test, participants judged whether each word had been read aloud or silently, on a four-point rating scale (“Sure Silent”, “Maybe Silent”, “Maybe Aloud”, “Sure Aloud”). In addition, halfway through the reading phase, the assignment of Production condition (Aloud, Silent) to the color of the sentence frames (red, blue) was reversed, followed by a four-sentence practice block with the reversed colors. This switch ensured that, during retrieval, recalling the color of a sentence context would not by itself be diagnostic of whether the corresponding critical word had been produced or read silently.

Results and Discussion

During the sentence reading phase, accuracy in reading aloud and silently according to the instructions was on average 98-99% in every condition (range in individual conditions in individual participants: 87-100%) and was not further analyzed. Naming latencies from correct trials were analyzed after adding a constant of 486 ms to offset a few negative values and then log-transforming to reduce skewness. Compared with unpredictable words (613 ms, de-logged), naming latencies of predictable words were shorter by 48 ms (95% CI [36, 60]), $d_z = 0.84$, t (69) = 7.015, $p < .0001$.

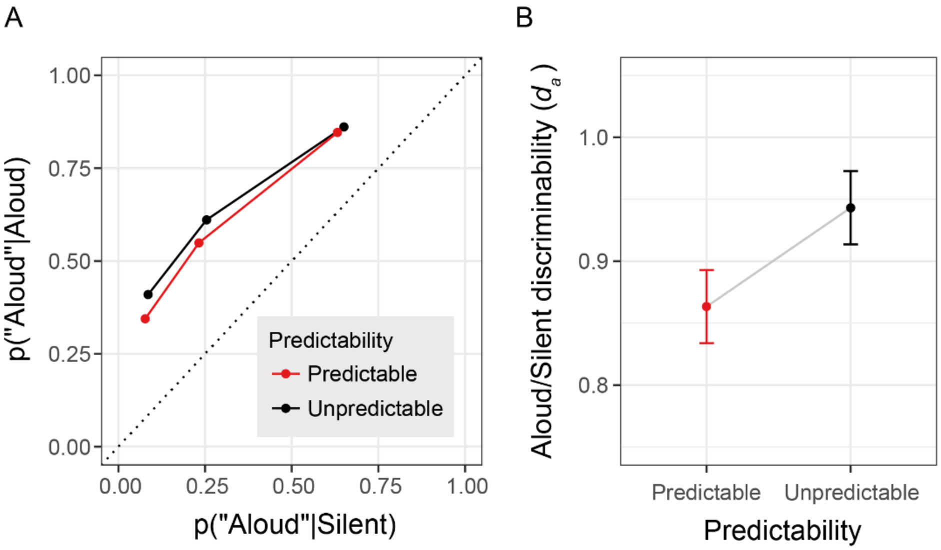

Figure 2. Grand average source memory results (Experiment 2). A) ROC curves. B) Aloud/silent discriminability.

In terms of overall memory performance for the words that had been read correctly, words read aloud received more “Sure Aloud” and “Maybe Aloud” judgments (58%) than did words read silently (24%), demonstrating accurate source memory (see Appendix for all proportions). In analyses of the full response scale, two discriminability values (d_a) were calculated for each participant: one for predictable words and one for unpredictable words, each based on separate hit and false alarm rates stemming from the aloud and silent trials, respectively. For two participants, discriminability could not be calculated because their responses were not sufficiently spread along the rating scale, leaving 68 participants for the analysis. As shown in Figure 2, compared with unpredictable words (d_a = 0.94), there was a small reduction in aloud/silent discriminability for predictable words, difference d_a = 0.08 (95% CI [-0.004, 0.16], d_z = 0.23), F (1,67) = 3.643, p = .061. The effect of interest was associated with a Bayes Factor of BF_{+0} = 1.421 (using a directional prior with a standard Cauchy scale of r = .707). The effect went in the predicted direction in 68% of the participants, an even greater proportion than in Experiment 1.[4] Although the reliability of novel results should always be examined in future work, it should be noted that this was a test of a directional hypothesis based on an existing theory, and the two experiments reported here converge on the same conclusion despite testing the theory in quite different ways.

As in Experiment 1, there was no evidence that naming difficulty improved subsequent memory independently of predictability. Within the unpredictable condition, naming latencies for subsequently remembered acts of production (judged “Sure Aloud” or “Maybe Aloud”; 608 ms) differed only numerically from those for subsequently forgotten acts of production (judged “Sure Silent” or “Maybe Silent”; 618 ms), t (69) = -1.291, p = .201, and in the opposite direction (-10 ms, 95% CI [-23, 3], d_z = -0.15).[5]

In summary, participants tended to be worse at telling the difference between production and silent reading when words were predictable than when they were unpredictable. This is consistent with the hypothesis that word predictability makes production and comprehension more similar in memory.

[4] Furthermore, ordinal regression analyses of the 4-point ratings using cumulative link mixed models (as implemented in the *ordinal* package in R; Christensen, 2019) with the maximal random effects structure confirmed the effects of interest in Experiment 1, β = -0.238, z = -2.247, $\chi2$ (1) = 4.944, p = 0.026, and Experiment 2, β = -0.199, z = -2.058, $\chi2$ (1) = 4.181, p = 0.041.

[5] As in Experiment 1, a linear mixed-effects regression model confirmed that there was only a numerical subsequent memory difference in the opposite direction, β = -0.00760 (-8 ms), t = -1.1, χ^2 (1) = 1.250, p = 0.264.

General Discussion

Two experiments used the production effect in memory to investigate the idea that reading predictable words involves covertly producing them (Chang et al., 2006; Federmeier, 2007; Pickering & Garrod, 2007; Van Berkum et al., 2005). Experiment 1 used old/new judgments and confirmed the hypothesis that the memory improvement from production should be smaller for predictable words than for unpredictable words; if predictable words are already produced covertly, then adding actual production to the task should have less of an effect. Experiment 2 used aloud/silent judgments and showed that word predictability tended to decrease participants' ability to discriminate between prior acts of production and silent reading. This suggests that word predictability can increase confusion between production and comprehension. Taken together, these results converge to support the hypothesis that word predictability blurs the lines between comprehension and production, consistent with the idea that predictions in sentence context are generated by the production system. The present experimental findings extend previous correlational support (Federmeier et al., 2002; Mani & Huettig, 2012) and converge with other recent experimental evidence suggesting reduced prediction when articulatory suppression prevented inner speech (Martin et al., 2018). The fact that the present results were obtained in memory tasks further suggests that the prediction-production link matters not only during rapid on-line processing but also for what language users ultimately retain.

As is common with manipulations of predictability, it is possible that processes other than prediction contributed to the observed effects, such as bottom-up integration of each word with its context (e.g., Kintsch, 1988). Such integration processes could influence memory because of stronger associations between sentence contexts and predictable words, compared with unpredictable words. In particular, one could argue that this led participants to rely on relatedness heuristics when recognizing predictable words, but on episodic details about production when recognizing unpredictable words. However, individuals are generally reluctant to use different strategies for different items on a recognition test, especially if the item itself does not inherently carry information about the class to which it belongs (Benjamin, 2003; Benjamin & Bawa, 2004; Rotello & Macmillan, 2007). In addition, production still considerably enhanced memory for predictable words, and the prediction explanation can be considered more parsimonious than a dual-mechanism memory account. Finally, unlike prediction, integration has not a priori been theorized to use the production system.

Similarly, although it is clear that reading aloud involves more production processes than does reading silently, factors other than the involvement of the production system might contribute to the production effect. Indeed, this would be one explanation for the fact that word predictability did not fully eliminate the production effect and source memory accuracy, but only decreased them. Related to this, production is probably not the only prediction mechanism (Mani & Huettig, 2013; Pickering & Garrod, 2013), and engagement of the production system might even be optional (Pickering & Gambi, 2018). In addition, predictions can be generated using various kinds of representations, such as event-based knowledge, independently of whether the words are covertly produced or not (Altmann & Mirković, 2009). We propose that the findings observed here reflect the memory consequences of processing overlap between the part of the production effect that reflects language production and the part of the predictability effect that reflects prediction.

On a methodological note, this study improved and extended previously used procedures for investigating the production effect in two ways. First, the produced speech was recorded and onset latencies were analyzed as a function of subsequent memory, albeit no correlation was observed in these data. Second, in Experiment 2 requiring aloud/silent judgments, the assignment of font color to aloud/silent condition was switched around halfway through the encoding phase, preventing it from being directly diagnostic of source judgments at test (although it could be indirectly diagnostic if participants managed to remember both color and item recency). As expected, source memory was still accurate, confirming the assumption that such judgments reflect memory for production beyond just memory for color (Ozubko, Gopie, & MacLeod, 2012; Ozubko, Major, & MacLeod, 2014).

As discussed previously, the possible memory mechanisms underlying the production effect include an increase in memory strength as well as the use of a distinctiveness heuristic (Bodner & Taikh, 2012; MacLeod et al., 2010; for recent review, see MacLeod & Bodner, 2017). For the purposes of the present study, both of these mechanisms are compatible with our interpretation that predictability made comprehension and production more similar in memory. Yet another way of looking at the production effect, however, is from the perspective of word production research. Word production involves many representations and processes that might drive the production effect, although memory research thus far has not investigated the effect from this perspective (for a recent exception, see Zormpa, Brehm, Hoedemaker, & Meyer, 2019). There is broad consensus that language production proceeds through the stages of conceptualization, lexical selection, phonological encoding and articulation (e.g., Dell, 1986; Levelt, Roelofs, & Meyer, 1999), but the involvement of each of these processes in the production effect on memory (and in prediction during comprehension) is an empirical issue. Regarding prediction during comprehension, which does not involve articulation, we can speculate that there might be a gradient whereby prediction during comprehension is more likely to involve production processes that are relatively far away from articulation, like conceptualization. That is, one may be more likely to predict upcoming meanings than upcoming sounds (for similar discussion, see Pickering & Gambi, 2018). But regarding the processes underlying the production effect, articulation is the mandatory process when words are read aloud, whereas the processing of meaning can sometimes be bypassed using a direct “route” from letter sequences to their corresponding speech sounds (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001). The proposal that reading aloud can bypass meaning would further imply that the production effect on memory might be dissociable from comprehension success, although this has yet to be investigated. In summary, a better understanding of the effects of language on memory is needed to pin down exactly at what levels production and comprehension may have overlapped to generate the effects observed here.

To conclude, the present study showed that, relative to unpredictable words, predictable words elicit less of a production effect and more confusion between production and comprehension in memory. Word predictability seems to blur the lines between comprehension and production, with downstream consequences that reach beyond rapid on-line processing.

Supplementary material

The data associated with this manuscript are available at https://osf.io/pj3hc/.

Acknowledgements

We are grateful to Faith Tan for data collection and speech onset measurements, and to Opal Harshe for data collection. JR was supported by the Dutch Research Council (NWO) grant 275-89-032. NWO played no role in the study design, the collection, analysis and interpretation of data, the writing of the report, or in the decision to submit the article for publication.

References

Altmann, G. T. M., & Mirković, J. (2009). Incrementality and Prediction in Human Sentence Processing. *Cognitive Science*, *33*(4), 583–609. https://doi.org/10.1111/j.1551-6709.2009.01022.x

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

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition*, *31*(2), 297–305. https://doi.org/10.3758/BF03194388

Benjamin, A. S., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language*, *51*(2), 159–172. https://doi.org/10.1016/j.jml.2004.04.001

Benjamin, A. S., & Diaz, M. (2008). Measurement of relative metamnemonic accuracy. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of Memory and Metamemory* (pp. 73–94). New York, New York: Psychology Press.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura, *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Bodner, G. E., & Taikh, A. (2012). Reassessing the basis of the production effect in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1711–1719. https://doi.org/10.1037/a0028466

Cairns, H. S., Cowart, W., & Jablon, A. D. (1981). Effects of prior context upon the integration of lexical information during sentence processing. *Journal of Verbal Learning & Verbal Behavior*, *20*(4), 445–453. https://doi.org/10.1016/S0022-5371(81)90551-X

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*(2), 234–272. https://doi.org/10.1037/0033-295X.113.2.234

Christensen, R. H. B. (2019). ordinal—Regression Models for Ordinal Data (Version 2019.4-25). Retrieved from http://www.cran.r-project.org/package=ordinal/

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*. https://doi.org/10.1017/S0140525X1500031X

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204–256. https://doi.org/10.1037/0033-295X.108.1.204

Conway, M. A., & Gathercole, S. E. (1987). Modality and long-term memory. *Journal of Memory and Language*, *26*(3), 341–361. https://doi.org/10.1016/0749-596X(87)90118-5

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, *105*(3), 658–668. https://doi.org/10.1016/j.cognition.2006.10.010

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*, 42–45.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321. https://doi.org/10.1037/0033-295X.93.3.283

Dell, G. S., & Chang, F. (2013). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120394–20120394. https://doi.org/10.1098/rstb.2012.0394

Dodson, C. S., & Schacter, D. L. (2001). "If I had said it I would have remembered it: Reducing false memories with a distinctiveness heuristic. *Psychonomic Bulletin & Review*, *8*(1), 155–161. https://doi.org/10.3758/BF03196152

Drake, E., & Corley, M. (2015). Articulatory imaging implicates prediction during spoken language comprehension. *Memory & Cognition*, *43*(8), 1136–1147. https://doi.org/10.3758/s13421-015-0530-6

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1016/0364-0213(90)90002-E

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505. https://doi.org/10.1111/j.1469-8986.2007.00531.x

Federmeier, K. D., & Kutas, M. (1999). Right words and left words: Electrophysiological evidence for hemispheric differences in meaning processing. *Cognitive Brain Research*, *8*(3), 373–392. https://doi.org/10.1016/S0926-6410(99)00036-1

Federmeier, K. D., Kutas, M., & Schul, R. (2010). Age-related and individual differences in the use of prediction during language comprehension. *Brain and Language*, *115*(3), 149–161. https://doi.org/10.1016/j.bandl.2010.07.006

Federmeier, K. D., Mclennan, D. B., De Ochoa-Dewald, 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*, 133–146. https://doi.org/10.1111/1469-8986.3920133

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research*, *1146*, 75–84. https://doi.org/10.1016/j.brainres.2006.06.101

Forrin, N. D., Jonker, T. R., & MacLeod, C. M. (2014). Production improves memory equivalently following elaborative vs non-elaborative processing. *Memory*, *22*(5), 470–480. https://doi.org/10.1080/09658211.2013.798417

Forrin, N. D., MacLeod, C. M., & Ozubko, J. D. (2012). Widening the boundaries of the production effect. *Memory & Cognition*, *40*(7), 1046–1055. https://doi.org/10.3758/s13421-012-0210-8

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Gathercole, S. E., & Conway, M. A. (1988). Exploring long-term modality effects: Vocalization leads to best retention. *Memory & Cognition*, *16*(2), 110–119. https://doi.org/10.3758/BF03213478

Goff, L. M., & Roediger, H. L. (1998). Imagination inflation for action events: Repeated imaginings lead to illusory recollections. *Memory & Cognition*, *26*(1), 20–33. https://doi.org/10.3758/BF03211367

Griffin, Z. M., & Bock, K. (1998). Constraint, Word Frequency, and the Relationship between Lexical Processing Levels in Spoken Word Production. *Journal of Memory and Language*, *38*(3), 313–338. https://doi.org/10.1006/jmla.1997.2547

Hintz, F., Meyer, A. S., & Huettig, F. (2017). Predictors of verb-mediated anticipatory eye movements in the visual world. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(9), 1352–1374. https://doi.org/10.1037/xlm0000388

Hopkins, R. H., & Edwards, R. E. (1972). Pronunciation effects in recognition memory. *Journal of Verbal Learning and Verbal Behavior*, *11*(4), 534–537. https://doi.org/10.1016/S0022-5371(72)80036-7

Kamide, Y. (2008). Anticipatory Processes in Sentence Processing. *Language and Linguistics Compass*, *2*(4), 647–670. https://doi.org/10.1111/j.1749-818X.2008.00072.x

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163–182. https://doi.org/10.1037/0033-295X.95.2.163

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). In M. Bar, *Predictions in the Brain: Using Our Past to Generate a Future* (pp. 190–207). Oxford University Press.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *BEHAVIORAL AND BRAIN SCIENCES*, 76.

Lindner, I., Drouin, H., Tanguay, A. F. N., Stamenova, V., & Davidson, P. S. R. (2015). Source and destination memory: Two sides of the same coin? *Memory*, *23*(4), 563–576. https://doi.org/10.1080/09658211.2014.911329

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition*, *6*(3), 312–319. https://doi.org/10.3758/BF03197461

MacLeod, C. M., & Bodner, G. E. (2017). The Production Effect in Memory. *Current Directions in Psychological Science*, *26*(4), 390–395. https://doi.org/10.1177/0963721417691356

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(3), 671–685. https://doi.org/10.1037/a0018785

Mani, N., & Huettig, F. (2012). Prediction during language processing is a piece of cake—But only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(4), 843–847. https://doi.org/10.1037/a0029284

Mani, N., & Huettig, F. (2013). Towards a complete multiple-mechanism account of predictive language processing. *Behavioral and Brain Sciences*, *36*(4), 365–366. https://doi.org/10.1017/S0140525X12002646

Marslen-Wilson, W. (1973). Linguistic Structure and Speech Shadowing at Very Short Latencies. *Nature*, *244*(5417), 522–523. https://doi.org/10.1038/244522a0

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

Matzen, L. E., & Benjamin, A. S. (2009). Remembering words not presented in sentences: How study context changes patterns of false memories. *Memory & Cognition*, *37*(1), 52–64. https://doi.org/10.3758/MC.37.1.52

Morey, R. D. (2008). Confidence Intervals from Normalized Data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*(2), 61–64. https://doi.org/10.20982/tqmp.04.2.p061

O'Brien, E. J., & Myers, J. L. (1985). When comprehension difficulty improves memory for text. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(1), 12–21. https://doi.org/10.1037/0278-7393.11.1.12

Ozubko, J. D., Gopie, N., & MacLeod, C. M. (2012). Production benefits both recollection and familiarity. *Memory & Cognition*, *40*(3), 326–338. https://doi.org/10.3758/s13421-011-0165-1

Ozubko, J. D., Hourihan, K. L., & MacLeod, C. M. (2012). Production benefits learning: The production effect endures and improves memory for text. *Memory*, *20*(7), 717–727. https://doi.org/10.1080/09658211.2012.699070

Ozubko, J. D., Major, J., & MacLeod, C. M. (2014). Remembered study mode: Support for the distinctiveness account of the production effect. *Memory*, *22*(5), 509–524. https://doi.org/10.1080/09658211.2013.800554

Perry, A. R., & Wingfield, A. (1994). Contextual encoding by young and elderly adults as revealed by cued and free recall. *Aging & Cognition*, *1*(2), 120–139. https://doi.org/10.1080/09289919408251454

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

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

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(04), 329–347. https://doi.org/10.1017/S0140525X12001495

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. https://doi.org/10.1038/4580

Rommers, J., & Federmeier, K. D. (2018a). Lingering expectations: A pseudo-repetition effect for words previously expected but not presented. *NeuroImage*, *183*, 263–272. https://doi.org/10.1016/j.neuroimage.2018.08.023

Rommers, J., & Federmeier, K. D. (2018b). Predictability's aftermath: Downstream consequences of word predictability as revealed by repetition effects. *Cortex*, *101*, 16–30. https://doi.org/10.1016/j.cortex.2017.12.018

Rommers, J., Meyer, A. S., & Huettig, F. (2015). Verbal and nonverbal predictors of language-mediated anticipatory eye movements. *Attention, Perception, & Psychophysics*, *77*(3), 720–730. https://doi.org/10.3758/s13414-015-0873-x

Rotello, C. M., & Macmillan, N. A. (2007). Response Bias in Recognition Memory. In A. S. Benjamin & B. H. Ross (Eds.), *Psychology of Learning and Motivation* (pp. 61–94). https://doi.org/10.1016/S0079-7421(07)48002-1

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General*, *110*(4), 474–494. https://doi.org/10.1037/0096-3445.110.4.474

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. https://doi.org/10.3758/BF03207704

Stanovich, K. E., & West, R. F. (1979). Mechanisms of sentence context effects in reading: Automatic activation and conscious attention. *Memory & Cognition*, *7*(2), 77–85. https://doi.org/10.3758/BF03197588

Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin*, *99*(1), 100–117. https://doi.org/10.1037/0033-2909.99.1.100

van Berkum, J. J. A. (2010). The brain is a prediction machine that cares about good and bad—Any implications for neuropragmatics? *Italian Journal of Linguistics*, *22*, 181–208.

Van Berkum, J. J. A., 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. https://doi.org/10.1037/0278-7393.31.3.443

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152–176. https://doi.org/10.1037/0033-295X.114.1.152

Wlotko, E. W., & Federmeier, K. D. (2007). Finding the Right Word: Hemispheric Asymmetries in the Use of Sentence Context Information. *Neuropsychologia*, *45*(13), 3001–3014. https://doi.org/10.1016/j.neuropsychologia.2007.05.013

Zormpa, E., Brehm, L. E., Hoedemaker, R. S., & Meyer, A. S. (2019). The production effect and the generation effect improve memory in picture naming. *Memory*, *27*(3), 340–352. https://doi.org/10.1080/09658211.2018.1510966

Appendix

Proportions of memory judgments

Table 1

Average proportion of each rating response in each condition in Experiment 1

	Predictable		Unpredictable		
Response	Aloud	Silent	Aloud	Silent	New
Sure New	.14 (.02)	.21 (.02)	.11 (.01)	.22 (.02)	.32 (.03)
Maybe New	.23 (.02)	.32 (.02)	.21 (.02)	.33 (.02)	.39 (.03)
Maybe Old	.25 (.02)	.27 (.02)	.24 (.01)	.24 (.01)	.22 (.01)
Sure Old	.39 (.02)	.19 (.02)	.44 (.02)	.21 (.02)	.07 (.01)

Note. The values between brackets represent unbiased within-participants SEM (Cousineau, 2005; Morey, 2008). Averages may not sum to 1 because of rounding.

Table 2

Average proportion of each rating response in each condition in Experiment 2

	Predictable		Unpredictable	
Response	Aloud	Silent	Aloud	Silent
Sure Silent	.15 (.01)	.37 (.02)	.14 (.01)	.35 (.02)
Maybe Silent	.30 (.02)	.40 (.02)	.25 (.02)	.40 (.02)
Maybe Aloud	.20 (.01)	.16 (.01)	.20 (.01)	.17 (.01)
Sure Aloud	.34 (.02)	.07 (.01)	.41 (.02)	.08 (.01)

Note. The values between brackets represent unbiased within-participants SEM. Averages may not sum to 1 because of rounding. Note that accuracy was higher for words read silently than words read aloud, but this reflects a bias to respond “silent”, which the analyses of discriminability in the main text circumvent.