The production effect over the long term: Modeling distinctiveness using serial positions

Véronique Cyr[1], Marie Poirier[2], James M. Yearsley[2], Dominic Guitard[1], Isabelle Harrigan[1], and Jean Saint-Aubin[1]

[1]School of Psychology, Université de Moncton

[2]Department of Psychology, City, University of London

Authors' Note

We have no known conflict of interest to disclose. This research was supported by Discovery grant RGPIN-2015-04416 from the Natural Sciences and Engineering Research Council of Canada to JSA. While working on the manuscript, DG was supported by a scholarship from NSERC, IH was supported by an undergraduate student research award from NSERC, and VC was supported by a joint scholarship from the New Brunswick Innovation Foundation and Université de Moncton.

Correspondence concerning this article should be addressed to Jean Saint-Aubin, School of Psychology, Université de Moncton, 18 ave Antonine-Maillet, Moncton, New Brunswick, E1A 3E9, Canada, Email: jean.saint-aubin@umoncton.ca

Open Practices Statement

The data for all experiments and the codes for the model are available on the Open Science Framework and will be made public upon publication (https://osf.io/p5gcb/?view_only=e6580cc760894eb3b86f45592d22e4b0).

Abstract

The production effect is a well-established finding: If some words within a list are read aloud, that is, produced, they are better remembered than their silently read neighbors. The effect has been extensively studied with long-term memory (LTM) tasks. Recently, using immediate serial recall and short-term order reconstruction, Saint-Aubin et al. (2021) reported informative interactions between the production effect and serial positions. Here, we asked whether these interactions would also be observed with the LTM tasks used in the field. In Experiment 1, pure and mixed lists of 8 words were presented in both order reconstruction and free recall tasks, with a 30-second filled retention interval. In Experiment 2, the list length was extended to 24 words; in Experiment 3, 10-word lists were used with a 2-minute retention interval. Results from all experiments aligned well with those observed in short-term memory. With mixed lists, where produced and silently read words alternated, produced items were better recalled, leading to sawtooth serial position curves. With pure lists, produced items were better recalled when studied in the last serial positions, but they were less well recalled for the primacy positions. Results were readily accounted for by the Revised Feature Model, originally developed to explain short-term memory performance. The findings and model suggest that produced items are encoded with more item-specific, modality-related features, that this generates a relative distinctiveness advantage in short- and long-term memory. However, the richer encoding comes at a cost: it appears to disrupt rehearsal.

The production effect over the long-term: Modeling distinctiveness using serial positions

Among the general principles that govern retrieval from memory, relative distinctiveness is often considered central (Crowder, 1976, 1993; Hunt, 2006; Surprenant & Neath, 2009). According to the relative distinctiveness principle, performance on a memory task is a function of the degree to which to-be-remembered items stand out relative to competing alternatives (Hunt, 2006). Relative distinctiveness has been called upon to explain what is known as the production effect: When words within a to-be-remembered list are produced –for example read aloud– they are better remembered than words within the same list that are read silently (see MacLeod & Bodner, 2017, for an overview). Here we revisit the production effect, present novel findings, and argue that this effect is observed in both short-term and long-term memory as a result of the influence of relative distinctiveness operating dynamically during both encoding and retrieval.

Over the last decade, the production effect has been extensively studied using long-term memory (LTM) tasks such as free recall and recognition. As predicted by the relative distinctiveness account, results have revealed a large production advantage in mixed lists containing both produced and silently read items (but see, Icht et al., 2014). However, with pure lists, where all items are either produced or silently read, produced items are no longer contrasted with contiguous, silent items. As a consequence, produced items are thought to lose much of their relative distinctiveness, and their advantage over silently read words is greatly reduced or disappears (Fawcett, 2013; Jones & Pyc, 2014).

Recently, Saint-Aubin et al. (2021) investigated the production effect in short-term memory (STM) and their results nicely extended those observed with LTM tasks. When mean recall performance for produced versus silently read words was examined, the pattern typical of

LTM was found: Mixed lists produced a clear advantage for produced items whereas pure lists saw this production effect significantly reduced or eliminated. Importantly, however, the production effect interacted rather dramatically with serial positions in ways that suggested both a benefit and a cost to producing items. We will return to this in more detail later in the introduction. Here we note that Saint-Aubin et al. (2021) accounted for their results by calling upon a few simple mechanisms: They tested their proposals experimentally and using a revised version of the Feature Model, a model of short-term recall effects in which relative distinctiveness is central (Nairne, 1988, 1990; Neath, 1999; Neath & Nairne, 1995).

The aim of the work herein was to test their proposals with tasks typical of the LTM production effect. As research relating to the production effect has mostly been conducted with classic LTM paradigms such as free recall and recognition, testing these ideas with such tasks is clearly relevant to our understanding of the production effect. To anticipate, the proposed mechanisms and the Revised Feature Model proved just as applicable to LTM tasks. One implication is that although a number of factors may distinguish LTM and STM, some principles cut across tasks and settings. Moreover, the findings suggest that production can have a cost in some circumstances, including in settings typical of LTM.

Serial positions have not been a focus of production effect research calling upon LTM paradigms. As a consequence, it was not possible to determine whether position-dependent effects like those reported by Saint-Aubin et al. (2021) would be observed with recall from LTM. Also, many would resist the suggestion that similar mechanisms can account for the production effect under STM and LTM conditions, considering the widespread view that the two systems are separable and differ in important ways (see, e.g., Norris, 2017, 2019). However, relative distinctiveness has been argued to be a general memory principle, that applies across time scales

and memory systems (see e.g., Brown et al., 2007; Surprenant & Neath, 2009), underscoring the usefulness of investigating the production effect and serial positions in LTM. Finally, establishing that comparable – or different – patterns of effects are observed for LTM would contribute to our understanding of the production effect and of the influence of relative distinctiveness more generally. Hence, in the current study, we pursued two aims: We asked to what degree the findings with STM tasks would also extend to LTM paradigms, and whether the Revised Feature Model (RFM) could also account for LTM findings.

Saint-Aubin et al. (2021): The cost of relative distinctiveness

In a series of six experiments, Saint-Aubin et al. (2021) investigated the production effect within classic short-term memory tasks, such as immediate serial recall. In the latter, a short sequence of items is sequentially presented; recall immediately follows presentation, and the order of recall must reproduce the order of presentation. Their results revealed a dramatic pattern of interactions between the production effect, serial positions, and list type (i.e., pure or mixed lists).

In their six-item mixed lists, words read aloud and words read silently systematically alternated. When responses were examined according to serial positions, a large sawtooth pattern emerged where the produced items were the peaks and the silently read items were the troughs (see Figure 1 for an illustration).

With pure lists, where all items were read either aloud or silently, overall, the production effect was small or absent, imitating findings in free recall (Forrin & MacLeod, 2016; Jones & Pyc, 2014; Jonker et al., 2014; Lambert et al., 2016). However, the serial positions told a different story. As shown in the left panel of Figure 1, there was a crossover interaction whereby

silently read items were better recalled than produced ones for the first positions, but the reverse was true for the last positions (see also Grenfell-Essam, Ward, & Tan, 2017; Macken et al., 2016).

Saint-Aubin et al. (2021) noted that the production advantage for the recency positions was very similar to the well-established modality effect, where the last item or last few items studied are better recalled when words are presented aurally, relative to visually (i.e., silently read; see Penney (1989), for a review). Given that production generates an auditory signal, one could expect a modality effect when produced items are compared to silently read ones. Saint-Aubin et al. tested this idea directly by comparing the produced items to items that were heard and seen simultaneously (audio-visual presentation). The size of the recency effect for produced items was indistinguishable from the modality effect observed with audio-visual presentation (see also, Crowder, 1970). In essence, the produced and audio-visual items showed the same recency advantage relative to the silently read items. Conversely, for primacy positions, the produced items differed from the audio-visual and silently read items: Words read aloud were at a disadvantage (see also Kappel et al., 1973). As suggested by Routh (1970), Saint-Aubin et al. hypothesized that this production disadvantage was due to subvocal rehearsal being disrupted by reading aloud. In a further experiment, in the silent condition, participants were required to say an irrelevant word aloud, once, after reading each to-be-remembered item silently. Under these conditions, produced words were better recalled in all positions, relative to words read silently. One could argue that the low-grade version of articulatory suppression used with the silent items may have led to a general dual-task decrement, resulting in a production advantage for all positions, instead of an advantage only for the last few positions. However, the findings do not clearly support this interpretation because the effect of saying one irrelevant word aloud

was pronounced for the early positions, but undetectable for the second half of the serial positions.

To explain the results obtained for both pure and mixed lists, Saint-Aubin et al. (2021) put forward an interpretation based on relative distinctiveness and on the cost of the richer encoding associated with reading words aloud. They suggested that once rehearsal was better equated, the increased item-specific encoding offered by production led to better recall across all positions. They argued that the enriched encoding came at a cost: To obtain the extra features afforded by production, resources that would otherwise be available for rehearsal are called upon. In essence, the benefit associated with aloud rehearsal is obtained at a cost to covert rehearsal. Importantly, Bhatarah et al. (2009) have shown that rehearsal is relied upon more heavily for the first items of a list, leading to the expectation that production would be more detrimental to the recall of earlier words.

To summarize, Saint-Aubin et al. (2021) offered an interpretation of their findings that relied on a few main ideas. One was increased covert rehearsal possibilities for silent (or audio-visual) conditions relative to production. A cornerstone of their view was the idea that production generates a greater amount of retrieval-relevant, reasonably distinct information, which benefits recall. Finally, they suggested that the recency advantage observed for produced items was analogous to the classic modality effect in STM (Penney, 1989).

To test this interpretation, Saint-Aubin et al. (2021) called upon a model that relied on these processes. The model made assumptions explicit and implemented the processes within a system where the influence of their simultaneous operation could be assessed. Saint-Aubin et al. then attempted to fit the data from their central experiments. The model, called the Revised Feature Model, was an adaptation of the Feature Model proposed by Nairne (1988, 1990) and

Neath and Nairne (1995), and it fit the data well. In discussing their findings, Saint-Aubin et al. surmised that the same principles might govern the production effect in LTM recall tasks (see, e.g., Crowder, 1993; Neath & Saint-Aubin, 2011; Neath et al., 2019; Surprenant & Neath, 2009). Others have suggested significant process similarities between immediate serial recall and free recall with list-length and task demands influencing rehearsal strategies and the initial output order (see, e.g., Bhatarah et al., 2008; Bhatarah et al., 2009; Grenfell-Essam & Ward, 2012; Grenfell-Essam et al., 2017). Under this view, with short lists, delayed free recall performance should be very similar to immediate serial recall, including at the level of serial positions.

Jonker et al. (2014) investigated the LTM production effect with 8-item pure and mixed lists in a delayed free recall task as well as a delayed order reconstruction task. In order reconstruction, a list of items is studied and is then presented again at the point of retrieval, but in a new random order. Participants are asked to reconstruct the original order in which the items were shown. In all cases, their presentation of the last to-be-remembered word was followed by a 30-second interfering task. Their overall findings can be compared with those of Saint-Aubin et al. (2021), who investigated immediate serial recall and immediate reconstruction. Within the free recall task in the Jonker et al. study, results revealed a robust production effect in mixed lists that vanished in pure lists. Saint-Aubin et al. (2021) obtained the same pattern in their findings. For the order reconstruction task, the results of Jonker et al. revealed an advantage of silently read items in pure lists that disappeared in the mixed list condition. This interaction was not reproduced in Saint-Aubin et al.: They obtained a mixed list advantage, which disappeared for pure lists in the immediate reconstruction task.

The aforementioned discrepancies between immediate and delayed tasks may be more apparent than real. As mentioned above, Saint-Aubin et al. (2021) showed that the production

effect interacts with serial positions: When results are examined averaging across positions, said interactions can lead to an advantage for produced items or for silently read items or to no difference (see also, Crowder, 1970; Greene & Crowder, 1984; Kappel et al., 1973; Macken et al., 2016). Jonker et al. (2014) did not report their serial position curves because, for their mixed lists, the within-list positions of produced and silently read words were randomized. Therefore, it is not possible to fully assess the similarities between immediate and delayed recall performance.

In the current study, we tested whether in LTM the production effect interacts with serial positions as it does in STM tasks. A comparable interaction would support the view that similar distinctiveness processes underpin the production effect in STM and LTM recall. We then tested the ability of the RFM to account for our findings. In Experiment 1, we used delayed recall and order recognition tasks with short lists like those used by Jonker et al. (2014). In Experiment 2, we used longer lists which are more frequently used to assess the production effect in LTM. Finally, in Experiment 3, a longer filled retention interval (2 minutes) was implemented to establish even more convincingly that the observed effects squarely belong in the realm of long-term memory.

Experiment 1

Our first experiment replicated the design of Jonker et al.'s (2014) Experiment 2 except that, in our mixed-list condition, produced and silently read items systematically alternated; this made results here directly comparable to those of Saint-Aubin et al. (2021) and made it possible to examine serial position curves. We reasoned that systematically alternating produced and silent items would also maximize any local distinctiveness effect—that is, the influence of having contiguous produced and silent items. As will be discussed later, the RFM includes a retroactive interference mechanism that predicts a heightened effect of local distinctiveness. As a

result, the size of the production effect was expected to be increased in mixed lists relative to what is observed when the position of produced and silent items is randomly determined [which leads to a lower frequency of contiguous, contrasting, produced and silent items]. According to the RFM, in pure lists, produced items should be less well recalled than silently read items for the initial serial positions. This disadvantage of produced items would be a consequence of pronouncing the items aloud which would interfere with covert rehearsal. As mentioned above, because the first items are usually rehearsed more frequently than other list items, the production disadvantage would be more pronounced for the initial serial positions (Bhatarah et al., 2009; Rundus, 1971; Tan & Ward, 2000; Ward, 2002). For the later serial positions, produced items should be better recalled than silently read items because the produced items would benefit from greater item distinctiveness provided by the extra modality-dependent components associated with production. In mixed lists, produced items should be systematically better recalled than silently read items. This would be evidenced by a sawtooth serial position curve.

With respect to delayed reconstruction, the predictions need to take into account that the task does not rely as heavily on item-specific information (see Neath, 1997). This is because, in this task, the to-be-retrieved words are provided at the point of recall, in new random positions, and participants are asked to reconstruct the original presentation order. Less reliance on the item-specific information encoded will reduce the influence of relative distinctiveness. We hence expected the same pattern of findings for the reconstruction task but with smaller discrepancies between silent and produced items across the board. For the first time, the RFM was called upon to model the reconstruction results, with a small change in the assumption governing the operation of the model; we will return to this when the model and simulations are presented following the description of the three empirical studies that we report.

Method

Participants. Sixty-four students (13 men, 51 women; mean age of 21.81 years) from Université de Moncton took part in this experiment for a small monetary honorarium. All participants were native French speakers and had normal or corrected to normal vision. The sample sizes for this and subsequent experiments were based on the effect size seen in Experiment 1 of Saint-Aubin et al. (2021) for the interaction between production and serial position in pure and mixed lists in the immediate serial recall task (η_p^2 = .43, η_p^2 = .63) and the order reconstruction task (η_p^2 = .59, η_p^2 = .78). An a priori power analysis computed with the Superpower package in *R* (Lakens & Caldwell, 2021) indicated that a sample of size of 48 participants would have power greater than .90 to detect the smallest effect size reported in Experiment 1 of Saint-Aubin et al. (2021). Given the uncertainties of replicating with long-term memory tasks findings observed with short-term memory tasks, we decided to overpower the first experiment with 64 participants. To further ascertain the sample size, we conducted a sensitivity analysis. Based on the partial eta squared reported above, the size of the effect would be between a Cohen's *f* of .87 and a Cohen's *f* of 1.88. We adopted a conservative approach by assuming the presence of only a medium effect size interaction (Cohen's *f* = .20). The sentivity analysis revealed that 64 participants would also have a power greater than .85 to detect the interaction between production and serial position (1 to 8).

Material. The stimuli were 448 French words taken from the database *Lexique* 3.83 (New et al., 2004). All words were one-syllable, between three and seven letters long, singular, and were common nouns. A total of 56 lists of 8 words were assembled by randomly drawing without replacement from this pool of 448 words. From these 56 lists, 8 were selected to serve as practice trials. The remaining 48 lists were used for experimental trials. The 8 practice lists and

48 experimental lists were the same for all participants, as was word order within the lists. Within a list, care was taken to ensure that words did not rhyme and were not semantically related to each other.

Design. A 2 x 2 x 2 within-participant design was used with list type (pure vs. mixed), presentation modality (read silently vs. produced) and recall task (free recall vs. order reconstruction) as factors. Participants undertook 8 practice trials (one per condition), followed by 48 experimental trials (6 per condition). Across participants, each experimental list was used as often for each of the 8 conditions. The presentation order of the lists was randomized for each participant. For the mixed-list condition, four items were read silently and four items were read aloud, in an alternating pattern. For half of these lists, odd items were read aloud; for the other half, even items were read aloud. In the pure-list condition, half of the lists were read silently whereas the other half was read aloud.

Procedure

To-be-remembered words were presented on a white background on a computer screen using E-Prime 2.0 software (Psychology Software Tools, 2016). The computer screen was located approximately 60 cm away from the participant. Eight words were sequentially presented at the center of the screen at a rate of 2 seconds per word (2000 msec on, 0 msec off). In the pure list condition, for half of the participants, all list words were displayed in blue, whereas they were displayed in red for the other half. In the mixed list condition, for half of the participants, blue words had to be read aloud and red words had to be read silently; it was the opposite for the other half. Two seconds after the offset of the last presented word in each list, a digit appeared in the center of the screen. This digit signaled the beginning of the 30-second interference task. During the interference task, a series of single digits (0-9) appeared individually in the center of

the screen. For each digit, participants indicated whether it was odd or even. Participants were instructed to press the 'Z' key if the stimulus was an odd number and the 'M' key if the stimulus was an even number. The task was self-paced, but lasted 30 seconds for all participants.

Immediately after the interference task, the identity of the memory task was revealed. Although presentation conditions varied randomly from trial to trial, all trials were answered in the same test booklet. For both tasks, recall was self-paced without time limit. For each trial, there were 8 lines displayed horizontally. For the order reconstruction task, the eight presented words were shown in alphabetical order in a vertical list at the center of the screen. Participants answered by writing the words in their presentation order, starting with the first presented item. They were instructed to write from left to right, and to avoid backtracking. For the free recall task, a white screen with three questions marks was shown. Participants were asked to recall the presented words in no particular order. Participants were told to recall the words on the answer sheets as they came to mind, without trying to place them in order. The researcher was present throughout the testing session to ensure compliance with the instructions.

Results

For all experiments, data from the order reconstruction and free recall tasks were analyzed separately. Within both tasks, participants produced misspellings. We first corrected misspellings by replacing them with the proper word as long as it could be unambiguously identified[1]. These corrections did not change the pattern of results, but slightly increased the overall level of performance. Data were then coded in the following manner. For the order

[1] The same person corrected the misspellings for all three experiments. The information about the conditions was available in the data file, but no specific instruction was provided about this information.

reconstruction task, an item was considered correct when it was produced at its exact serial position. For the free recall task, an item was considered correct if it was recalled, irrespective of its recall position. However, serial positions were produced by examining performance as a function of word position at study. In other words, if the fifth presented word was recalled second, credit was given at the fifth serial position.

Before presenting the results of the main analyses, we verified performance at the parity judgment task to ensure that participants were adequately engaged in it. For all experiments, the proportion of correct parity judgment and the number of parity judgment attempts as function of the recall task (free recall vs. order reconstruction) and list type (silent, aloud, AS, SA) are shown in Table 1. As can be seen in the table, participants were actively engaged in the distracting task and their performance level did not differ across conditions. The 2 X 4 repeated-measures ANOVA with recall task and list type as factors revealed that the proportion of correct responses was consistent across conditions, all *F*s < 1. Furthermore, the ANOVA with recall task and list type as factors on the number of parity judgment attempts revealed a main effect of list type, $F(3,189) = 4.70$, $p = .003$, $\eta_p^2 = .07$, but neither the main effect of recall task, $F(1,63) = 1.94$, $p = .168$, $\eta_p^2 = .03$, nor the interaction were significant, $F(3,189) = 1.17$, $p = .324$, $\eta_p^2 = .02$. Post hoc Tukey's HSD tests revealed that there were slightly more parity judgments in the silent than then produced condition or the AS condition.

For the main analyses, the proportion of correct recall as a function of input modality, list type, and task is shown in Figure 2, and serial positions are presented in Figure 3. An inspection of Figure 2 reveals the standard production effect in the mixed list condition with a large advantage of produced words over words read silently. With pure lists, this advantage is

abolished in free recall and reversed in the order reconstruction task. This pattern of results nicely reproduces the results found in the Jonker et al. (2014) study.

An examination of performance as a function of serial position provides further insight into the effect of production. As observed by Saint-Aubin et al. (2021) in short-term ordered recall tasks, with mixed lists, sawtooth serial position curves emerged for both tasks, with larger effects for the free recall task. Furthermore, again mimicking the Saint-Aubin et al. findings, in the pure-list condition, for both tasks, words read silently were better recalled for the first serial positions and produced words were better recalled for the last serial positions.

Order reconstruction task. To facilitate a comparison with the findings of Jonker et al. (2014), a 2 X 2 repeated-measures ANOVA with list type (pure vs. mixed) and presentation modality (read silently vs. produced) as factors revealed a main effect of list type, $F(1,63) = 8.88$, $p < .001$, $\eta_p^2 = .12$, and an interaction between list type and presentation modality, $F(1,63) = 31.34$, $p < .001$, $\eta_p^2 = .33$. The main effect of presentation modality was not significant, $F(1,63) = 3.82$, $p > .05$, $\eta_p^2 = .06$. Post-hoc comparisons showed that produced items were better reordered than silently read items in mixed lists, $p < .001$, Cohen's $d = 0.709$, whereas produced items were less well reordered than silently read items in pure lists, $p = .027$, Cohen's $d = 0.284$. In addition, silently read items were better reordered in pure than in mixed lists, $p < .001$, Cohen's $d = 0.716$, but there was no difference for produced items, $p = .15$, Cohen's $d = 0.181$.

The data were further analyzed by means of two 2 X 8 repeated measures ANOVAs with presentation modality (read silently vs. produced) and serial position (1 to 8) as factors. For pure lists, the ANOVA revealed significant main effects of presentation modality, $F(1,63) = 5.15$, $p < .05$, $\eta_p^2 = .08$, and of serial position, $F(7,441) = 34.31$, $p < .001$, $\eta_p^2 = .35$, and an interaction between presentation modality and serial position, $F(7,441) = 5.07$, $p < .001$, $\eta_p^2 = .07$. Post hoc

Tukey's HSD tests revealed that silently read words were significantly better reordered than produced items at Position 2, $p < .05$, Position 3, $p < .05$, Position 4, $p < .001$, and Position 5, $p < .05$, whereas produced items were better reordered at Position 8, $p < .05$. There was no significant difference at Positions 1, 6 and 7, all *p*s > .06. For the mixed list condition, the ANOVA revealed a significant main effect of serial position, $F(7,441) = 32.47$, $p < .001$, $\eta_p^2 = .34$, and an interaction between presentation modality and serial position, $F(7,441) = 10.31$, $p < .001$, $\eta_p^2 = .14$. There was no significant main effect of presentation modality, $F < 1$. Post hoc Tukey's HSD tests revealed that words read aloud were better reordered at Position 1, $p < .001$, Position 2, $p < .05$, Position 3, $p < .01$, Position 7, $p < .001$, and Position 8, $p < .001$. There was, however, no significant difference for Positions 4 through 6, all *p*s > .14.

Free recall task. As was done with the order reconstruction task, a 2 X 2 repeated-measures ANOVA with list type (pure vs. mixed) and presentation modality (read silently vs. produced) as factors revealed main effects of list type, $F(1,63) = 4.09$, $p < .05$, $\eta_p^2 = .06$., and of presentation modality, $F(1,63) = 163.60$, $p < .001$, $\eta_p^2 = .72$, and an interaction between list type and presentation modality, $F(1,63) = 166.64$, $p < .001$, $\eta_p^2 = .73$. Post-hoc comparisons revealed that produced items were better recalled than silently read items in mixed lists, $p < .001$, Cohen's $d = 2.02$, whereas there was no significant difference between produced words and words read silently in pure lists, $p = .44$, Cohen's $d = 0.097$. In addition, silently read items were better recalled in pure than in mixed lists, $p < .001$, Cohen's $d = 1.385$, but it was the reverse for produced items, with better recall in mixed than in pure lists, $p < .001$, Cohen's $d = 0.972$.

Once again, two 2 X 8 repeated measures ANOVAs with presentation modality (read silently vs. produced) and serial position (1 to 8) as factors were computed. For pure lists, the ANOVA revealed a main effect of serial position, $F(7,441) = 12.65$, $p < .001$, $\eta_p^2 = .17$, and an

interaction between presentation modality and serial position, $F(7,441) = 13.19$, $p < .001$, $\eta_p^2 = .17$. There was no main effect of presentation modality, $F < 1$. Post hoc Tukey's HSD tests revealed that words read silently were better remembered than words read aloud at Position 1, $p < .05$, Position 2, $p < .05$, and Position 4, $p < .01$; the reverse was found at Position 7, $p < .01$, and Position 8, $p < .001$. There were no significant differences for Position 3, 5, and 6, all ps $> .05$. For mixed lists, the ANOVA revealed a main effect of serial position, $F(7,441) = 16.73$, $p < .001$, $\eta_p^2 = .21$, and an interaction between presentation modality and serial position, $F(7,441) = 88.73$, $p < .001$, $\eta_p^2 = .58$. There was no main effect of presentation modality, $F < 1$. Post hoc Tukey's HSD tests revealed that words read aloud were better recalled than words read silently at all serial positions, all ps $< .001$.

Discussion

When averaged across positions, the results from Experiment 1 revealed a significant mixed-list production advantage for both order reconstruction and free recall tasks, with a stronger effect in the latter. With pure lists, the effect was abolished in free recall whereas order reconstruction showed a reversal. This pattern nicely replicated the results reported by Jonker et al. (2014). Most importantly, for both tasks, the analysis of serial positions in mixed lists revealed a systematic advantage for produced words compared to words read silently; in pure lists, there was a cross-over effect with an advantage for words read silently in primacy positions and an advantage for words read aloud in recency positions.

Results at the parity judgment task indicated that participants were actively engaged in the distractor activity and equally so in all conditions, despite a small difference across conditions in the number of attempts. Nevertheless, the pattern of results at the LTM tasks mimics the results reported by Saint-Aubin et al. (2021) in immediate serial recall – i.e. a classic

STM paradigm. As previously mentioned, Saint-Aubin et al. (2021) accounted for the crossover observed for pure lists by suggesting that the production advantage for the last serial positions derives from the auditory features that are generated by aloud pronunciation; the latter are thought to mimic classic modality effects. In the case of the production disadvantage for the initial positions, the authors suggested that rehearsal was differentially influenced by reading aloud versus silently. They hypothesized that pronouncing aloud is associated with a certain degree of covert rehearsal suppression, thought to interfere with performance for words read aloud. As the results of Experiment 1 are very similar to those reported by Saint-Aubin et al. (2021), it is straightforward to suggest that the same explanation could apply to the free recall findings, albeit based on results with short lists of words. In support of this view, Bhatarah et al. (2009) found identical patterns of rehearsal in immediate serial recall and in free recall for 8-item lists. In mixed lists, as observed here, Saint-Aubin et al. reported a sawtooth pattern across positions, with an advantage for words read aloud compared to words read silently. They interpreted their findings using a distinctiveness account, according to which producing words aloud involves distinctive processing that supports retrieval of the produced items. Moreover, because production is thought to hinder rehearsal, silently read words would be less well recalled in mixed lists than in pure lists.

Experiment 2

To follow the above replication of Jonker et al. (2014), and in keeping with concerns raised by Forrin and MacLeod (2016) and Lambert et al. (2016) regarding the length of lists used by Jonker et al., Experiment 2 aimed to extend these findings to LTM recall with longer lists—specifically lists of 24 words. Such list lengths are much more typical of free recall tasks (Bhatarah et al., 2009). Given the added difficulty of the free recall task with longer lists, similar

but attenuated serial position patterns are expected for mixed lists, with an advantage for produced relative to silently read items. In short, a sawtooth serial position pattern is predicted for mixed lists, even when these lists are much longer. With respect to pure lists, as found in Experiment 1 and in STM tasks, the first silent items should be better recalled than the first produced items. In effect, even with long lists and a free recall task, previous studies have shown that early items are still more frequently rehearsed than later items (see, e.g., Rundus, 1971; Tan & Ward, 2000; Ward, 2002).

To provide comparability with Experiment 1, we included an order reconstruction task. However, based on the immediate serial recall results of Bhatarah et al. (2009) with 12-item lists, we anticipated that participants would have low performance levels. Furthermore, Mulligan and Lozito (2007) tested free recall for lists of 8, 16, and 24 words. Although they found a close link between order information and free recall for 8-item lists, the relation totally vanished with 24-item lists. Therefore, in addition to potential floor effect challenges, it remains to be seen whether the pattern of results observed with 8-item lists can be reproduced in attenuated form with 24 items.

Method

Participants. Forty-eight students (8 men, 40 women; mean age of 19.81 years) from Université de Moncton took part in this experiment and received course credits for their participation. All participants were native French speakers and had normal or corrected to normal vision. None had taken part in Experiment 1. The number of participants was based on the a priori power analysis reported in Experiment 1 indicating that 48 participants provide a power greater than .90 for detecting the critical interactions. A sentivity analysis also revealed that 48

participants would also have a power greater than .90 to detect a medium effect size interaction (Cohen's f = .20) between production and serial position (1 to 24).

Material. One-hundred and twenty-eight French words taken from the database *Lexique* 3.83 (New et al., 2004) were added to those used in Experiment 1. All words were one-syllable, and between three and seven letters long, singular, and were common nouns. Twenty-four lists of twenty-four words were assembled by randomly drawing without replacement from this pool of 576 words. From these 24 lists, 8 were selected to serve as practice trials. These lists were the same for all participants. The remaining 16 lists were used for experimental trials. The 16 experimental lists were the same for all participants, as was word order within the lists. As was done in Experiment 1, within a list, care was taken to ensure that words did not rhyme and were not semantically related to each other.

Design and Procedure. The design and procedure were the same as in Experiment 1, except for the following changes. The testing session lasted for approximately 90 minutes. Participants undertook 8 practice trials, followed by 16 experimental trials. The lists were composed of 24 words. Participants recalled words from the list in a 3 X 8 grid in an answer booklet.

Results and Discussion

The proportion of correct recall as a function of input modality, list type, and task is shown in Figure 2, and serial positions are presented in Figure 4. Unsurprisingly, performance with lists of 24 words is much lower than what was observed in Experiment 1 with 8-item lists; this is particularly true for the order reconstruction task, where performance is around or below 10% correct in all conditions. However, the pattern presented for free recall is similar to what we

reported in Experiment 1 and to previous findings with longer lists (Forrin & MacLeod, 2016; Jones & Pyc, 2014; Lambert et al., 2016). It is worth noting that Experiment 2 was a conceptual replication of Forrin and MacLeod. The main differences being our usage of French stimuli, the systematic alternation of produced and silently read items in the mixed list condition and our inclusion of an order reconstruction task on half of the trials. Despite these changes, results at the free recall task are almost identical to those of Forrin and MacLeod demonstrating the reproducibility of the effect. In addition, the similarity of results across studies suggest that our inclusion of an order reconstruction task did not change how participants processed information in free recall.

We now turn to Figure 4 presenting performance as a function of serial position. With the order reconstruction task and mixed lists, results suggest a small benefit of produced items. With pure lists, there was no clear difference between the two presentation modalities, apart from an advantage for words read aloud compared to words read silently on the first and last list items. With free recall, produced words are almost systematically better remembered than are the words read silently, again resulting in a distinctive sawtooth curve. In pure lists, words read silently are better recalled than words read aloud on some of the early serial positions, whereas the reverse is observed for the second half of the list; however, these trends are not systematically observed throughout the list. Overall, this equivalence between LTM results in free recall and those observed in STM (Grenfell-Essam et al., 2017; Macken et al., 2016; Saint-Aubin et al., 2021) suggests that similar principles govern the influence of production on memory irrespective of the time scale involved in the task (Crowder, 1993; Surprenant & Neath, 2009).

As in Experiment 1, results shown in Table 1 revealed that participants were actively engaged in the distractor task. The 2 X 4 repeated-measures ANOVA with recall task and list

type as factors for the proportion of correct responses at the parity judgment task revealed that neither the main effect of list type $F(3,141) = 1.58$, $p = .197$, $\eta_p^2 = .03$, nor of recall task, $F < 1$, nor the intereaction were significant, $F < 1$. The ANOVA on the number of attempts also failed to reaveal a main effect of list type $F < 1$, recall task, $F(1,47) = 2.22$, $p = .143$, $\eta_p^2 = .05$, or the intereaction, $F(3,141) = 1.03$, $p = .383$, $\eta_p^2 = .02$.

Order reconstruction task. A 2 X 2 repeated-measures ANOVA with list type (pure vs. mixed) and presentation modality (read silently vs. produced) as factors revealed a main effect of presentation modality, $F(1,47) = 7.77$, $p < .01$, $\eta_p^2 = .14$. The main effect of list type was not significant, $F(1,47) = 1.14$, $p = .29$, $\eta_p^2 = .02$, nor was the interaction between list type and presentation modality, $F(1,47) = 2.52$, $p = .12$, $\eta_p^2 = .05$.

The data were further analyzed by means of two 2 X 24 repeated measures ANOVAs, one for each list type (pure vs. mixed), with presentation modality (read silently vs. produced) and serial position (1 to 24) as factors. For pure lists, the ANOVA revealed a main effect of serial position, $F(23,1081) = 11.86$, $p < .001$, $\eta_p^2 = .20$. There was no main effect of presentation modality, and no interaction between presentation modality and serial position, both Fs < 1. For the mixed lists, the ANOVA revealed a main effect of serial position, $F(23,1081) = 13.79$, $p < .001$, $\eta_p^2 = .23$, plus an interaction between presentation modality and serial position, $F(23,1081) = 3.43$, $p < .001$, $\eta_p^2 = .07$. There was no main effect of presentation modality, $F(1,47) = 2.31$, $p = .14$, $\eta_p^2 = .05$. Post hoc Tukey's HSD tests on the mixed lists revealed that words read aloud were better reordered than words read silently at Position 1, $p < .01$, Position 9, $p < .05$, Position 22, $p < .01$, and Position 24, $p < .01$, whereas words read silently were better reordered than words read aloud at Position 10 and Position 15, both ps $< .05$. There were no significant differences for all other serial positions, all ps $> .05$.

Free recall task. The 2 X 2 repeated-measures ANOVA with list type (pure vs. mixed) and presentation modality (read silently vs. produced) as factors revealed a main effect of presentation modality, $F(1,47) = 33.63$, $p < .001$, $\eta_p^2 = .42$, of list type, $F(1,47) = 4.67$, $p < .05$, $\eta_p^2 = .09$, and an interaction between list type and presentation modality, $F(1,47) = 46.95$, $p < .001$, $\eta_p^2 = .50$. Post-hoc comparisons revealed that produced words were better recalled than words read silently in mixed lists, $p < .001$, Cohen's $d = 1.483$, whereas there was no significant difference between the two input modalities in pure lists, $p = .865$, Cohen's $d = 0.025$. In addition, silently read items were better recalled in pure than in mixed lists, $p < .001$, Cohen's $d = 0.975$, whereas it was the reverse for produced items where recall was better in mixed than in pure lists, $p < .001$, Cohen's $d = 0.531$.

For pure lists, the 2 X 24 repeated-measures ANOVA with presentation modality (read silently vs. produced) and serial position (1 to 24) as factors revealed a main effect of serial position, $F(23,1081) = 7.13$, $p < .001$, $\eta_p^2 = .13$, and an interaction between presentation modality and serial position, $F(23,1081) = 4.13$, $p < .001$, $\eta_p^2 = .08$. There was no main effect of presentation modality, $F < 1$. Post hoc Tukey's HSD tests revealed that words read silently were better remembered than words read aloud at Position 5, $p < .05$, Position 7, $p < .001$, and Position 10, $p < .05$, whereas the reverse was found at Position 23, $p < .05$, and Position 24, $p < .001$. There were no significant differences at the other serial positions, all ps $> .05$. For mixed lists, the ANOVA revealed main effects of presentation modality, $F(1,47) = 5.35$, $p < .05$, $\eta_p^2 = .10$, and of serial position, $F(23,1081) = 5.89$, $p < .001$, $\eta_p^2 = .11$, and an interaction between presentation modality and serial position, $F(23,1081) = 9.68$, $p < .001$, $\eta_p^2 = .17$. Post hoc Tukey's HSD tests revealed that words read aloud were better recalled than words read silently

in 16 of the 24 positions, namely 1, 2, 3, 6, 7, 8, 11, 12, 14, 16, 18, 20, 21, 22, 23, 24, all *p*s < .05; the difference was not significant at the remaining 8 positions, all *p*s > .05.

Overall, the general pattern of results is similar to the one observed in Experiment 1, but it is noisier. Two factors can account for the greater noise in Experiment 2. First, there were fewer observations per data point than in Experiment 1. In effect, 64 participants took part to Experiment 1, while only 48 participants took part to Experiment 2. In addition, there were only 2 trials per participant per condition, to keep the duration of the experiment within reasonable limits with longer lists, while there were 6 trials per participant per condition in Experiment 2. Second, the usage of long lists combined with a filled retention interval severely reduced performance. Experiment 3 was designed to overcome these two potential issues and to implement a more stringent test of the implication of LTM.

Experiment 3

The results of the first two experiments extend previous findings with immediate serial recall and immediate order reconstruction to LTM paradigms. The experiments were modeled on previously published studies of the production effect in LTM (Forrin & MacLeod, 2016; Jonker et al., 2014). However, one could argue that 30 seconds of distractor activity does not provide a stringent test of the influence of production on retrieval from LTM. In effect, the duration for which the information can be maintained in short-term or working memory is controversial (Cowan, 2017a; Voyer, Saint-Aubin, Altman, & Gallant, 2021) and some theorists suggest that some portion of working memory may remain active for up to a minute (see Cowan, 2017b, for a review). The duration of the retention interval was less of a concern in Experiment 2 in which

long lists were used, but as mentioned above, some of those effects could have been clouded by floor level performance.

To address both potential issues, and to better establish the robustness and reliability of the pattern of findings, Experiment 3 used a two-minute filled retention interval with the same distractor activity as used in Experiment 1 and 2, along with 10-item lists. If the first two experiments relied on retrieval from LTM, we should observe a similar pattern of results with sawtooth serial positions in mixed lists. As for pure lists, we expect a deleterious effect of production on the first serial positions and a beneficial effect on the last serial positions. Overall performance level should be somewhat between what was observed in Experiments 1 and 2.

Method

Participants. Seventy-two students (19 men, 51 women, 2 participants preferred not to report; mean age of 21.58 years) from Université de Moncton were paid $20 for their participation. All participants were native French speakers and had normal or corrected to normal vision. None of the participants took part in the previous experiments. Although the a priori power analysis revealed that 48 participants would have been enough to achieve a power of .90 to uncover the interaction between the production effect and serial position, we selected 72 participants to achieve more stable performance estimates for the simulations, because there were only two trials per condition per participant. In addition the sentivity analysis revealed that 72 participants would have a power greater than .90 to detect a medium effect size interaction (Cohen's f = .20) between production and serial position (1 to 10).

Materials. Two-hundred and twenty words were taken from Experiment 2. Twenty-two lists of ten words were assembled using the same constraints as before. From these 22 lists, 6

were selected to serve as practice trials. These lists were the same for all participants. The remaining 16 lists were used for experimental trials. As in Experiments 1 and 2, the experimental lists were the same for all participants, as well as word order within the lists, although the order of lists was randomly determined for each participant.

Design and Procedure. The design and procedure were the same as in Experiments 1 and 2, except for the following changes. The experiment was programmed with PsyToolKit (Stoet, 2010, 2017). The testing session lasted for approximately 90 minutes. In accordance with public health guidelines, the participants were tested remotely on their personal computers while the experimenter was present via Microsoft Teams or Zoom throughout the session to ensure compliance with the instructions and to monitor the testing environment. Participants were required to turn on their camera and microphone and to share their screen. The duration of the retention interval filled with the parity judgment task was increased from 30 seconds to 2 minutes. Participants recalled words from the list using their keyboard, pressing the enter key after each word. Participants' responses were displayed on a 2 X 5 grid.

Results and Discussion

The proportion of correct recall as a function of input modality, list type and task is shown in Figure 2, and serial positions are presented in Figure 5. As expected, performance with lists of 10 words was superior to performance with 24 words in Experiment 2 and lower than performance with 8 words in Experiment 1. Despite the methodological changes—10-item lists and 2 minutes of parity judgment—the pattern of results was similar to the pattern observed in the first two experiments. With both the order reconstruction task and the free recall task, a benefit of produced items was observed for mixed lists and no clear difference was observed for pure lists between the two presentation modalities.

As shown in Figure 5, presenting performance as a function of serial position, results were again similar to those observed in the previous experiments. For the order reconstruction task and mixed lists, a small benefit of produced items was observed, and with pure lists, produced items were slightly less well recalled than silent items on the first serial positons. With free recall, a large sawtooth serial position curve was observed with mixed lists. For pure lists, again consistent with our previous demonstrations, words read silently were better recalled than words read aloud on the early serial positions, whereas the reverse was observed for the last serial positions. The results of the first two experiments are consistent with retrieval from LTM.

An examination of Table 1 revealed that participants were engaged in the parity judgment task for the full two minutes, and were so to the same extent in all conditions. Accordingly, the repetead-measures ANOVA on the proportion of correct decisions with recall task and list type as factors revealed the absence of a main effect of list type, $F(3,213) = 1.25$, $p = .293$, $\eta_p^2 = .02$, of recall task, $F(1,71) = 1.22$, $p = .273$, $\eta_p^2 = .02$, or of the intereaction, $F(3,213) = 2.51$, $p = .060$, $\eta_p^2 = .03$. The ANOVA on the number of parity judgment attempts revealed that neither the main effects nor the interaction were significant, all Fs< 1.

Order reconstruction task. Results from the 2 X 2 repeated-measures ANOVA with list type (pure vs. mixed) and presentation modality (read silently vs. produced) revealed a main effect of list type, $F(1,71) = 8.10$, $p < .01$, $\eta_p^2 = .10$. The main effect of presentation modality was not significant, $F(1,71) = 1.58$, $p = .21$, $\eta_p^2 = .02$, but there was a significant interaction between list type and presentation modality, $F(1,71) = 10.35$, $p < .001$, $\eta_p^2 = .13$. Post-hoc comparisons showed that produced items were better reordered than silently read items in mixed lists, $p < .001$, Cohen's $d = 0.557$, but there was no significant difference between produced items and silently read items in pure lists, $p = .232$, Cohen's $d = 0.142$. In addition, silently read

items were better reordered in pure than in mixed lists, $p < .001$, Cohen's $d = 0.486$, whereas there was no difference for produced items, $p = .677$, Cohen's $d = 0.049$.

We further analyzed the data by conducting two 2 X 10 repeated measures ANOVAs, one for each list type (pure vs. mixed), with presentation modality (read silently vs. produced) and serial position (1 to 10) as factors. For pure lists, the ANOVA revealed a main effect of serial position, $F(9, 639) = 17.72$, $p < .001$, $\eta_p^2 = .20$. There was no main effect of presentation modality, $F(1,71) = 1.45$, $p = .23$, $\eta_p^2 = .02$, and no interaction between presentation modality and serial position, $F(9,639) = 1.32$, $p = .22$, $\eta_p^2 = .02$.

For mixed lists, the ANOVA revealed a main effect of serial position, $F(9,639) = 18.54$, $p < .001$, $\eta_p^2 = .21$, and an interaction between presentation modality and serial position, $F(9,639) = 3.83$, $p < .001$, $\eta_p^2 = .05$. However, there was no main effect of presentation modality, $F(1,71) = 2.42$, $p = .12$, $\eta_p^2 = .03$. Post hoc Tukey's HSD tests revealed that words read aloud were better reordered than words read silently at Position 1, $p < .001$, and Position 3, $p < .05$; there were no significant differences for the remaining positions, all ps > .05.

Free recall task. The 2 X 2 repeated-measures ANOVA with list type (pure vs. mixed) and presentation modality (read silently vs. produced) as factors revealed a main effect of presentation modality, $F(1,71) = 81.98$, $p < .001$, $\eta_p^2 = .54$, of list type, $F(1,71) = 10.23$, $p < .01$, $\eta_p^2 = .13$, and an interaction between these factors, $F(1,71) = 74.66$, $p < .001$, $\eta_p^2 = .51$. Post-hoc comparisons revealed that produced words were better recalled than words read silently in mixed lists, $p < .001$, Cohen's $d = 11.446$, while there was no difference between the two input modalities in pure lists, $p = .557$, Cohen's $d = 0.070$. Furthermore, silently read items were better recalled in pure than in mixed lists, $p < .001$, Cohen's $d = 1.06$, whereas the reverse was

observed for produced items, which were better recalled in mixed than in pure lists, $p < .001$, Cohen's $d = 0.402$.

For pure lists, the 2 X 10 repeated-measures ANOVA with presentation modality (read silently vs. produced) and serial position (1 to 10) as factors revealed a main effect of serial position, $F(9,639) = 9.95$, $p < .001$, $\eta_p^2 = .12$, and an interaction between presentation modality and serial position, $F(9,639) = 3.52$, $p < .001$, $\eta_p^2 = .05$. However, there was no main effect of presentation modality, $F < 1$. Post hoc Tukey's HSD tests revealed that words read silently were better remembered than words read aloud at Position 2, $p < .05$, whereas the reverse was found at Position 6, $p < .05$, Position 9, $p < .05$, and Position 10, $p < .01$. There were no significant differences at the other serial positions, all *p*s > .05.

For mixed lists, the ANOVA revealed a main effect of presentation modality, $F(1,71) = 2.11$, $p = .15$, $\eta_p^2 = .03$, and of serial position, $F(9,639) = 5.19$, $p < .001$, $\eta_p^2 = .07$, and an interaction between presentation modality and serial position, $F(9,639) = 24.85$, $p < .001$, $\eta_p^2 = .26$. Post hoc Tukey's HSD tests revealed that words read aloud were better recalled than words read silently in all positions, all *p*s < .05.

The Revised Feature Model

As its name implies, the Revised Feature Model (RFM) is an adaptation of the Feature Model (Nairne, 1988, 1990; Neath & Nairne, 1995; Neath & Surprenant, 2007). In their adaptation of the Feature model, Saint-Aubin et al. (2021) retained its key elements while adding a rehearsal process and slightly modifying the overwriting process that the original model included. We now turn to a succinct description of the main characteristics of the model. In this

section we are concerned with how information about the items is encoded; in the next section we will describe the retrieval process in order reconstruction and free recall.

In the RFM, as in the original Feature Model, items are represented by two types of features. On one hand, encoding is thought to generate modality-dependent features, related to physical presentation conditions such as item color or voice quality. On the other hand, items also produce modality-independent features, generated by internal processes of categorization and identification (e.g. gender of speaker, meaning of word, category of the visual item presented). This characteristic of the RFM naturally lends itself to modelling differences related to presentation modalities such as those involved in the experiments reported here.

At presentation, items simultaneously generate traces in primary and secondary memory. In both cases, items are represented by vectors of features, with each (randomly generated) feature taking values 1-3. Traces in primary memory are subject to degradation through overwriting from subsequent items. This retroactive interference process is similarity-based: If feature f of item n is identical to feature f of item $n-m$, then this feature of item $n-m$ will be overwritten (set to 0) with probability $e^{-\lambda(m-1)}$. This form of overwriting allows retroactive interference to operate further back than just the most recently presented item, which seems appropriate for a few reasons. First, it seems reasonable to assume that the current encoding could disrupt more than just the immediately preceding trace. There are also empirical findings that support this idea. For example, the modality effect extends further than the final item in the lists (Saint-Aubin et al., 2021). Note, however, that the probability of overwriting remains highest for the immediately preceding item, a characteristic that leads to local distinctiveness effects: Features that are common between successive items are set to zero, meaning that the features that survive in item n-1 are maximally distinct from item n. In contrast, representations

in secondary memory are assumed to remain intact. After presentation of all items, a final overwriting of only modality-independent features takes place due to continuing internal thought activity in preparation for recall.

If overwriting degrades traces in primary memory, in the RFM we assume that a process of rehearsal can act to restore these overwritten features. Specifically, after every item presentation, there is a rehearsal which attempts to rehearse all previously presented items. This rehearsal cycle which runs after presentation of item n will successfully restore any overwritten feature with probability,

$$p = r \times e^{-\frac{(n-1)^2}{9}}$$

Where r is a parameter encoding the effectiveness of rehearsal, and the value of 9 in the exponent comes from previous work suggesting a significant drop in rehearsal for lists longer than four items (Bhatarah et al., 2009).

Order information is encoded in the same way as for the original FM: In particular, each presented item is tagged with its position in the list. This positional encoding is allowed to drift slightly according to a parameter θ which is set here to the default value from Neath and Surprenant (2007).

Thus far, we have described the encoding of presented items in a way which is identical to that of the RFM presented in Saint-Aubin et al. (2021), which was developed as a model of immediate serial recall. However, the experiments reported here used either a delayed reconstruction task or a free recall task, so we need to modify the retrieval process in the RFM slightly to account for this.

Accounting for the additional delay or filler task is simple – we will simply assume an extra overwriting process which affects all features between encoding and retrieval. Adapting the model to deal with different retrieval conditions is more challenging, and we turn to this below.

The Revised Feature Model for Order Reconstruction and Free Recall

The RFM was developed by Saint-Aubin et al. (2021) to account for the production effect in immediate serial recall. However, nothing in our description of the model thus far has concerned retrieval; the model as it stands is simply a way of encoding presented items. Saint-Aubin et al. (2021) went on to describe a mechanism for extracting this information at the point of recall to attempt a serial recall of the presented items. However, in the experiments reported above, the retrieval process was either delayed reconstruction or free recall, so the version of the RFM originally reported is not immediately applicable.

In the context of experiments like ours in which the recall task is only revealed at retrieval, we assumed that information is encoded similarly whether participants perform a free recall task or an order reconstruction task. Therefore, we suggest, a good model of memory ought to have encoding and retrieval elements that can be separated, and it should be possible to combine the same encoding stage with multiple retrieval processes to model different tasks. It is in this spirit that we proceed.

Before we explain the way that the model deals with order reconstruction and free recall, we briefly recap the way serial recall is handled in the RFM. This will serve to introduce some relevant concepts. After encoding is completed, we have a set of items in secondary memory consisting of vectors of features taking values 1-3, and a set of traces in primary memory which partially match the items, but where some of the features are missing (set to 0 by overwriting).

These traces will serve as cues to enable recall of the items, by a process in which the similarity between cue and stored item serves to activate that item. The exact details of this depend on the type of recall process (serial, free, or reconstruction) but the basic ingredients are the same across all recall types.

In serial recall, we begin with the item number to be recalled, i, and use that to retrieve the relevant degraded cue (i.e., representations that have been over-written). We then compute the similarities between this cue and all current list items stored in secondary memory. The similarities are related to the feature-to-feature correspondence between primary and secondary traces, via Shepard's Law (1987):

$$s(i,j) = e^{-d_{ij}}$$

This distance, d_{ij}, is taken to be a scaling constant, a, times the proportion of mismatching features between items i and j. The probability of retrieving item j given cue i is then given by,

$$p(j|i) = \frac{e^{\frac{s(i,j)}{\tau}}}{\sum_k e^{\frac{s(i,k)}{\tau}}}$$

This form for the probability is known as a soft-max function (technically a soft-argmax) and τ is a so-called 'temperature' parameter that controls how deterministically the item with the highest similarity is chosen. When the temperature tends to zero, then the item j with the highest similarity to cue i is chosen with probability 1; when the temperature is very large, all items have equal probability of being chosen, regardless of their similarity to the cue.

We also allow for the possibility that no secondary memory trace matches the primary memory trace well enough to be recalled. We do this by including an extra 'null'

possibility which has constant similarity between itself and all primary memory traces. For the set of features described here, this is approximately $s(null, j) = 1.7 \times 10^{-3}$. The process of recalling an item does not directly alter the cues, however the model includes a step where multiple recalls of the same item are suppressed by a factor e^{-cr} where r is the number of times an item has previously been recalled, and c is a constant, resulting instead in an omission. Full details are given in Saint-Aubin et al., (2021).

The retrieval process relevant for serial recall is adapted from the original Feature Model, and is not immediately relevant for order reconstruction or free recall. However, the basic idea, that similarity between cue and item is effectively the extent to which a cue 'activates' an item, and items with higher activations will tend to be chosen first, can be adapted to other types of retrieval.

We begin by considering order reconstruction. Reconstruction is, in many ways, the inverse of serial recall – in the RFM for serial recall, we start from the item number (or position) that we want to recall, pick the associated cue from primary memory, and then attempt to match that with the correct item from secondary memory. In reconstruction, we are presented with the list of items which, according to the RFM, ought to be stored in secondary memory, and our task is to match each item with the associated cue in primary memory, and from that extract the item position. The extent to which a cue activates an item is based on the similarity between feature vectors, which is symmetric, so the probability of remembering that the 4th item was *'chat'* (serial recall) is the same as the probability of remembering that the item *'chat'* was presented in position 4 (reconstruction). In a similar way to serial recall, a 'null' response is possible when an item fails to match any cues, in which case we make an omission, declining to give a label to that item.

Thus, it looks as though the retrieval probabilities for order reconstruction should be essentially identical to those for serial recall. However, there are two notable differences. First, because participants are given the list of items written on the screen, we assume that this triggers an extra rehearsal of modality-independent features. Second, the items that participants are attempting to order are stripped of any modality-dependent features they may have had if they were produced at the encoding stage. It would presumably be possible for a participant to match list items with the full item stored in secondary memory, including any modality-dependent features, but it seems likely this would require more effort and might even be unreliable. For that reason, we assume that the actual retrieval process for reconstruction is essentially identical to that for serial recall, except that a much smaller number of modality-dependent auditory features is accessible. For immediate reconstruction, we set this number to six, or a third of the possible features; for delayed reconstruction we assume that only two of the possible features are used.

We turn now to free recall. In the original RFM, for serial recall, the model works by picking a cue, computing the similarities between that cue and all items in secondary memory, which can be thought of as the extent to which that cue activates each secondary memory trace, and then picking an item with probability related to this activation. For free recall, instead of thinking of activations as being produced by a single cue, we can imagine that all the cues activate items stored in secondary memory in a way which depends on the similarities between cue and item. Each item is then activated by all cues and the item with the highest total activation will tend to be picked first. Specifically, each item i in secondary memory gets an activation,

$$p(i) = \frac{e^{\frac{s(i)}{\tau}}}{\sum_j e^{\frac{s(j)}{\tau}}}$$

Where $s(i) = \sum_j s(j,i)$ is the sum over the similarities between the item i and all cues. Once an item has been recalled, we set that activation to zero and choose from the remaining items. This is roughly equivalent to the suppression of multiple recalls of the same item which occurs in serial recall, but is computationally much less demanding than allowing the multiple recalls of the same item but then suppressing the output. Recall terminates either once all items have been recalled, or after an omission, described below.

As for serial recall and order reconstruction, we also have a 'null' option, which corresponds to a case where the summed similarities between cues and items is too low to activate any item. In this case, the model makes an omission, after which the retrieval process terminates. This null similarity is set to $s(null) = 1.36 \times 10^{-2}$, which is equivalent to the summed similarity between eight randomly chosen vectors of features (this was based on expectations from Experiment 1, where this 'null' possibility may be activated by each of the cues from the eight items). For Experiments 2 and 3, we left this constant unchanged for simplicity.

To summarize, there are two sources of difference between the original application of the RFM for immediate serial recall and the model as used here. To deal with the delay and filler task, we assume an extra overwriting step affecting all features. To model the different retrieval conditions, for order reconstruction we assume that presenting the items written on the screen triggers a rehearsal of modality independent features, followed by a retrieval process essentially identical to serial recall but making use of a smaller number of modality-dependent features. For free recall, we assume that activation is related to the summed similarities between all cues and an item in secondary memory. These changes, we argue, represent natural generalizations of the

basic retrieval process of the original and revised Feature Models, and allow the same basic model to account for all three retrieval types.

General Information about the Model Fitting

Model fitting for all experiments called upon Approximate Bayesian Computation (see Turner & Van Zandt, 2012, or Marin et al., 2012, for a review), using a version of sequential Monte Carlo sampling known as Partial Rejection Control (Sisson et al., 2007), hereafter referred to as ABC-PRC. Full details are given in the appendix and Code to fit the model can be found on the OSF page. Our general approach was to fit all data from a single experiment at once, which means fitting pure (aloud / silent) and mixed lists (ASAS / SASA), with both order reconstruction and free recall tasks, for a total of eight conditions per experiment for Experiments 1, 2, and 3. In addition, to provide a more stringent test of the RFM, we also modeled Experiment 1b of Saint-Aubin et al. (2021) which was an immediate order reconstruction task in which there were four conditions: pure (aloud / silent) and mixed lists (ASAS / SASA). To do this, we assumed that the encoding process did not depend on the retrieval task, and so these processes shared parameters. This is much more challenging for any model than allowing all parameters to vary between list or retrieval conditions. Model simulation was carried out using 1000 particles, and performed on City, University of London's SOLON cluster.

Although the model contains many possible parameters that could be varied, only a small number were allowed to vary in the model fitting. In particular, the number of possible feature values, the number of modality-dependent and independent features, and details of the recovery and perturbation parameters were fixed for all simulations. Values of all nonvarying parameters are given in the appendix. The parameters which we attempted to fit can be grouped into

encoding and retrieval parameters. The encoding parameters were λ, which controls how many items can be affected retroactively by overwriting, and $r_S, r_p,$ and r_m, which control how effectively rehearsal can restore overwritten features, in the silent, produced, and mixed list conditions. The retrieval parameters were the distance scaling parameters, $a_{p-R}, a_{m-R}, a_{p-F}$ and a_{m-F}, which control the overall ease of matching cues to items (one for each of pure list reconstruction, mixed list reconstruction, pure list free recall, and mixed list free recall), and finally $\tau_R,$ and $\tau_F,$ which control how deterministically the item with the highest activation is chosen in the order reconstruction and free recall tasks respectively. This gives ten parameters for each of Experiments 1, 2, and 3, four associated with encoding and six associated with retrieval. For Experiment 1b from Saint-Aubin et al. (2021), we only have a reconstruction condition, and so only seven parameters.

For all fits, we report the results by showing the means and 90% HDIs of the posteriors of the model predictions for each serial position in the different conditions, Figures 6 to 12. We report medians and 95% HDIs for the posteriors of the parameters allowed to vary in Table 2, and we also plot the posteriors for the parameters in Figure 13. Overall match between data and model seems very good. There is some misfitting in places, but the qualitative and quantitative agreement is excellent. Some of the disagreement is probably explained by the fact that the model exclusively attempts forward serial recall for order reconstruction, and recall without a privileged order for free recall. If some participants use other strategies, for example attempting serial recall in the free recall condition, or choosing to start with the final items in the reconstruction condition, this will tend to resulting in some systematic misfitting for the initial or final items.

Overall, the fits provide good evidence that the RFM, incorporating the basic ingredients of relative distinctiveness, similarity-based overwriting, and rehearsal, can account as well for the production effect in long-term memory as it does for the same effect in immediate serial recall and immediate order reconstruction. It also confirms the basic idea that we can model the encoding process independently from the retrieval process for these tasks.

We are not engaging here in a formal model comparison exercise. Instead, our aim is to show that a model with the key components of relative distinctiveness and rehearsal can account for the data in these experiments. However, we can increase our confidence in our explanations for the data by examining how our model behaves when we impose certain constraints on the parameters. For example, since our proposed explanation for the distinctive sawtooth pattern seen in mixed lists involves relative distinctiveness, we can attempt to fit a version of the model where we fix the number of features to be the same for produced and silently read items, thus removing this property. In the Appendix we show the results of fitting three different alternative versions of the model to the data from Experiment 1. The first alternative model fixes the number of features for produced items to be equal, the second sets all three rehearsal parameters to zero, and the third approximates the n-1 overwriting behavior of the original Feature Model by setting [lambda]=10. The patterns of misfitting produced by each of these reduced models shed light on the role of the different features of the RFM. As expected, a model without relative distinctiveness cannot account for the sawtooth pattern seen in mixed lists (a similar result was noted by Saint-Aubin et al. (2021) for immediate serial recall). A model with no rehearsal cannot reproduce the higher recall probabilities for the first few items—again, echoing a finding reported by Saint-Aubin et al. (2021). Finally, fixing overwriting to only occur for the previously presented item (as in the original Feature Model) gives rise to a number of odd effects, most

notable of which is that, since any activity after list presentation can now only overwrite features from the final item, recall is worse for the final item compared with the next to last item, the opposite of what we see in the data.

In summary, the RFM, with some sensible adjustments to handle order reconstruction and free recall, can account for the patterns of data reported from Experiments 1, 2, and 3, as well as data previously reported in Experiment 1b in Saint-Aubin et al. (2021). Fits were done assuming the same encoding processes and parameters at play in both retrieval tasks, meaning the model suggests that the slightly different production effects seen in order reconstruction and free recall are the result of the same basic encoding processes, depending on relative distinctiveness, similarity-based overwriting, and rehearsal. Removing any of relative distinctiveness, rehearsal, or longer horizon overwriting leads to a model unable to reproduce qualitative features of the data. Finally, the encoding elements of this model are exactly the same as the model used for immediate serial recall in Saint-Aubin et al. (2021), suggesting some commonality between immediate and long-term recall, at least with respect to this effect.

General discussion

Almost all of the research on the production effect has relied on LTM tasks such as free recall and item recognition. It can be argued that the importance of the production effect rests on its status as a straightforward example of relative distinctiveness (MacLeod & Bodner, 2017). Because the latter can be considered as a general memory principle (Surprenant & Neath, 2009), Saint-Aubin et al. (2021) examined the production effect in STM tasks. In addition to replicating the basic effects, they reported significant interactions with serial positions which shed light on the mechanisms underpinning the production effect. By relying on a revised Feature Model (Nairne, 1988, 1990; Neath & Nairne, 1995; Neath & Surprenant, 2007), they were able to

account for the entire pattern of results. With this in mind, the current study was aimed at investigating whether these interactions with serial positions can also be observed in LTM and whether the RFM can also account for the production effect in LTM tasks. Under the assumption that the same principles would operate across memory tasks, we hypothesized that this would be the case (see, e.g., Crowder, 1993; Neath & Saint-Aubin, 2011; Neath et al., 2019).

Here, using the design of Jonker et al. (2014), Experiment 1 examined the production effect in a delayed free recall task. In the second experiment, we investigated the production effect with much longer lists (e.g., 24 items) more typical of those used in free recall. Finally, Experiment 3 extended the duration of the filed interval to 2 minutes and called upon 10-item lists. Results of the parity judgment task indicate that the participants were actively engaged in the distractor activity during the interval and by doing so confirm that our memory tasks called upon LTM. In our work, there was one critical adjustment: For mixed lists, produced and silently read words were systematically alternated instead of being randomly mixed. Alternating the items maximized the contrast between items read aloud versus silently and made it much easier to consider what the serial position pattern would reveal about the production effect. When the results for mixed lists are considered overall, across all experiments, the findings show that produced items are better remembered; this suggests that adding relevant dimensions to the studied items (e.g., articulatory programming, motor output, auditory feedback) improves performance, leading to the observed sawtooth patterns. The favored interpretation of these findings is that reading a visually presented item aloud generates more useful modality-related features than does silent reading (Saint-Aubin et al., 2021).

The second pattern that mimics STM findings relates to the pure list conditions. Although the results were perhaps slightly more systematic with shorter lists, overall, any differences in the

first serial positions favored the silently read items whereas the reverse was true for the later serial positions (see, e.g., Crowder, 1970; Grenfell-Essam et al., 2017; Greene & Crowder, 1984; Kappel et al., 1973; Macken et al., 2016; Saint-Aubin et al., 2021). The latter finding was interpreted by Saint-Aubin et al. (2021) as a modality effect and such an interpretation also fits well here. This effect relates to the finding that items that are presented aurally are better remembered than items presented silently for the last or last few studied items. We will return to this when discussing the RFM below. With respect to the advantage found for silent items in the earlier positions, Saint-Aubin et al. (2021) offered that early items are typically rehearsed more (Bhatarah et al., 2009; Rundus, 1971; Tan & Ward, 2000; Ward, 2002) and that producing the items would hinder the covert rehearsal process. Because early items are typically rehearsed more, the negative effect of production on rehearsal will be felt more acutely for items in the first positions.

With the free recall task, in all experiments, we observed the predicted pattern of results, albeit with more noise in Experiment 2, where participants struggled in particular with the LTM order reconstruction task. This pattern is in line with results of Mulligan and Lozito (2007) who found a sizeable relation between order information and free recall with 8-item lists, but not with 24-item lists. That said, overall, the results of all experiments nicely extend those observed in immediate serial recall.

The findings reported here are well aligned with the expectations derived from the work of Saint-Aubin et al. (2021) and the interpretation put forward by those authors seems appropriate for the current work. In essence, they suggested that, relative to silently reading items, producing items generates useful, distinctive features. These extra features are thought to bring about an advantage at the point of retrieval – the more distinctive items, possessing extra,

relatively item-specific features, would be easier to retrieve against the backdrop of silent items that do not benefit from said production-dependent features. This relative distinctiveness advantage comes at a cost, however, as production is thought to interfere with covert rehearsal (see also, Macken et al., 2016).

We will turn to the more specific delineation of these mechanisms suggested by the RFM. Before however, it seems important to point out that Saint-Aubin et al. (2021) were not the first to suggest a distinctiveness account of the production effect. MacLeod and Bodner (2017) summarized the distinctiveness view of the production effect as follows: "The idea is that producing items increases their distinctiveness in memory relative to unproduced items. The processing operations applied during a production task constitute part of the encoding for items (Conway & Gathercole, 1987), and at the time of test, the distinctiveness of these operations can facilitate access to produced items relative to unproduced items" (MacLeod & Bodner, 2017, p. 392). MacLeod and Bodner argued, as we have here, that producing items entails distinctive processing that enhances retrieval of the produced items. However, they also pointed out that for free recall, the production effect seems mostly attributable to an increased cost to silent items when going from pure to mixed lists – rather than to an increased benefit to produced items when going from pure to mixed lists. They suggested that this poses a problem for a distinctiveness account of the production effect, as the latter predicts a benefit for produced items in mixed lists, through enhanced distinctiveness processing, rather than a cost to silent ones. We examined this for our free recall findings without considering serial position. In all experiments, as previously found, silently read items were less well recalled in mixed than in pure lists. However, we also observed the predicted benefit for produced items that were better recalled in mixed than in pure lists. How does one reconcile the idea of a distinctiveness processing advantage with a drop in

performance for silently read items? If one considers that there is a cost to said distinctive processing, then the drop in performance for silent items when going from pure to mixed lists is not as surprising.

As mentioned above, in the data reported here, we have better performance for produced items and a drop for silently read words when going from pure to mixed lists. The RFM can account for both effects because of the similarity-based retroactive interference central to encoding in the model – hence, our formal definition of distinctiveness also predicts a benefit for produced words and a cost for silent items, with alternating lists.

Our implementation of the relative distinctiveness principle in the RFM and our discussion of its role in the production effect is consistent with ideas developed in other models. More specifically, there has been one other attempt to account for the production effect with a formal model of memory. Using MINERVA 2, Jamieson et al. (2016) modeled the production effect in *recognition* by adding sensory feedback features to the vector representing a produced item whereas they did not add any to the vector representing a silently read item. This is akin to the addition of modality-dependent features to the vector representing a produced item in the RFM. Furthermore, in MINERVA 2, retrieval is assumed to be cue-driven, with each trace activated in proportion to its similarity to the retrieval cue. This retrieval process also shares many similarities with the RFM. However, contrary to the RFM, MINERVA 2 cannot account for serial positions, for the interaction between production and serial positions, or for the cost of producing the items. As Jamieson et al. noted, MINERVA 2 was developed to account for item recognition and has only been applied to the production effect in this context. In contrast, the Feature Model was developed to account for recall data and the RFM has been expanded to also account for order reconstruction. Therefore, it is difficult to provide a complete comparative

analysis of both models in the context of the production effect, but it is encouraging that the two formal models applied to the production effect called upon the distinctiveness principle.

Here, the RFM, originally developed to account for immediate serial recall performance, was able to account for delayed order reconstruction and delayed free recall with short and long lists; this was possible with no change to the model's basic assumptions and operation. The main assumptions are as follows. Produced items generate features that are absent when reading items silently and these features can support retrieval. In essence, produced items are more 'richly' encoded. Also, in mixed lists, there is a local distinctiveness effect related to these extra features; retroactive interference affects all items, but the distinctive, modality-dependent features associated with produced items lead to a relative advantage that is maximal in alternating, mixed lists. Finally, the richer encoding of produced items comes at a cost, in that saying items aloud may interfere with covert rehearsal. These three mechanisms account for the findings in all the experiments reporeted here, as well as in the six experiments reported by Saint-Aubin et al. (2021).

A few further points are worth noting. First, the RFM can account for order reconstruction and free recall with few adjustments to the retrieval processes while keeping encoding operations the same. Second, the only adjustment needed to account for delayed performance relative to immediate recall was the addition of one overwriting cycle. The latter takes the distractor task occupying the delay into account. In effect, with this simple adjustment, the RFM was able to account for both immediate order reconstruction (Experiment 1b of Saint-Aubin et al., 2021) and delayed order reconstruction with short lists (Experiments 1 and 3), long lists (Experiment 2), and short delays (Experiment 1 and 2) and long delays (Experiment 3), as well as the free recall data from all experiments, while keeping all other parameters the same. Third, the RFM is based

on the Feature Model; as mentioned, the latter was developed to account for immediate serial recall performance (Nairne, 1990). It is remarkable that the RFM has been able to fit the serial position curves and their interaction with list type (pure and mixed) and condition (aloud and silent) for both the order reconstruction and the free recall tasks. For the order reconstruction task, it was assumed that, contrary to recall, participants compare partial retrieval cues with the represented items, rather than the representation in secondary memory. Therefore, fewer modality-dependent features are involved which resulted in an attenuated production effect. For free recall, instead of using a single cue, it is assumed that all cues activate items in secondary memory and the most activated item is output first. These demonstrations are important as they show that a single model can account for performance in a variety of paradigms with different time frames and list lengths.

Basically, the same assumptions and modelled processes with very few adjustments, were able to account for both the data in a classic STM task and performance in classic LTM settings. Moreover, the same assumptions and model were able to account for what appear to be very different and quite complex patterns of performance. The latter points highlight one of the advantages of having a precise model of the mechanisms that are assumed to generate responses. One of the implications is that the same mechanisms can generate quite different patterns depending on task difficulty, or at least list length. In essence, the work presented here suggests that the same processes underpin the production effect in short-term and long-term memory. This makes sense if one assumes that the production effect is driven by a basic, domain-general, relative distinctiveness process and that the latter applies across time scales and memory systems.

Concluding remarks

Our discipline has often been criticized for being overly empirical, with a collection of apparently unrelated effects, while lacking a coherent and general explanation of behavior (see, e.g., Crowder, 1993; Jamieson et al., 2016; Surprenant & Neath, 2009). Here, we hoped to address this criticism by establishing that the complex patterns of findings related to the production effect within STM paradigms were also observed in LTM settings. The same basic relative distinctiveness processes can account for both sets of findings. Furthermore, we showed that the RFM, originally developed to account for short-term recall, can handle recall and order reconstruction data from both STM and LTM tasks. Due to its computational nature, the RFM further offers new and promising avenues for understanding the role of distinctiveness in memory.

References

Bhatarah, P., Ward, G., Smith, J., Hayes, L. (2009). Examining the relationship between free recall and immediate serial recall: Similar patterns of rehearsal and similar effects of word length, presentation rate, and articulatory suppression. *Memory & Cognition, 37*, 689-713. https://doi.org/10.3758/MC.37.5.689

Bhatarah, P., Ward, G., & Tan, L. (2008). Examining the relationship between free recall and immediate serial recall: The serial nature of recall and the effect of test expectancy. *Memory & Cognition, 36(1)*, 20–34. https://doi.org/10.3758/MC.36.1.20

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review, 114*, 539-576. https://doi.org/10.1037/0033-295X.114.3.539

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

Cowan, N. (2017a). Working memory: The information you are now thinking of. In. Wixted, J. T. (Ed.), *Cognitive Psychology of Memory, Vol. 2 of Learning and Memory: A comprehensive reference, 2nd edition,* Byrne, J.H. (Ed.). pp. 147-161. Oxford: Academic Press

Cowan, N. (2017b), The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review, 24*, 1158-1170. https://doi.org/10.3758/s13423-016-1191-6

Crowder, R. G. (1970). The role of one's own voice in immediate memory. *Cognitive Psychology, 1*, 157–178. https://doi.org/10.1016/0010-0285(70)90011-3

Crowder, R. G. (1976). *Principles of learning and memory*. Lawrence Erlbaum.

Crowder, R. G. (1993). Short-term memory: Where do we stand? *Memory & Cognition, 21*(2), 142–145. https://doi.org/10.3758/BF03202725

Fawcett, J. M. (2013). The production effect benefits performance in between-subject designs: A meta-analysis. *Acta Psychologica, 142*, 1–5. https://doi.org/10.1016/j.actpsy.2012.10.001

Forrin, N. D., & MacLeod, C. M. (2016). Order information is used to guide recall of long lists: Further evidence for the item-order account. *Canadian Journal of Experimental Psychology, 70*, 125–138. https://doi.org/10.1037/cep0000088.supp (Supplemental)

Greene, R. L., & Crowder, R. G. (1984). Modality and suffix effects in the absence of auditory stimulation. *Journal of Verbal Learning & Verbal Behavior, 23*, 371–382. https://doi.org/10.1016/S0022-5371(84)90259-7

Grenfell-Essam, R., & Ward, G. (2012). Examining the relationship between free recall and immediate serial recall: The role of list length, strategy use, and test expectancy. *Journal of Memory and Language, 67*(1), 106–148. https://doi.org/10.1016/j.jml.2012.04.004

Grenfell-Essam, R., Ward, G., & Tan, L. (2017). Common modality effects in immediate free recall and immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(12), 1909–1933. https://doi.org/10.1037/xlm0000430.supp (Supplemental)

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In R. R. Hunt, & J. B. Worthen (Eds.), *Distinctiveness and memory*. (pp. 3–25). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195169669.003.0001

Icht, M., Mama, Y., & Algom, D. (2014). The production effect in memory: Multiple species of distinctiveness. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.00886

Jamieson, R. K., Mewhort, D. J. K., & Hockley, W. E. (2016). A computational account of the production effect: Still playing twenty questions with nature. *Canadian Journal of Experimental Psychology*, *70*(2), 154–164. https://doi.org/10.1037/cep0000081

Jones, A. C., & Pyc, M. A. (2014). The production effect: Costs and benefits in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 300–305. https://doi.org/10.1037/a0033337

Jonker, T. R., Levene, M., & MacLeod, C. M. (2014). Testing the item-order account of design effects using the production effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 441–448. https://doi.org/10.1037/a0034977

Kappel, S., Harford, M., Burns, V. D., & Anderson, N. S. (1973). Effects of vocalization on short-term memory for words. *Journal of Experimental Psychology, 101*, 314–317. https://doi.org/10.1037/h0035247

Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science, 4(1)*. https://doi.org/10.1177/2515245920951503

Lambert, A. M., Bodner, G. E., & Taikh, A. (2016). The production effect in long-list recall: In no particular order? *Canadian Journal of Experimental Psychology, 70*, 165–176. https://doi.org/10.1037/cep0000086

Macken, B., Taylor, J. C., Kozlov, M. D., Hughes, R. W., & Jones, D. M. (2016). Memory as embodiment: The case of modality and serial short-term memory. *Cognition, 155*, 113–124. https://doi.org/10.1016/j.cognition.2016.06.013

MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science, 26*, 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

Marin, J.-M., Pudlom, P., Robert, C. P., & Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing, 22*, 1167–1180. https://doi.org/10.1007/s11222-011-9288-2

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial in Quantitative Methods for Psychology, 4*, 61-64.

Mulligan, N. W., & Lozito, J. P. (2007). Order information and free recall: Evaluating the item-order hypothesis. *The Quarterly Journal of Experimental Psychology*, *60*(5), 732–751. https://doi.org/10.1080/17470210600785141

Nairne, J. S. (1988). A framework for interpreting recency effects in immediate serial recall. *Memory & Cognition*, *16*(4), 343-352. https://doi.org/10.3758/BF03197045

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, *18*(3), 251-269. https://doi.org/10.3758/BF03213879

Neath, I. (1997). Modality, concreteness, and set-size effects in a free reconstruction of order task. *Memory & Cognition*, *25*, 256–263. https://doi.org/10.3758/BF03201116

Neath, I. (1999). Modelling the disruptive effects of irrelevant speech on order information.

International Journal of Psychology, *34*(5-6), 410-418. https://doi.org/10.1080/002075999399765

Neath, I., & Nairne, J. S. (1995). Word-length effects in immediate memory: Overwriting trace decay theory. *Psychonomic Bulletin & Review*, *2*(4), 429-441. https://doi.org/10.3758/BF03210981

Neath, I., & Saint-Aubin, J. (2011). Further evidence that similar principles govern recall from episodic and semantic memory: The Canadian prime ministerial serial position function. *Canadian Journal of Experimental Psychology, 65*(2), 77–83. https://doi.org/10.1037/a0021998

Neath, I., Saint-Aubin, J., Bireta, T. J., Gabel, A. J., Hudson, C. G., & Surprenant, A. M. (2019). Short- and long-term memory tasks predict working memory performance, and vice versa. *Canadian Journal of Experimental Psychology, 73*(2), 79–93. https://doi.org/10.1037/cep0000157

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 516-524. https://doi.org/10.3758/BF03195598

Norris, D. (2017). Short-term memory and long-term memory are still different. *Psychological Bulletin, 143*(9), 992–1009. https://doi.org/10.1037/bul0000108

Norris, D. (2019). Even an activated long-term memory system still needs a separate short-term store: A reply to Cowan (2019). *Psychological Bulletin*, *145*(8), 848–853. https://doi.org/10.1037/bul0000204

Penney, C. G. (1989). Modality effects and the structure of short-term verbal memory. *Memory & Cognition*, *17*(4), 398-422. https://doi.org/10.3758/BF03202613

Poirier, M., Yearsley, J.M., Saint-Aubin, J., Fortin, C., Gallant, G. and Guitard, D. (2019). Dissociating visuo-spatial and verbal working memory: It's all in the features. *Memory & Cognition, 47(4)*, 603–618. doi:10.3758/s13421-018-0882-9.

Psychology Software Tools. (2016). *E-Prime* (Version 3.0) [Computer software].

Routh, D. A. (1970). "Trace strength," modality, and the serial position curve in immediate memory. *Psychonomic Science, 18*, 355–357. https://doi.org/10.3758/BF03332397

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology*, *89*(1), 63–77. https://doi.org/10.1037/h0031185

Saint-Aubin, J., Yearsley, J., Poirier, M., Cyr, V., & Guitard, D. (2021). A model of the production effect over the short-term: The cost of relative distinctiveness. *Journal of Memory and Language*, *118*. https://doi.org/10.1016/j.jml.2021.104219

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science, 237(4820)*, 1317–1323. http://dx.doi.org/10.1126/science.3629243

Sisson, S. A., Fan, F., & Tanaka, M. A. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, *104*(6), 1760–1765. https://doi.org/10.1073/pnas.0607208104

Sisson, S.A., Fan, Y., & Tanaka, M.M. (2009). Correction for Sisson et al., Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 16889.

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods, 42*, 1096–1104. http://dx.doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolKit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology, 44*, 24–31. http://dx.doi.org/10.1177/0098628316677643

Surprenant, A. M., & Neath, I. (2009). *Principles of memory*. Psychology Press.

Tan, L., & Ward, G. (2000). A recency-based account of the primacy effect in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(6), 1589–1625. https://doi.org/10.1037/0278-7393.26.6.1589

Turner, B. M., & Van Zandt, T. (2012). A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology, 56*, 69–85. https://doi.org/10.1016/j.jmp.2012.02.005

Voyer, D., Saint-Aubin, J., Altman, K., & Gallant, G. (2021). Sex differences in verbal working memory: A systematic review and meta-analysis. *Psychological Bulletin, 147(4)*, 352–398. https://doi.org/10.1037/bul0000320

Ward, G. (2002). A recency-based account of the list length effect in free recall. *Memory & Cognition, 30*(6), 885–892. https://doi.org/10.3758/BF03195774

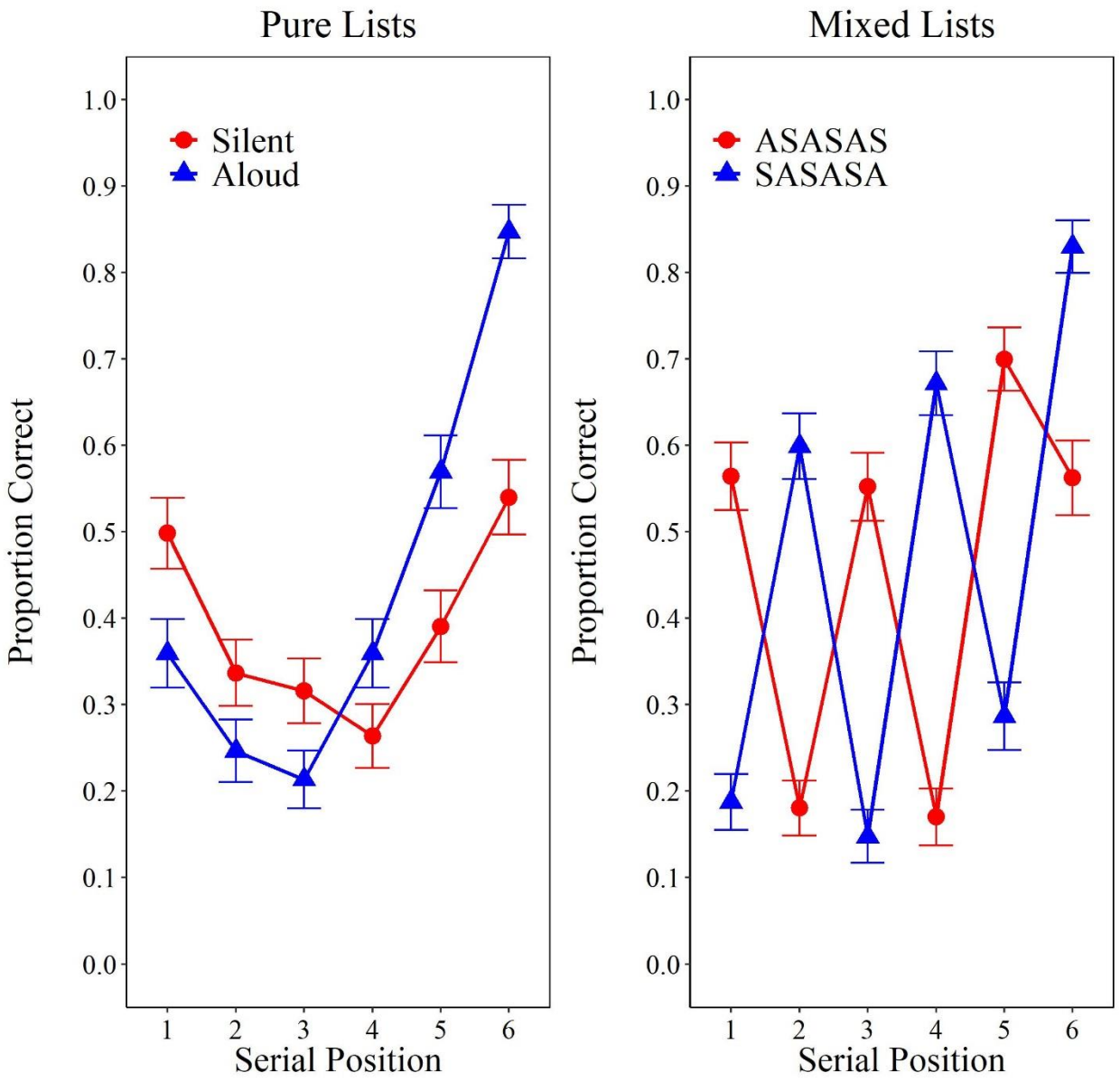

Figure 1. Proportion of correct recall in Experiment 1a of Saint-Aubin et al. (2021) with an immediate serial recall task. In the legend, the letter A indicates that the word holding that serial position was read aloud, while the letter S indicates that the word was read silently. Error bars represent confidence intervals at 95% for the repeated measures factor computed after Morey's (2008) method. The figure has been redrawn from the original data.

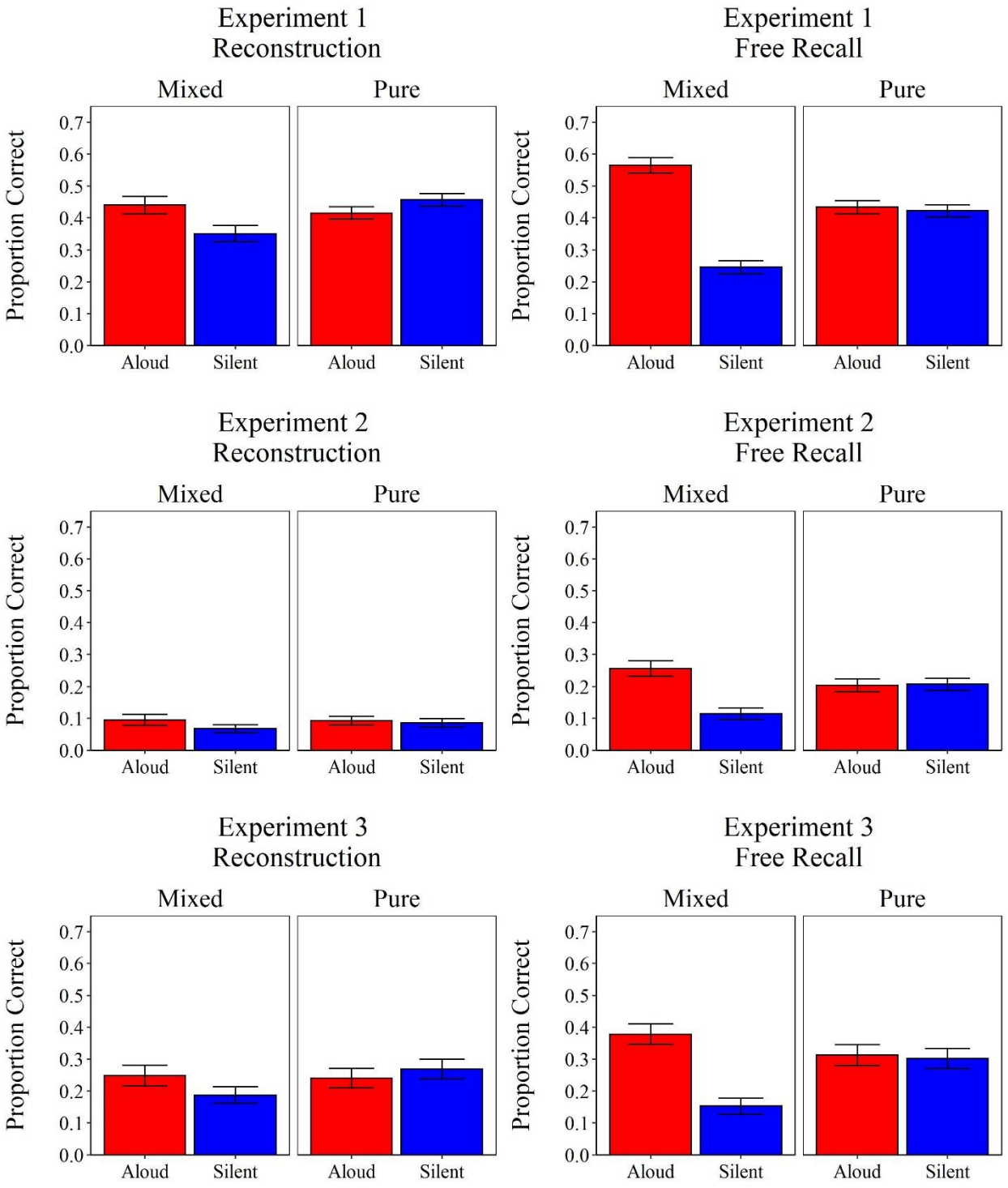

Figure 2. Proportion of correct reordering with an order reconstruction task (left column) and of correct recall with a free recall task (right column) as a function of list type, presentation modality (words read aloud or silently), and experiment. Error bars represent confidence intervals at 95% for the repeated measures factor computed after Morey's (2008) method.

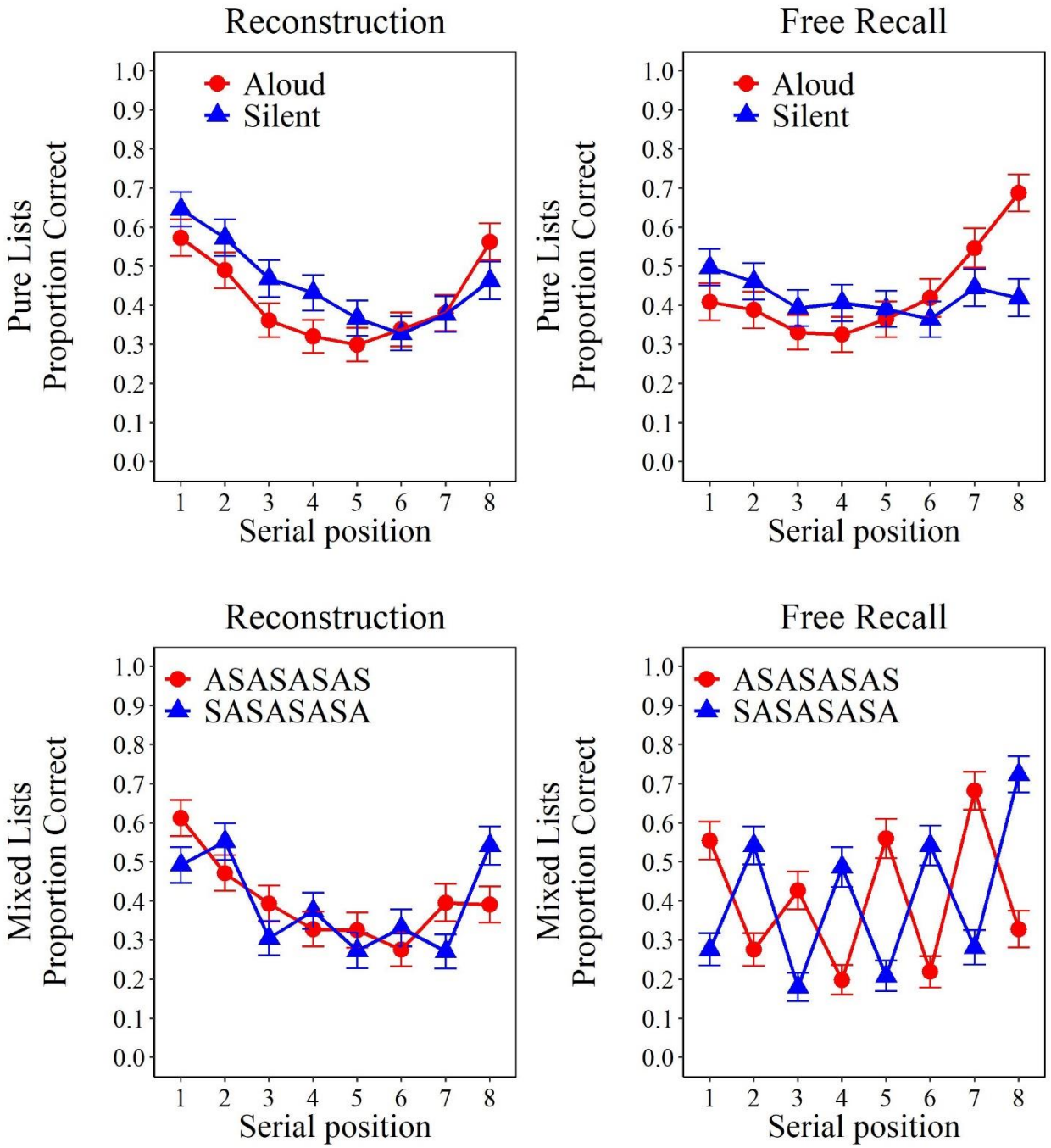

Figure 3. Proportion of correct reordering with an order reconstruction task and of correct recall with a free recall task in Experiment 1 as a function of serial position and input modality for mixed and pure lists. In the legend, the letter A indicates that the word holding that serial position was read aloud, while the letter S indicates that the word was read silently. Error bars represent confidence intervals at 95% for the repeated measures factor computed after Morey's (2008) method.

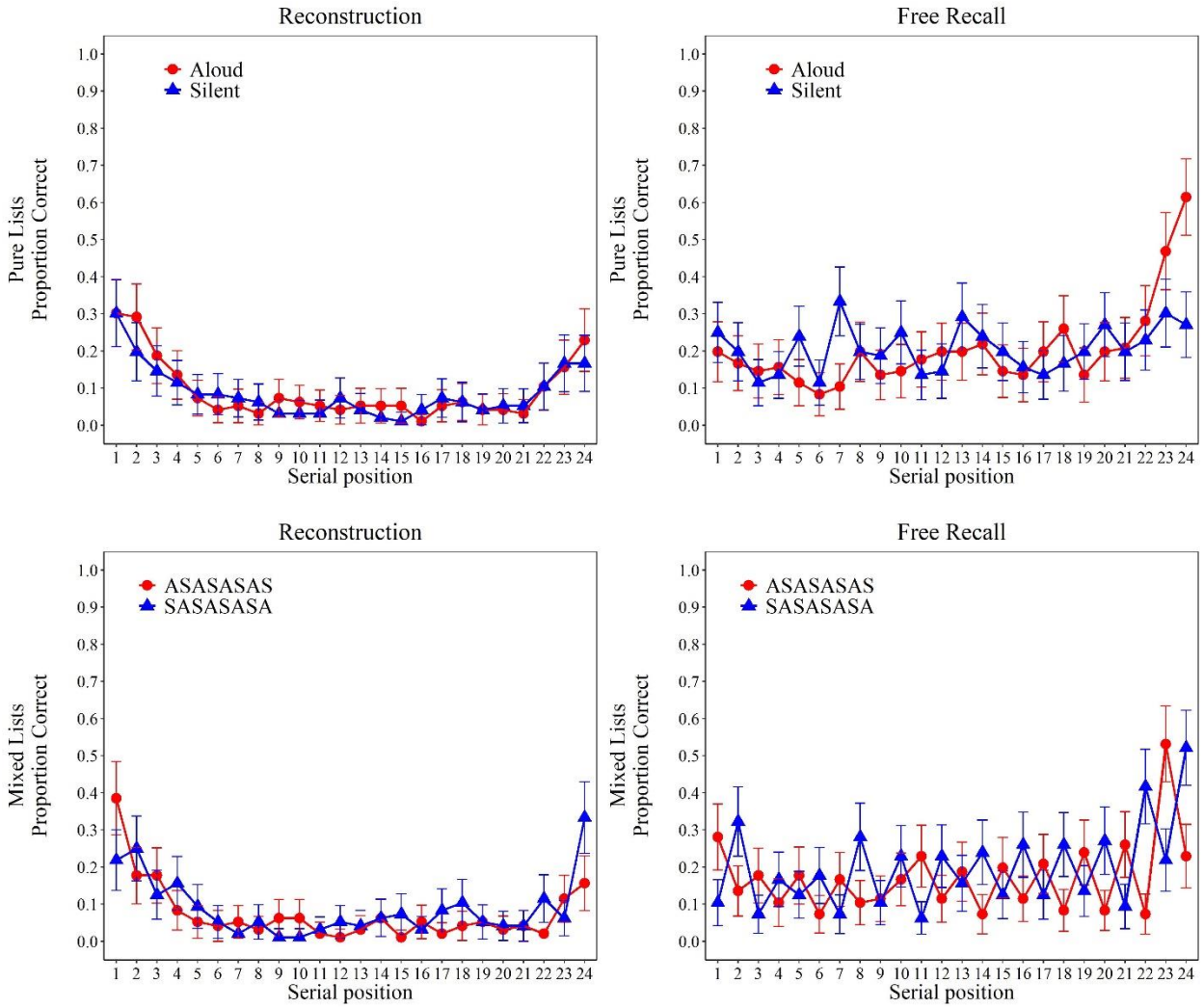

Figure 4. Proportion of correct reordering with an order reconstruction task and of correct recall with a free recall task in Experiment 2 as a function of serial position and input modality for mixed and pure lists. In the legend, the letter A indicates that the word holding that serial position was read aloud, while the letter S indicates that the word was read silently. Error bars represent confidence intervals at 95% for the repeated measures factor computed after Morey's (2008) method.

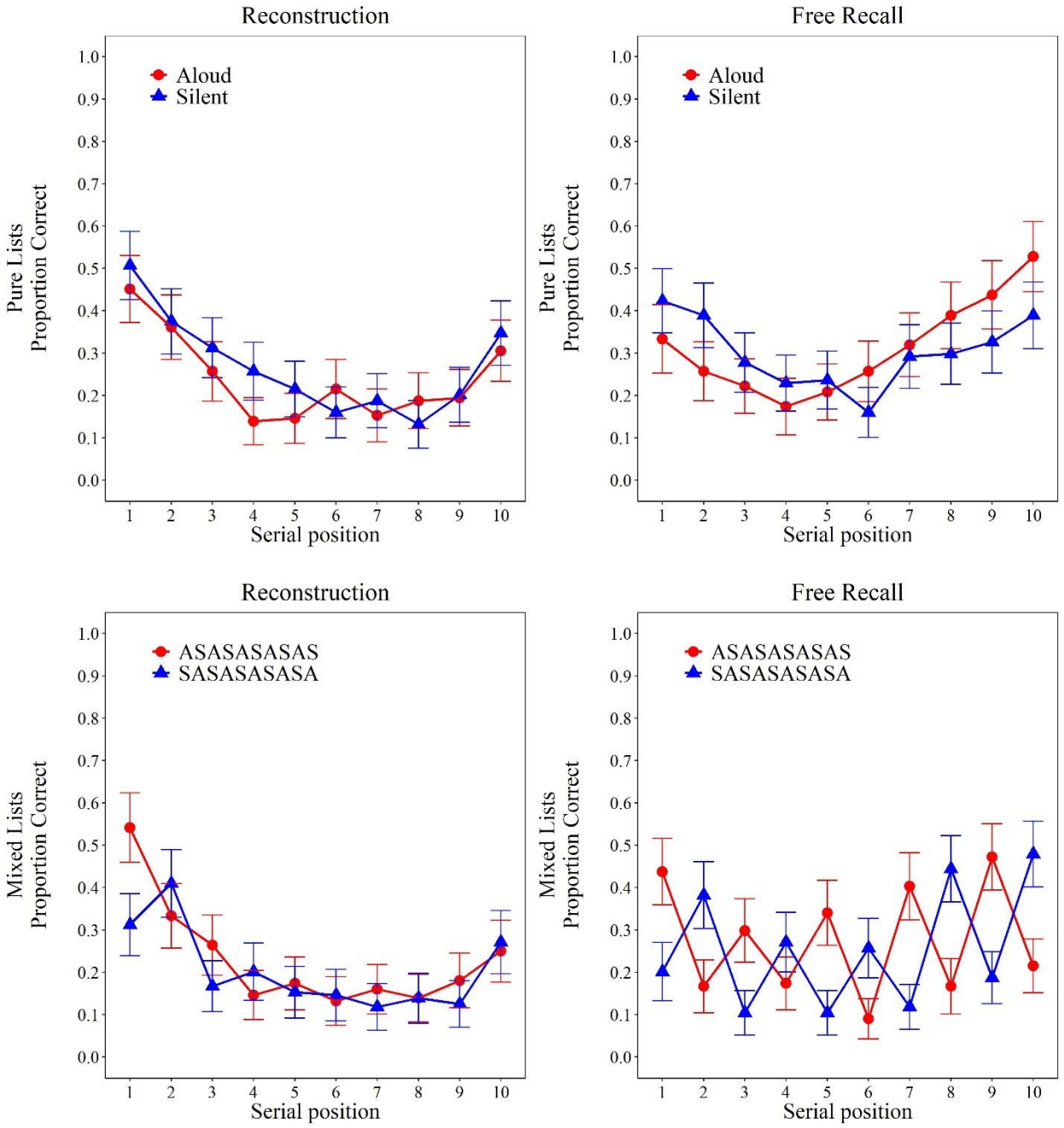

Figure 5. Proportion of correct reordering with an order reconstruction task and of correct recall with a free recall task in Experiment 3 as a function of serial position and input modality for mixed and pure lists. In the legend, the letter A indicates that the word holding that serial position was read aloud, while the letter S indicates that the word was read silently. Error bars represent confidence intervals at 95% for the repeated measures factor computed after Morey's (2008) method.

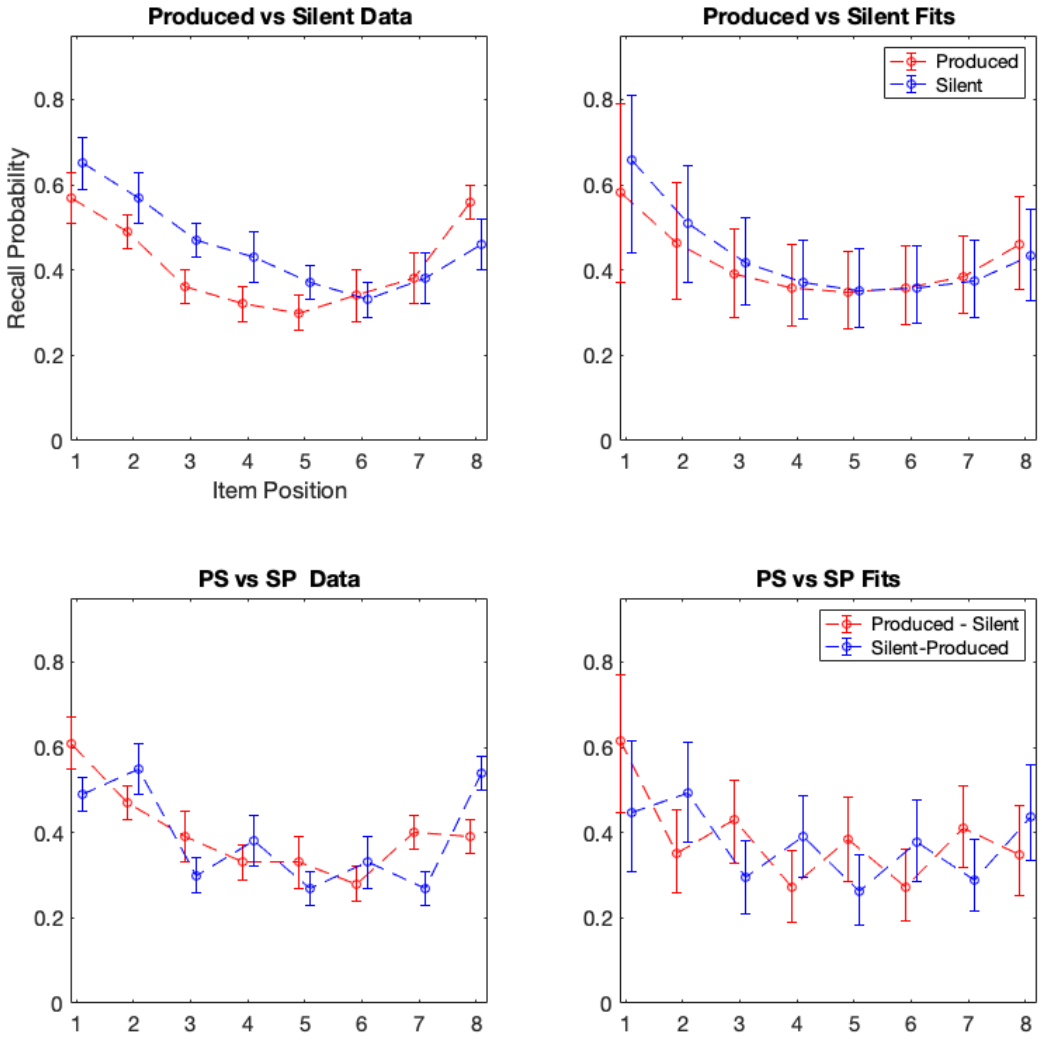

Figure 6. Data and model fits for the order reconstruction task used in Experiment 1. For the model fits, error bars are 95% HDIs of the posterior predictions. Although there is some misfitting in places, on the whole the qualitative and quantitative match between model and data seems good.

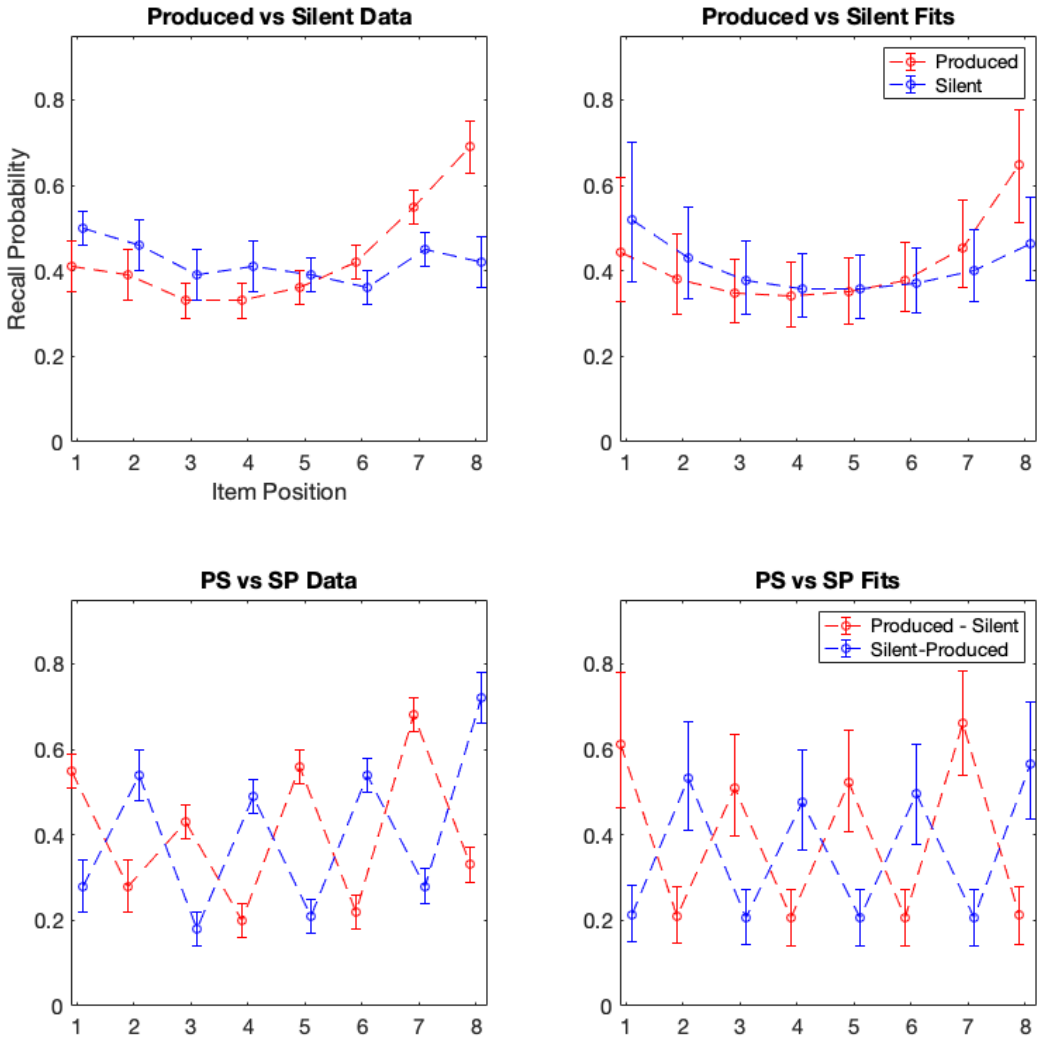

Figure 7. Data and model fits for the free recall task used in Experiment 1. For the model fits, error bars are 95% HDIs of the posterior predictions. Although there is some misfitting in places, on the whole the qualitative and quantitative match between model and data seems good.

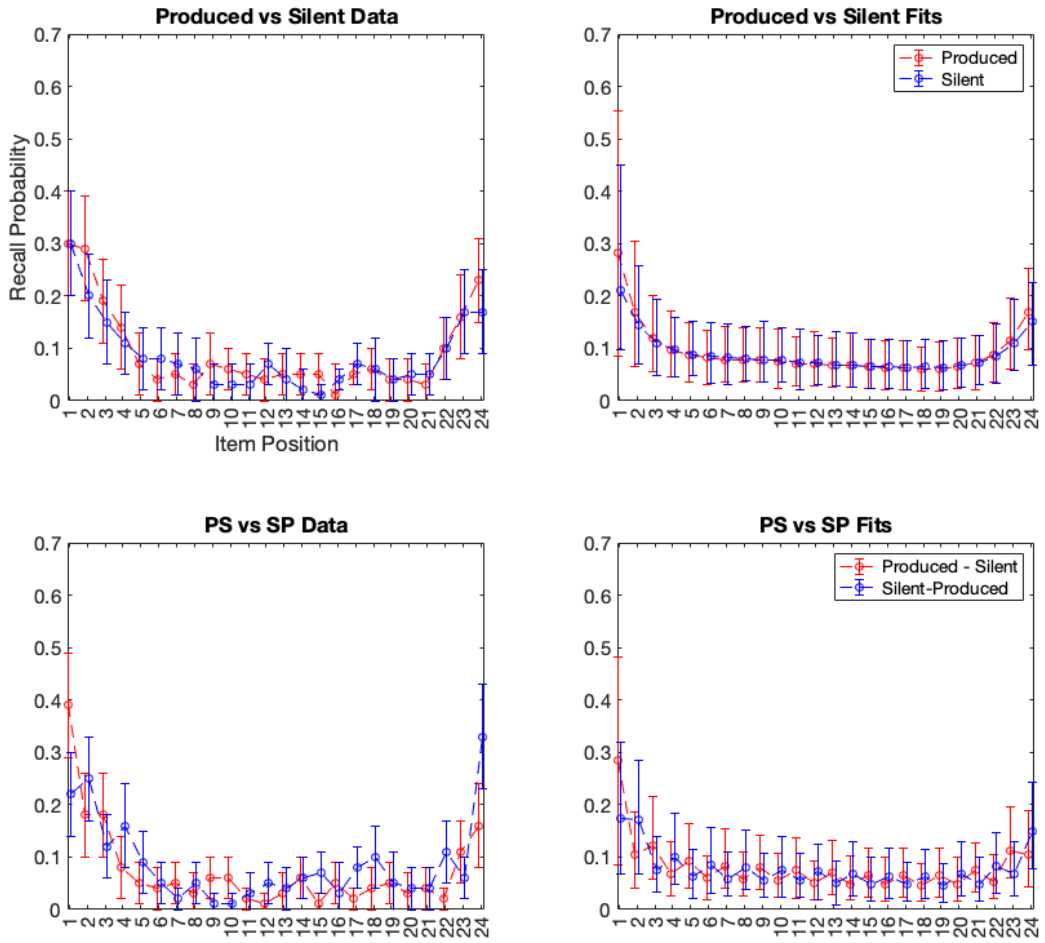

Figure 8. Data and model fits for the order reconstruction task used in Experiment 2. For the model fits, error bars are 95% HDIs of the posterior predictions. Although there is some misfitting in places, overall the match between data and model is good. Error bars for order reconstruction are somewhat larger here, which does seem to match the greater variability seen in this data.

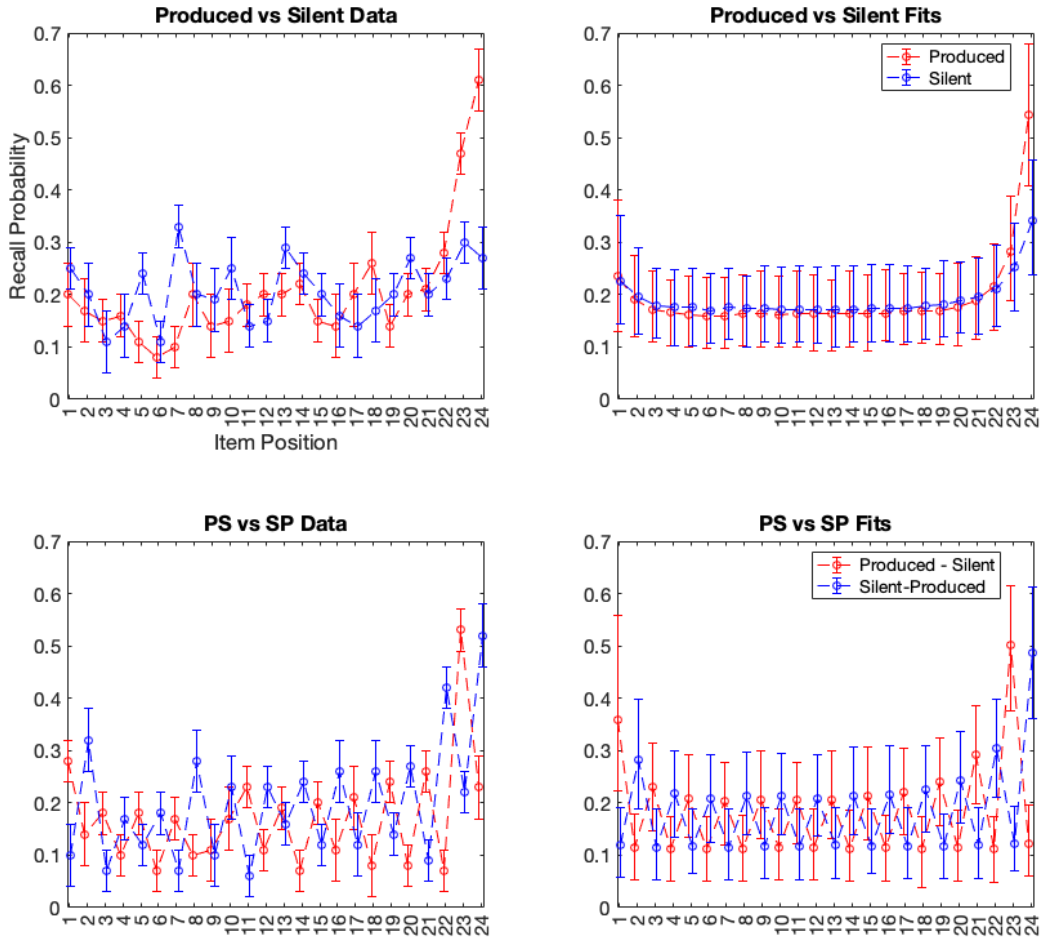

Figure 9. Data and model fits for the free recall task used in Experiment 2. For the model fits, error bars are 95% HDIs of the posterior predictions. Although there is some misfitting in places, overall the match between data and model is good. Error bars for Free Recall are somewhat larger here, which does seem to match the greater variability seen in this data.

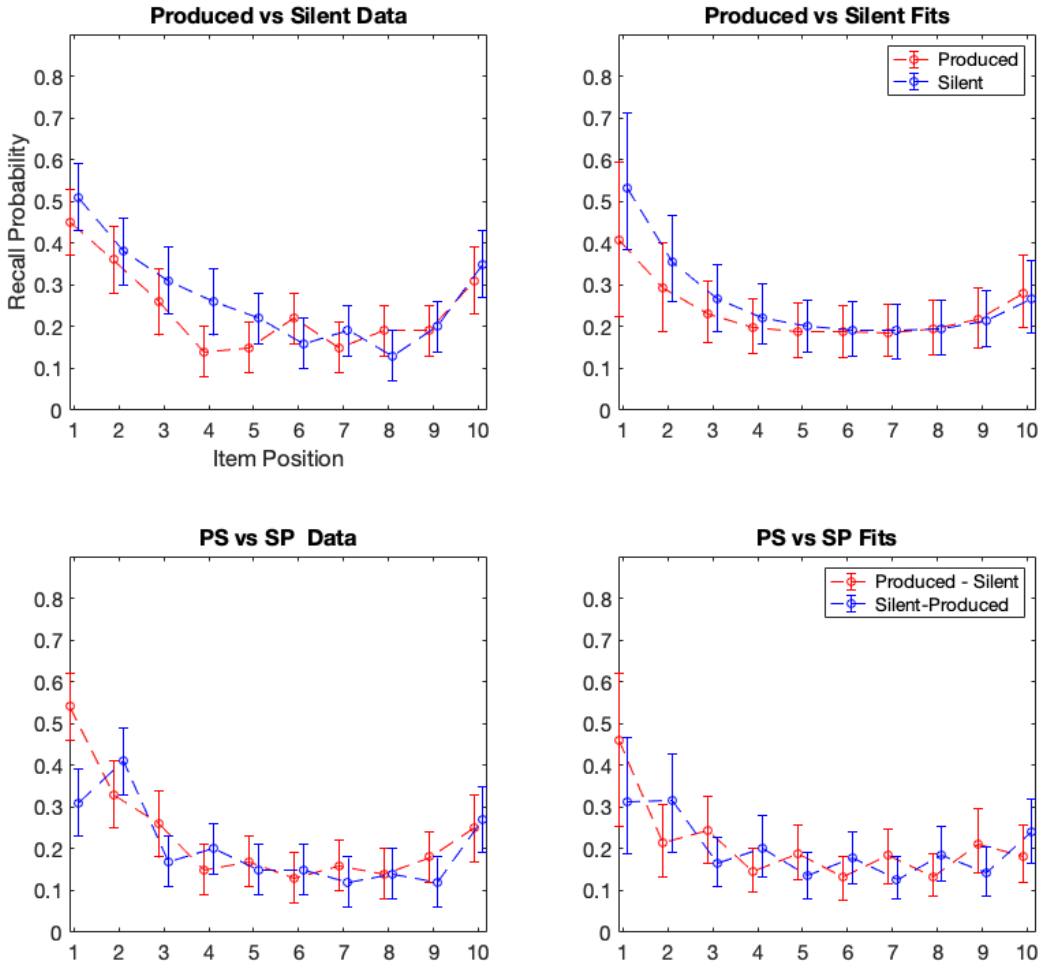

Figure 10. Data and model fits for the order reconstruction task used in Experiment 3. For the model fits, error bars are 95% HDIs of the posterior predictions. As for Experiments 1 and 2, the qualitative and quantitative match between model and data is very good.

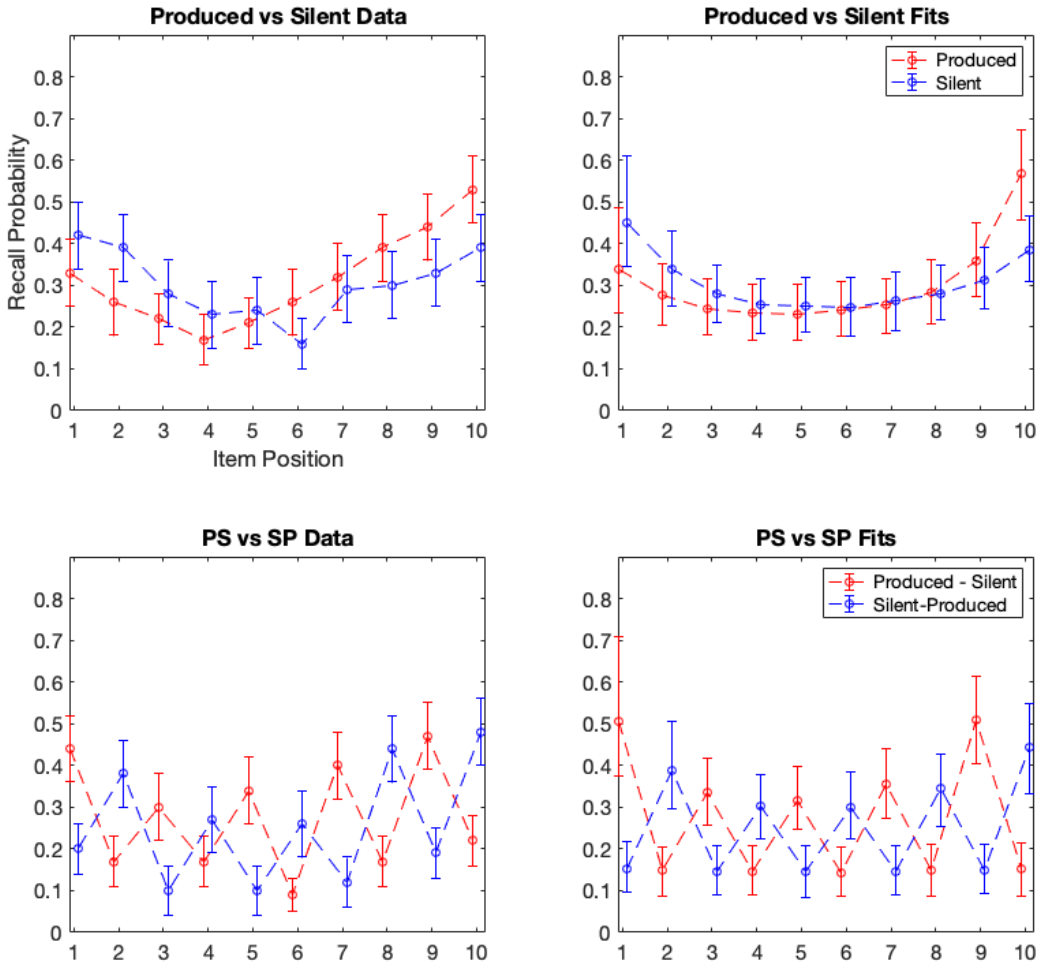

Figure 11. Data and model fits for the free recall task used in Experiment 3. For the model fits, error bars are 95% HDIs of the posterior predictions. As for Experiments 1 and 2, the qualitative and quantitative match between model and data is very good.

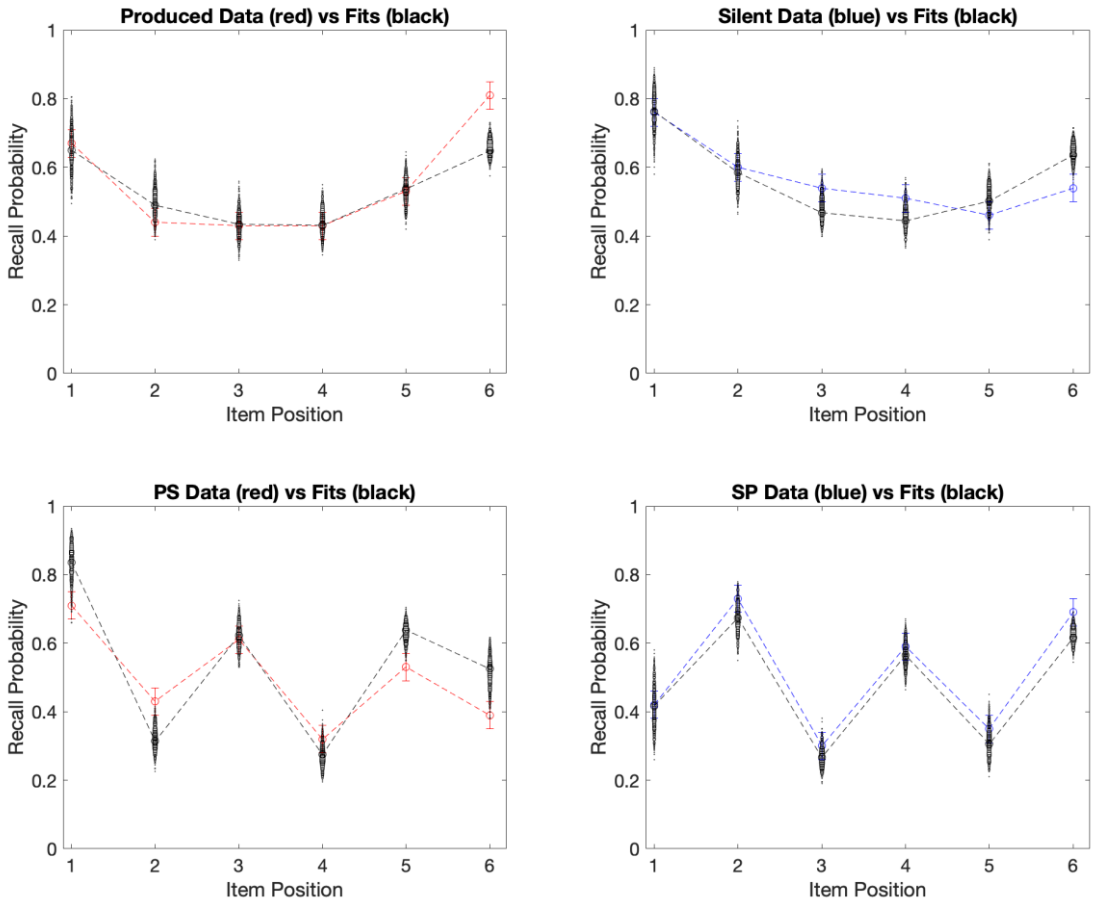

Figure 12. Data and model fits for Experiment 1b from Saint-Aubin et al. (2021). Model fits are black lines and boxes. Boxes represent histograms of the posterior predicted distributions. Black lines are the results of simulating the model with the best fitting parameters. (The slightly different way of plotting the data is to facilitate comparison with the other model fits in Saint-Aubin et al. (2021)). Although there is some misfitting in places, overall the level of both qualitative and quantitative agreement is good.

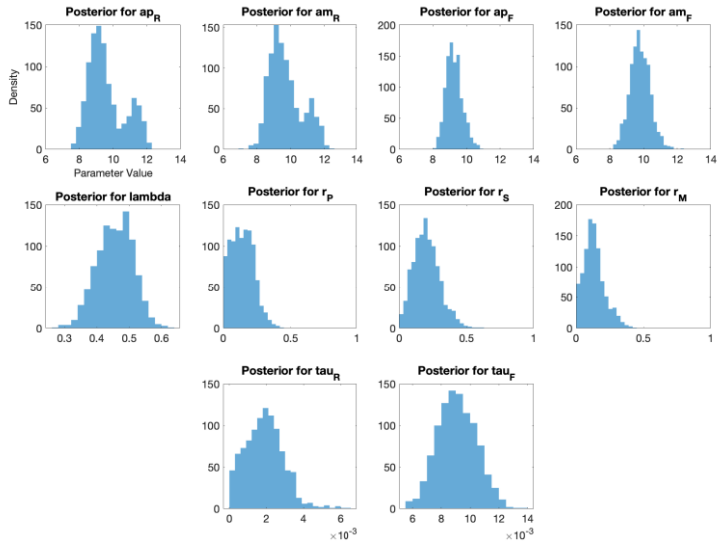

Figure 13a: Parameter posteriors for Experiment 1.

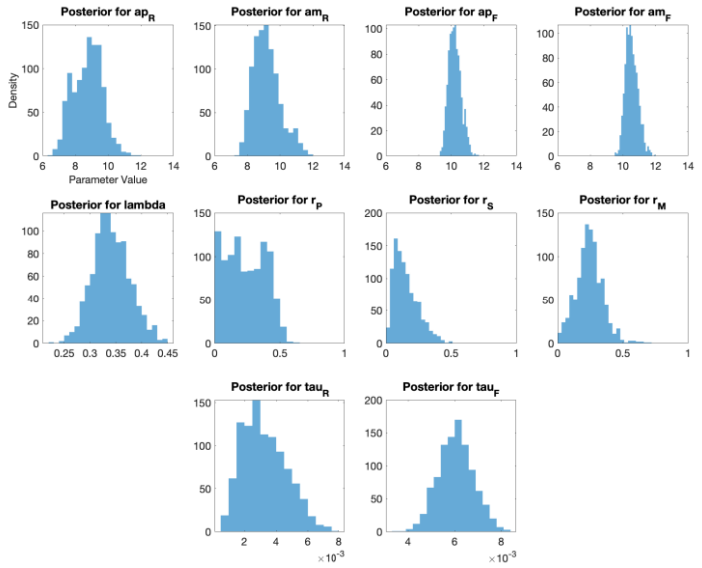

Figure 13b: Parameter posteriors for Experiment 2.

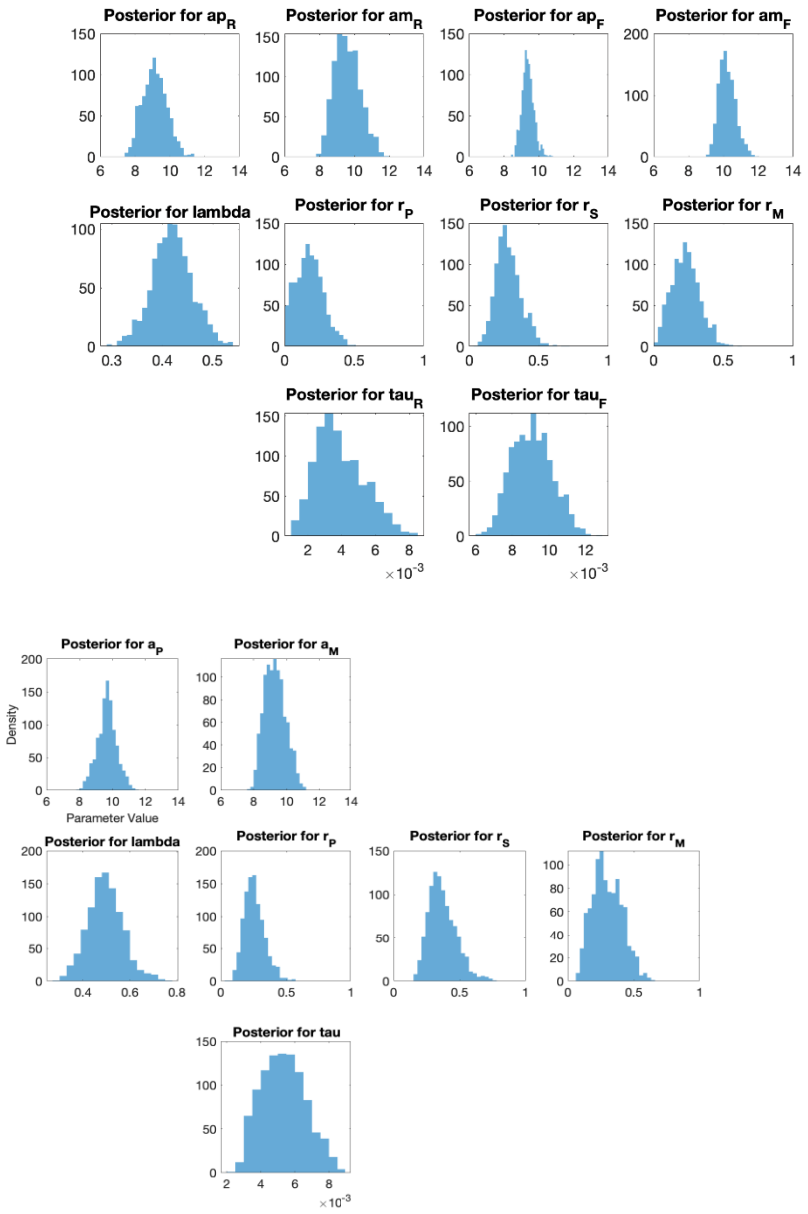

Figure 13c: Parameter posteriors for Experiment 3.

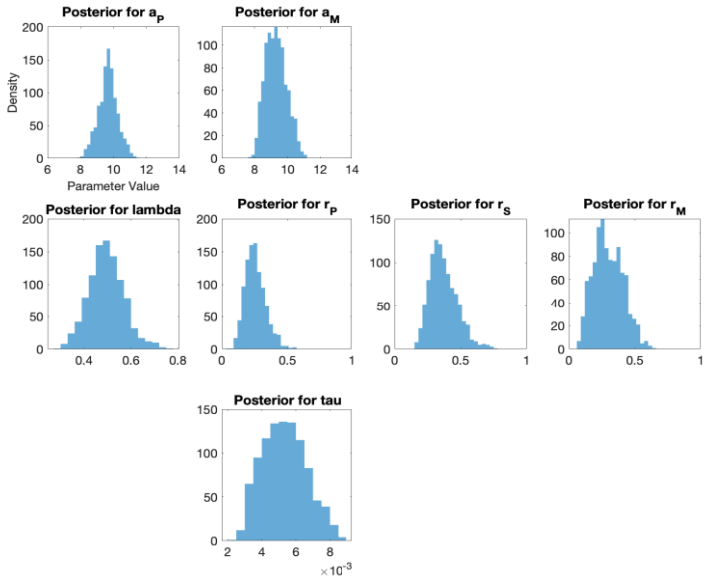

Figure 13d: Parameter posteriors for Experiment 1b of Saint-Aubin et al. (2021).

Figure 13: Parameter posteriors for the four experiments. Overall there is fairly good agreement between parameter sets, which is reassuring.

Table 1

Means and standard deviations (in parentheses) for the proportion of correct parity judgment and the number of parity judgment attempts

Proportion Correct						
	Experiment 1		Experiment 2		Experiment 3	
Condition	Reconstruction	Free Recall	Reconstruction	Free Recall	Reconstruction	Free Recall
Silent	0.90 (.07)	0.89 (.10)	0.86 (.17)	0.87 (.15)	0.94 (.05)	0.94 (.05)
Aloud	0.90 (.08)	0.89 (.09)	0.89 (.11)	0.90 (.08)	0.95 (.04)	0.95 (.03)
AS	0.89 (.10)	0.90 (.07)	0.88 (.13)	0.88 (.11)	0.95 (.05)	0.94 (.04)
SA	0.89 (.07)	0.89 (.10)	0.87 (.15)	0.89 (.11)	0.95 (.04)	0.95 (.04)
Number of Attempts						
	Experiment 1		Experiment 2		Experiment 3	
	Reconstruction	Free Recall	Reconstruction	Free Recall	Reconstruction	Free Recall
Silent	23.59 (4.33)	23.52 (4.45)	24.02 (5.27)	23.99 (4.65)	161.80 (32.19)	162.28 (33.10)
Aloud	24.18 (3.96)	23.79 (4.17)	23.91 (4.70)	24.96 (2.79)	162.50 (33.12)	163.07 (31.28)
AS	24.05 (4.30)	23.95 (4.12)	24.15 (4.25)	24.69 (4.02)	162.78 (32.88)	162.36 (31.67)
SA	23.74 (4.10)	23.85 (4.37)	24.12 (4.88)	24.58 (4.00)	162.81 (30.91)	163.84 (33.60)

Parameter	**Experiment 1 (Median, 95% HDI)**	**Experiment 2 (Median, 95% HDI)**	**Experiment 3 (Median, 95% HDI)**	**Experiment 1b Saint-Aubin et al. (2021) (Median, 95% HDI)**
λ Overwriting Parameter	0.46 [0.35, 0.56]	0.34 [0.28, 0.42]	0.42 [0.34, 0.50]	0.49 [0.35, 0.66]
R_P Rehearsal parameter for Produced items	0.14 [0.00, 0.31]	0.23 [0.02, 0.49]	0.18 [0.02, 0.40]	0.25 [0.12, 0.44]
R_S Rehearsal parameter for Silently read items	0.20 [0.04, 0.42]	0.13 [0.03, 0.37]	0.27 [0.12, 0.48]	0.36 [0.21, 0.61]
R_M Rehearsal parameter for Mixed lists	0.13 [0.01, 0.32]	0.24 [0.05, 0.46]	0.22 [0.06, 0.44]	0.29 [0.11, 0.53]
a_{P-R} Distance scaling parameter for Pure lists in Reconstruction condition	9.3 [8.1, 11.8]	8.8 [7.2, 10.5]	9.1 [7.9, 10.5]	9.7 [8.4, 10.9]
a_{M-R} Distance scaling parameter for Mixed lists in Reconstruction condition	9.5 [8.2, 11.7]	9.1 [7.9, 11.1]	9.5 [8.4, 11.2]	9.3 [8.2, 10.6]
a_{P-F} Distance scaling parameter for Pure lists in Free Recall condition	9.2 [8.4, 10.4]	10.2 [9.6, 11.0]	9.3 [8.7, 10.1]	-
a_{M-F} Distance scaling parameter for Mixed lists in Free Recall condition	9.8 [8.7, 11.1]	10.5 [9.9, 11.4]	10.2 [9.4, 11.3]	-
τ_R Temperature parameter for Reconstruction	0.0019 [0.0002, 0.0039]	0.0031 [0.0011, 0.0062]	0.0037 [0.0016, 0.0070]	0.0053 [.0032, .0080]
τ_F Temperature parameter for Free Recall	0.0091 [0.0066, .0118]	0.0061 [0.0046, .0075]	0.0091 [0.0071, 0.0112]	-

Table 2: Median and 95% HDI of the parameter posteriors for Experiments 1, 2, 3, and Experiment 1b from Saint-Aubin et al. (2021). There are few substantial differences, although some of the posteriors are rather wide.

Appendix: Model Fitting Details.

Since our model is too complex for an analytic expression for the likelihood to be derived, we used a version of Approximate Bayesian Computation (ABC) to carry out model fits (see Turner & Van Zandt, 2012, or Marin et al., 2012, for a review). ABC methods allow for Bayesian model fitting even in cases when the likelihood cannot be computed, by using simulated data to obtain an approximate likelihood. Specifically, we used a procedure known as ABC Partial Rejection Control (ABC-PRC) (Sisson et al., 2007, 2009) which we have previously used to fit the original Feature Model (Poirier et al., 2019) and the Revised Feature Model (Saint-Aubin et al., 2021).

ABC-PRC works by repeatedly sampling from a prior over the parameter space until it finds a set of parameters which generate a set of summary statistics (in our case serial recall curves) sufficiently close to the data. When this happens, the algorithm stores these parameter values, and moves on to the next particle in the generation. Once all particles in a generation have been associated with parameter sets, the algorithm gives each particle a weight depending on the prior, and then begins a new generation, sampling from the previous generation with probabilities given by the weights, and repeatedly perturbing around the previous parameter values until a set is found producing summary statistics even closer to the data. For full details see Sisson et al. (2007) (Note also the errata, Sisson et al., 2009).

Under ABC-PRC, the posterior estimates for the parameters are just the fraction of particles in the final generation with that parameter value. Posterior predicted distributions of the summary statistics are also easily obtained.

The important parameters for ABC-PRC are the number of particles (set to 1000 for all fits reported here), the details of the prior, the proposal distributions, and the minimum tolerances for each fit. Setting the number of generations and the tolerances requires some trial and error. Lower tolerances will tend to result in a better match between model and data, but at some point the computational cost becomes prohibitive. The choice of distance function is also important – for Experiment 1, 3, and Experiment 1b from Saint-Aubin et al. (2021) we made use of all data points, whereas for Experiment 2 we only used the first and last eight items, since that is where the majority of the interesting behavior lies. For Experiments 1, 2, and 3, we also implemented a distance function inspired by robust regression techniques, where we computed the squared difference between simulated and real data for each of the 64, 80, or 128 data points we used, but then dropped the four largest values before summing, square rooting, and comparing with the acceptance criterion at each generation. In effect, the ABC algorithm can choose four data points to be ‘outliers’ whenever it compares simulated and real data. The aim is to make the fitting less sensitive to noise in the data. This was not done for Experiment 1b from Saint-Aubin et al. (2021) to match the way the fits were done to the other data sets reported there.

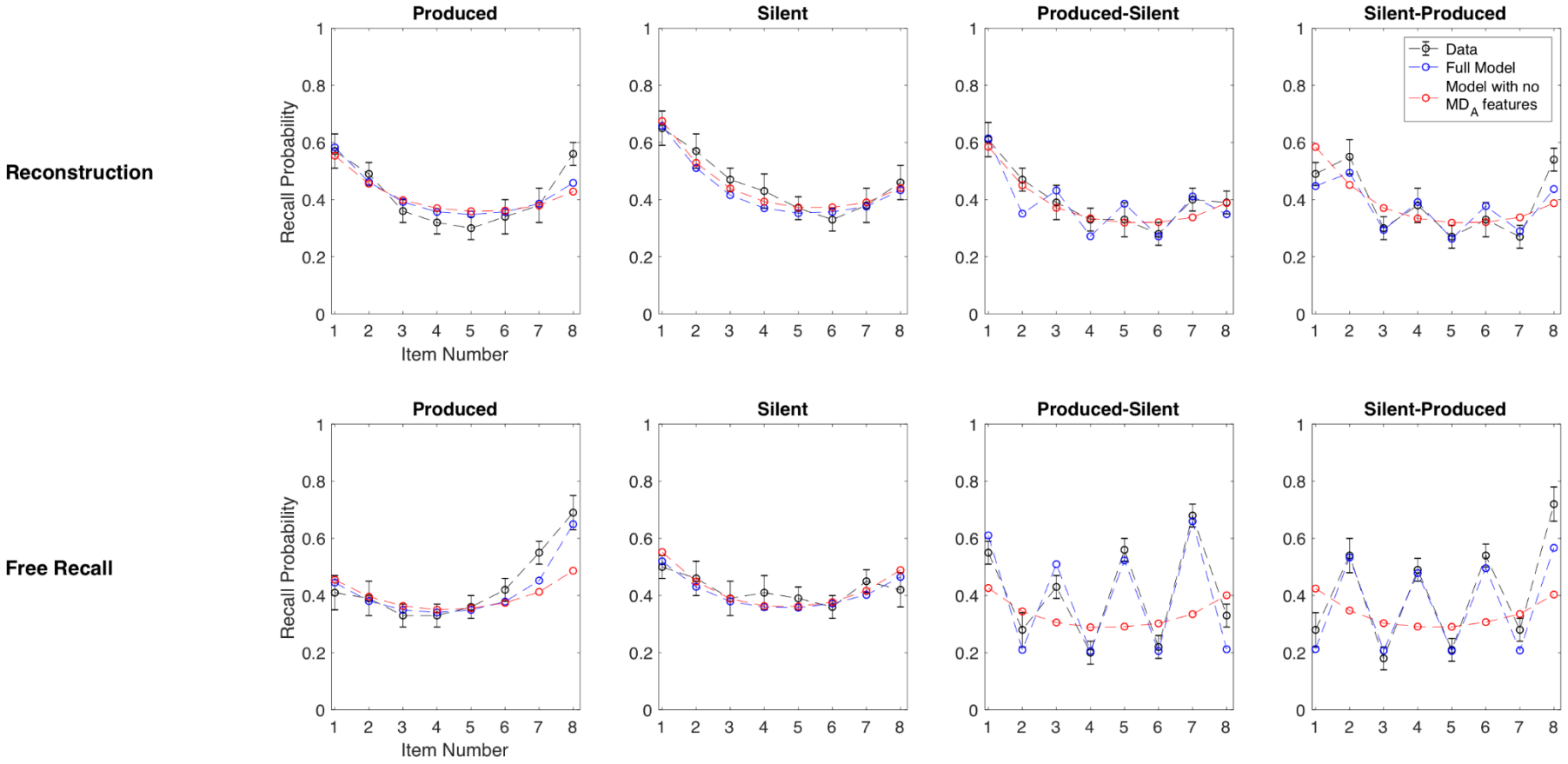

Figure S1: Model fits (Blue) to the data for Experiment 1 (Black) together with fits obtained from a model with the number of modality dependant (auditory) features set to zero (Red). In other words, the number of features is the same for produced and silent items. The most obvious qualitative difference is the inability of the model with no relative distinctiveness (Red) to produce the characteristic sawtooth pattern for mixed lists.

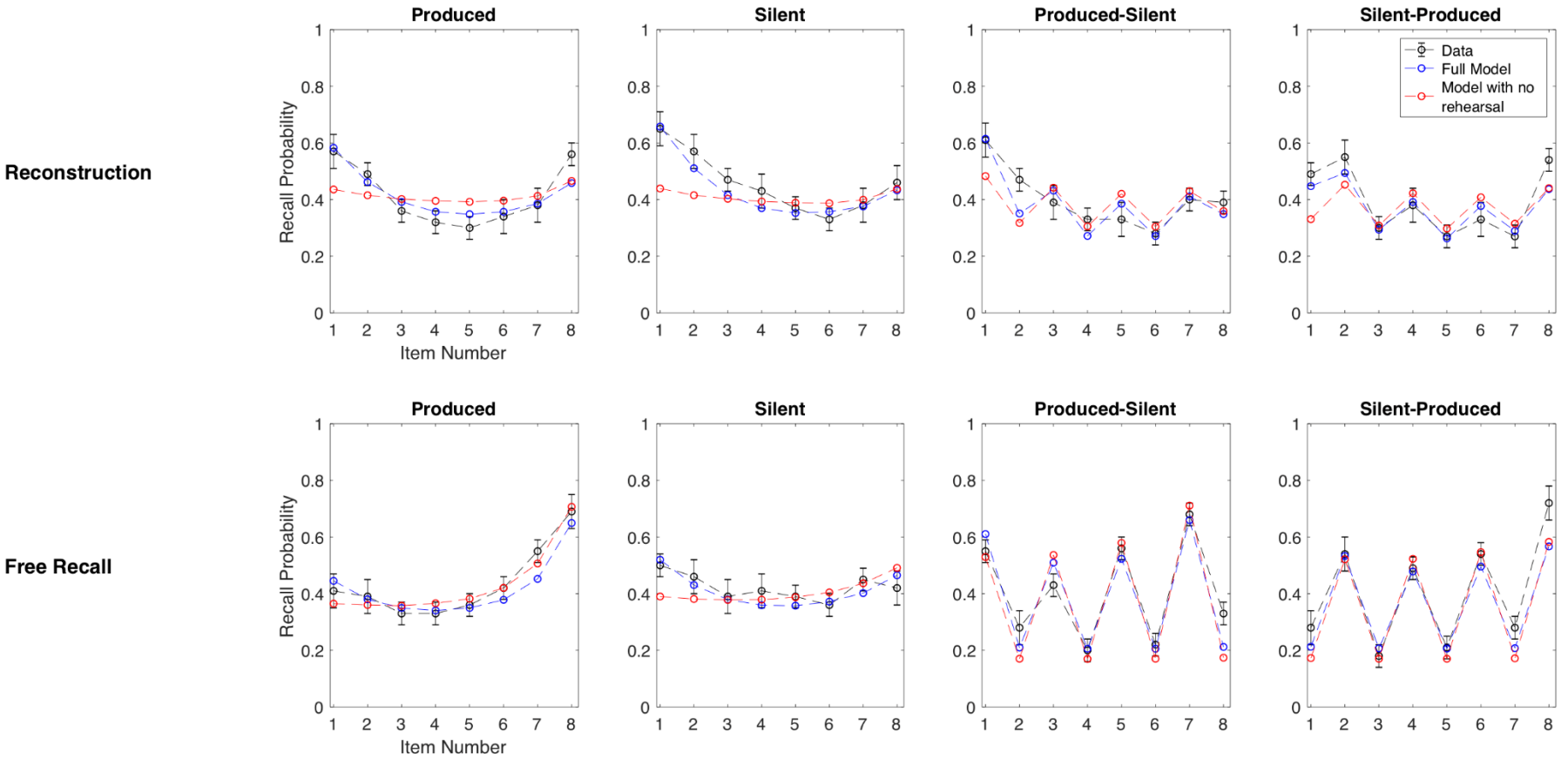

Figure S2: Model fits (Blue) to the data for Experiment 1 (Black) together with fits obtained from a model with rehearsal parameters all set to zero (Red). The most obvious qualitative difference is the inability of the model with no rehearsal (Red) to explain the higher accuracy for items early in the list, particularly in the pure list reconstruction cases.

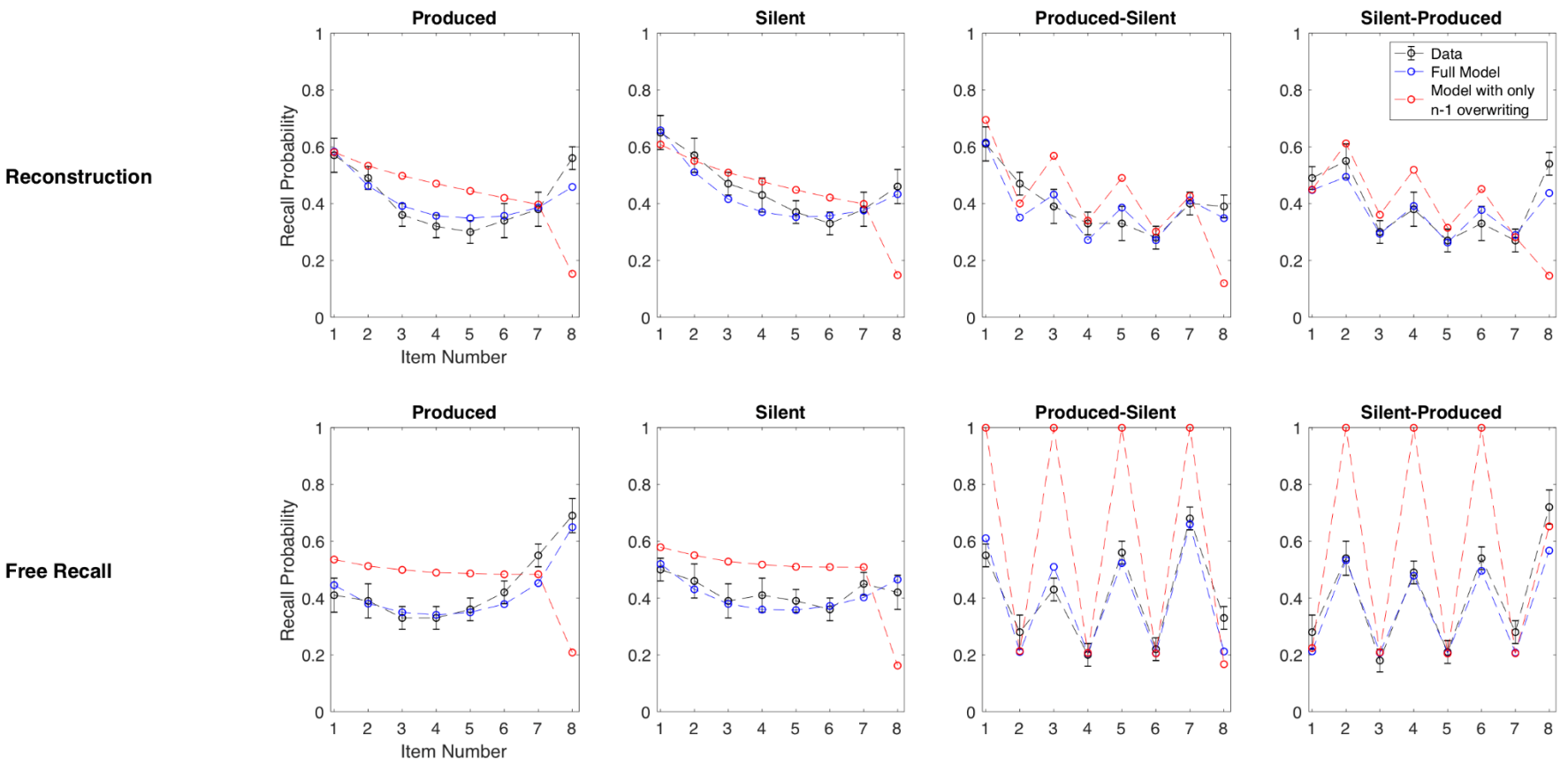

Figure S3: Model fits (Blue) to the data for Experiment 1 (Black) together with fits obtained from a model with lambda set >>1 (Red). In other words overwriting only occurs on the most recently presented item. There are some clear qualitative differences, firstly, since activity after the list presentation only overwrites features of the final item, there is a drop in recall probability for the final item in the pure list free recall conditions. Secondly, because the modality dependent auditory features in mixed lists cannot be overwritten at all (excepting maybe the final item) the model predicts very high recall probabilities for these items in free recall.