A Model of the Production Effect over the Short-Term:

The Cost of Relative Distinctiveness

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

[1]School of Psychology, Université de Moncton

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

Authors Note

Jean Saint-Aubin https://orcid.org/0000-0002-4799-6912

James M. Yearsley https://orcid.org/0000-0003-4604-1839

Marie Poirier https://orcid.org/0000-0002-1169-6424

Dominic Guitard https://orcid.org/0000-0002-4658-3585

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.

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 Practice Statement

The data is available on the Open Science Framework project page (https://osf.io/7zdws/?view_only=dab05dd71219493f80ffc60ffdeea4fa).

Abstract

The production effect relates to the better memory of words read aloud during a study phase compared to silently read items. Here, we examined the production effect for memory over the short-term. In long-term memory tasks, the effect generates a complex pattern of results where production interacts with memory task and list composition. Within an immediate ordered recall paradigm, involving both item and order information, we tested the item-order account, recently called upon to explain the production effect. We also analysed results as a function of serial position. Results of the first five experiments were highly consistent, but hard to reconcile with the item-order account. Instead, we put forward an interpretation based on relative distinctiveness and the costs of the richer encoding associated with production. The predictions we derived from this interpretation were supported in the final experiment. Moreover, we tested the interpretation through a new version of the Feature Model. Overall, the work highlights the value of the production effect as a prototypical distinctiveness phenomenon illuminating the interaction of encoding and retrieval processes, the value of feature-rich representations, and the costs that can be associated with feature-generating distinctive processing.

Keywords: Production Effect; Feature Model; Immediate Serial Recall

A Model of the Production Effect over the Short Term:

The Cost of Relative Distinctiveness

Recent research on episodic memory has seen interest in the production effect increase (see MacLeod & Bodner, 2017, for a review). Simply put, when some of the words within a list are pronounced aloud – i.e. produced – they tend to be better remembered than those read silently. First identified in the sixties and seventies (e.g., Crowder, 1970; Hopkins & Edwards, 1972; Pollack, 1963), the phenomenon was systematically studied more recently by MacLeod and collaborators, who coined the term production effect (MacLeod et al., 2010). As is the case for several encoding effects, the overall pattern of findings proved to be complex, with the strength of the effect depending on the memory task called upon, as well as on list composition and design. Other encoding effects include the generation effect (Serra & Nairne, 1993), the bizarreness effect (McDaniel & Einstein, 1986), the enactment effect (Engelkamp & Dehn, 2000), and others (see, e.g., Nairne, 1988a). In comparison to these, the production effect can be seen as relatively simple – it involves little extra processing compared to generating a bizarre image, for example. Instead, production involves a highly compatible and easy to observe response: reading verbal material aloud. In that sense, it is one of the best encoding effects to explore as it makes it easier to get to the fundamentals of what is causing the obtained patterns. In sum, the production effect can be thought of as a prototypical encoding effect, one that can help to elucidate the principles underlying the influence of such effects on memory.

In the present study, we had two main aims: The first was to determine if the production effect also held for short-term memory (STM), as the work alluded to above focussed on long-term memory. Should our results reveal the basic effect, we wished to systematically document

the phenomenon in STM. Our second aim was originally to test one of the promising explanations of encoding effects: The item-order account (McDaniel & Bugg, 2008). As applied to the production effect, the account explains the interaction between list composition, memory task, and production by proposing that a) item-specific and order information compete for encoding resources and b) in some tasks, disruption of order encoding leads to predictable recall difficulties due to the role of order information in guiding retrieval (Jonker et al., 2014; McDaniel & Bugg, 2008). We return to this explanation of the production effect later. At this point, we note that our original objective was to test how well these ideas can account for the production effect in STM; to do so, we called upon well-known short-term memory tasks where item and/or order information are considered central. Below, we briefly review the relevant data in the field as well as how the item-order account might explain the known pattern of findings; we then present five experiments as well as a quantitative model of our results.

To anticipate, the impact of production on some aspects of STM performance was spectacular, but unpredicted by the item-order account. The follow-up studies suggest a compelling interpretation of the production effect in STM, where limited encoding resources interact with relative distinctiveness to produce the rich and complex mnemonic behaviour that we observed.

The account we propose emphasises the importance of local distinctiveness effects as well as trade-offs at the point of encoding, i.e., distinctive processing comes at a cost. We use the expression ‘local distinctiveness’ because our findings and modelling underline the fact that differences between contiguous items—a form of contrast effect—have a very significant effect on STM performance. We argue that the principles highlighted in explaining these findings reveal some core truths about encoding effects and STM functioning more generally. Some of

these interpretations may also apply to long-term memory, although we did not test this directly; hence, the latter suggestions are more speculative.

The Production Effect

In a number of circumstances, relative to silent reading, production enhances future recall and the effect can last up to a week (Grohe & Weber, 2018; Ozubko et al., 2012) and occur even with an incidental learning procedure (see, e.g., Greene & Pearlman, 1998). Moreover, different forms of production have a positive effect, including singing, mouthing items, spelling, writing, typing, and even imagining typing; that said, except for singing, the latter effects are typically smaller than what is observed when items are read aloud (Forrin et al., 2012; Jamieson & Spear, 2014; MacLeod et al., 2010; Quinlan & Taylor, 2013, 2019). Importantly however, the size of the effect depends on list composition and task (MacLeod & Bodner, 2017). In recognition tasks, when lists comprise a mixture of items read aloud versus silently, a sizable production effect is obtained (see, e.g., Gathercole & Conway, 1988; Hopkins & Edwards, 1972; MacLeod et al., 2010). However, the effect is much reduced and less reliable when the two item types are studied separately in what are called pure lists – i.e., lists containing only silently read or only read aloud items (see Fawcett, 2013, for a meta-analysis). With recall tasks, the reported pattern is slightly different; with mixed lists, containing both silently read and aloud items, there is again a strong production advantage, but the effect disappears when pure lists are tested (Forrin & MacLeod, 2016a; Jonker et al., 2014; Lambert et al., 2016).

One promising hypothesis to account for the pattern across tasks and list types is the item-order account (Jonker et al., 2014; McDaniel & Bugg, 2008). According to this view, producing items involves more item-specific processing relative to silently reading items; when pronouncing words aloud, the motor features as well as the auditory features involved would

enhance / add to the encoded event. In mixed lists, these more elaborate records would create a relative distinctiveness advantage. The item-order account also emphasises the importance of relational information in recall tasks; more specifically, order information is thought to play a central role in retrieval processes by guiding retrieval in tasks like free recall (see, Beaman & Jones, 1998; Grenfell-Essam, et al., 2017).

With respect to order information, the suggestion is that producing items has a negative effect (Forrin & MacLeod, 2016a; Jonker et al., 2014; Lambert et al., 2016). The item-specific processing is thought to have a cost: It disrupts order encoding. In item recognition tasks, order information is not relevant but item-specific information is central; hence, according to the item-order account, in pure lists, produced items should be better recognised than silent items because the extra item-specific information will support recognition. In the case of mixed lists, the same is true; however, the relative distinctiveness of produced items compared to silently read words boosts the recognition advantage for the produced items. In recall tasks, the same relative distinctiveness produces an advantage in mixed lists; it is thought that this advantage overwhelms any negative impact of production on order information. In pure lists, however, the influence of relative distinctiveness is much reduced for produced items; moreover, any advantage of item-specific processing is offset by the disruption of order information encoding that production causes. The results can go in any direction depending on the influence of item-specific processing and of how disruptive it is to order encoding. Because the effects of item-specific processing and disruption of order encoding are in opposite directions, any effect will be small and so far, only null effects have been reported (see Forrin & MacLeod, 2016a).

These relatively precise predictions of the item-order account were tested by Jonker et al. (2014). More specifically, they tested the idea that differential order information processing

could account for the production effect findings in recall tasks. In their first experiment, they examined the influence of list type (i.e., mixed or pure lists) on performance using an order reconstruction task. In such a task, the studied items are provided anew at the point of recall and the participants attempt to reproduce the order in which the words were studied. Based on the item-order account, order reconstruction performance should be better for lists read silently than for lists read aloud because reading aloud is thought to disrupt order encoding. Moreover, as the items are provided at the point of recall, any item-specific processing advantage is likely to be reduced. The reasoning in the case of mixed lists is that reading items aloud would hinder order encoding of silent items, while encoding of silent items would support the order encoding of produced items. The result of the latter would be that any order advantage associated with silently read words would be reduced or eliminated. Based on this analysis, Jonker et al (2014) also predicted that order reconstruction for silent items in mixed lists would be poorer than in pure lists. Conversely, they expected more order errors for aloud items in pure relative to mixed lists. Their first experiment involved lists of eight words followed by a thirty-second distractor task after which participants completed the order reconstruction test. The results of the experiment supported the item-order account predictions. For pure lists, participants were able to remember the studied order of silent items better than the order of aloud items; in contrast, in mixed lists, performance for silent and aloud items was equivalent.

In a second experiment, Jonker et al. (2014) wanted to verify that the use of a reconstruction task had not altered encoding and processing strategies in such a way that the typical production effect was not observable. They used the same order reconstruction task for half of the trials but for the other half, the task was free recall. Participants were only told at the point of retrieval which task they needed to accomplish. The results for the reconstruction task

were similar to those of the previous experiment. There was better order reconstruction for pure silent lists, as expected based on the item-order account, while for the mixed list, there was no significant difference between the silently read and produced items (although there was a trend toward better order recall for items read aloud). With respect to free recall, aloud items were better remembered than silent items when they were presented in mixed lists but for pure lists, recall of silent and aloud items could not be distinguished. The authors also examined two order recall measures for the free recall task. These analyses showed that the order of the items was better reproduced in the free recall of pure lists read silently than for pure lists read aloud; moreover, silent reading produced better order memory than what was observed in mixed lists. Order memory in pure aloud lists was not distinguishable from order memory in mixed lists.

The Jonker et al. (2014) study provided clear support for an item-order account with the predictions derived from the account buttressed by the main data of both experiments and when measures of order memory in free recall were considered. The type of episodic memory task called upon was somewhat atypical however, in that the list length was reduced (8 items) relative to standard free recall tasks (15+ items). There is some evidence that the importance of order information in recall increases with shorter lists (Grenfell-Essam et al., 2017).

To address this issue, Lambert et al. (2016) and Forrin and MacLeod (2016a) followed up on the Jonker et al. (2014) study. They set out to further test the item-order account with a task involving longer lists and free recall. The general pattern of results in all experiments supported the predictions of the item-order account in that there was no difference between silent and aloud lists for the groups studying pure lists whereas a reliable production effect was found for the mixed list group. However, Lambert et al. (2016) found little evidence that order memory could account for the production effect pattern; they examined two measures of order memory and

found no support for the expected differences in order memory predicted by the item-order account. However, as mentioned above, Grenfell-Essam et al. (2017) have shown that in free recall, order information is most critical with shorter lists.

From the above, we conclude that the item-order account seems like a promising explanation for the production effect in general and even more in STM where item and order information are heavily called upon. However, there are inconsistencies in the reported findings; replications are needed, and further research should test the specific predictions attached to the item-order account in a variety of settings.

These were an important part of the aims that we initially pursued when conducting the first experiment described below. We set out with two objectives in mind. The first was to test the item-order account of the production effect within paradigms recognised for calling upon item and order information – namely, immediate serial memory tasks. A pattern of results similar to those reported by Jonker et al. (2014) and by Forrin and MacLeod (2016a) would provide support for the item-order account, while a discrepant pattern of results like the one reported by Lambert et al. (2016) would provide evidence against the account. Second, as the tasks are typically associated to memory over the short-term, we wished to better establish the nature of the production effect within such a setting. The presence of a production effect in typical STM tasks analogue to the production effect observed in typical long-term memory tasks would provide support for the view that the same set of principles govern memory irrespective of the tasks (Surprenant & Neath, 2009). In short-term ordered recall tasks, with pure lists, an advantage of produced items has been shown in some studies (e.g., Conrad & Hull, 1968), but not in others (e.g., Kappel et al., 1973). However, mixed list designs have not yet been

investigated in immediate serial recall so the complete pattern of findings remains to be established.

Experiment 1

In Experiment 1, we investigated the production effect in immediate serial recall (Experiment 1A) and immediate order reconstruction tasks (Experiment 1B). In the immediate serial recall task, participants were sequentially presented with six words; immediately following this presentation, their task was to recall the items in their exact order of presentation, starting with the first studied item, continuing to the second, and so on. To perform well in this task, both item and order information must be remembered. In the case of the immediate reconstruction task, the presentation of the to-be-remembered items proceeds exactly as in the immediate recall task; however, at the point of retrieval, all the studied items are provided simultaneously in a new random order. The participant's task is to re-order these items to reproduce the sequence of the just studied list. Again, the first item studied needs to be positioned, followed by the second, etc. In this case, both item and order are important but as the items are provided, successful performance relies more heavily on memory for the order in which the items originally appeared.

Both tasks were included because together they provide a more complete empirical picture of the production effect in STM. In effect, although item and order information can be isolated in immediate serial recall by analysing item and order errors, Guitard, Saint-Aubin, and Cowan (2020) recently showed an asymmetrical interference between item and order information in STM. More specifically, compared to item information, order information was more vulnerable to interference. Therefore, the order reconstruction task—which is a purer measure of order information—allowed a more stringent test of the item-order account with respect to order information. In both Experiments 1A and 1B, pure and mixed lists were used. In

the pure list conditions, participants read all six words in the same manner, either aloud or silently. In the mixed list condition, participants alternated between reading words aloud and silently; in other words, every other word was read aloud, while the remaining words were read silently.

In the sixties and seventies, within research on the modality effect in STM, the production effect was investigated in immediate serial recall. The modality effect refers to the large recall advantage for the last serial position (or last 2 positions) which is observed with auditory presentation relative to visual presentation (see, Penney, 1989, for a review). Interestingly, because of the technical difficulties related to presenting items auditorily at the time, it was common to ask participants to read aloud in the auditory condition. With this procedure, the items read aloud were systematically better recalled than the items read silently, but only for the last serial positions (Conrad & Hull, 1968; Crowder, 1970; Greene & Crowder, 1984; Kappel et al., 1973; Murray, 1965; Nairne & Walters, 1983; Poirier et al., 2005; Routh, 1970). For the early serial positions, there was no difference between items read silently or aloud (Conrad & Hull., 1968; Nairne & Walters, 1983; Poirier et al., 2005; Routh, 1970) or a disadvantage was observed for the items read aloud (Crowder, 1970; Greene & Crowder, 1984; Kappel et al., 1973).

While many studies at the time investigated the production effect with pure lists in immediate serial recall, mixed lists were mostly ignored. Greene's (1989) study is the exception. In his study, the first four items were read aloud and the last four were read silently or vice-versa. Results revealed a large production effect with an advantage for items read aloud compared to silently read items. However, it is difficult to extrapolate these results to lists in which produced and silently read items systematically alternate. In effect, previous studies with word frequency revealed a large effect with pure lists (Guérard & Saint-Aubin, 2012) and half-lists (Watkins,

1977), but no effect with alternating lists (Hulme et al., 2003). This pattern of results differs from that observed with the word length effect for which there was an effect with pure (Baddeley et al., 1975), half (Cowan et al., 1992), and alternating (Cowan et al., 2003) lists. Within a standard immediate serial recall task, Experiment 1 allowed us to directly compare memory for pure lists of silently read and produced items as well as mixed, alternating, lists.

Based on the predictions of the item-order account, the immediate serial recall (Exp. 1A) of mixed-modality lists should lead to better recall for the produced items. With pure lists, there should be little or no effect of production because the advantage of enhanced item-specific encoding will be offset by worse order information encoding. Also, when comparing pure list performance with mixed list results, we expected to see an improvement in produced item recall going from the pure to the mixed condition; this would be due to better order encoding for the aloud items in mixed lists relative to pure lists afforded by the introduction of silent items. Moreover, because produced items benefit from more elaborate traces than silently read items, this should lead to a relative distinctiveness benefit of produced items compared to silently read items within the same lists. By contrast, in pure produced lists, all items benefit from these more elaborate traces and there is no relative distinctiveness advantage. As for the silent items, we expected performance to drop when going from pure to mixed lists because order encoding will be disrupted by the item-specific processing of the produced items. We predicted that measures of order memory such as order errors would also conform to the predictions of the item-order account.

With respect to immediate reconstruction (Exp. 1B), following Jonker et al. (2014), we made the simplifying assumption that item recall would not contribute significantly to performance (but see, Neath, 1997). Therefore, with pure lists, performance should be superior

for silently read items than for produced items, and this advantage should vanish with mixed lists (Fawcett, 2013; Forrin & MacLeod, 2016a; Jonker et al., 2014).

Experiment 1A

Method

Participants. Forty-eight students from Université de Moncton (9 men, 39 women; mean age of 20 years old), took part in this experiment, and received course credits for their participation. Participants were randomly assigned to one of the two list types (pure or mixed), with the restriction that 24 participants were assigned to each group. All participants were native French speakers and had normal or corrected to normal vision.

Materials. Three hundred French words were used. The words were all nouns comprising two phonological and orthographic syllables, with frequency ranging from 5.07 to 497.43 occurrences per million (M = 36.16, SD = 57.90), according to Lexique 3 (New et al., 2004). Fifty lists, each containing six words, were generated by drawing without replacement from this pool with the constraint that no words within a list could rhyme. Two of the 50 lists were randomly selected to serve as practice trials. The remaining 48 lists were used for experimental trials. The experiment was controlled with E-Prime 2.0 (Psychology Software Tools, 2016).

Design. A 2 x 2 mixed-design was used with list type (pure vs. mixed) as the between-participants factor and presentation modality (read silently vs. read aloud) as the repeated factor. Each participant completed 2 blocks of 24 experimental trials preceded by 1 practice trial, for a total of 50 trials. Each of the 50 trials involved the immediate serial recall of a six-item list. For the group assigned to pure lists, in one block of 25 trials, items were read silently; in the other block of 25 trials, they were read aloud. This means that the practice list was silently read in the

silently read condition and was read aloud in the other condition. Conditions were blocked to emulate the procedure used in typical long-term memory tasks in which a single list is presented to participants in a condition (see, e.g., MacLeod et al., 2010). Furthermore, it has been shown that the anticipation of an upcoming oral reading can disrupt memory for the preceding items (see, e.g., Brenner, 1973; Forrin et al., 2019). Therefore, by blocking the conditions, any difference across conditions could be attributed to the condition itself rather than to inter-trial effects. For the group assigned to mixed lists, in each list, three of the six items were read silently, and the other three items were read aloud. In half of the lists, Items 1, 3, and 5 were read aloud, while it was Items 2, 4, and 6 for the other half. Again, in each block, the practice list has the same structure as the 24 experimental lists. Across participants, the order of presentation of the blocks was systematically counterbalanced. In addition, across all participants, each word was presented as often in each presentation modality. Finally, while the order of presentation of the words within the list was the same across participants, the order of presentation of the lists was randomized for each participant.

Procedure

Participants were tested individually in a quiet testing room in one session lasting approximately 30 minutes. The words were presented sequentially at the center of the computer screen, at a rate of 1 word per second (1000 msec on, 0 msec off). Words were displayed in a 24-point Times New Roman font. In the pure list condition, for half of the participants, all words were displayed in blue, while they were displayed in black for the other half. In the mixed list condition, for half of the participants, blue words had to be read aloud, and black words had to be read silently, while it was the reverse for the other half. Following the presentation of the last word, three question marks were presented on the screen to indicate the recall period.

Participants recalled the words by writing them on an answer sheet. Participants were told that if they forgot a word in a given serial position, they should leave the space blank. They were further instructed that they would not be penalized for spelling errors. Backtracking was not allowed, and the experimenter was present throughout the testing session to ensure compliance with the instructions.

At the beginning of the experiment, instructions were presented on the screen. The instructions stipulated that participants had to memorize lists of six words and to perform a written recall on a response sheet. Participants were instructed to recall the words in the order in which they were presented, from the first word to the last word. For the mixed lists, the instructions specified the colour that had to be read aloud.

Results & Discussion

Responses were first scored with a strict recall criterion. With this criterion, to be considered correct, words had to be recalled in their original presentation position. The proportion of correct responses as a function of input modality and list type is shown in Figure 1, and the serial positions are shown in Figure 2. As predicted by the item-order account, inspection of the top row of Figure 1 reveals a large advantage of produced words over silently read words in the mixed list condition (right panel), which is much reduced in the pure list condition (left panel). Furthermore, compared to the mixed list condition, in pure lists, silently read items are better recalled, while produced items are less well recalled. The overall pattern of results nicely extends previous findings reported with free recall (Forrin & MacLeod, 2016a; Jonker et al., 2014; Lambert et al., 2016) and fits well with the item-order account (McDaniel & Bugg, 2008).

However, when performance was split by serial positions, a different picture emerged (first row, Figure 2). In mixed lists, sawtooth serial position curves were observed with a systematic advantage for produced items over silently read items. The novel effect is in line with previous findings observed with lexicality and phonological similarity, although its magnitude is much larger here; it diverges from results with word frequency (Baddeley, 1968; Hulme et al., 2003). With pure lists, produced items were better recalled on the last three serial positions, but the reverse was observed for the first three. The better recall of produced items on the last serial positions is reminiscent of the modality effect (see, e.g., Conrad & Hull, 1968), while the cost of saying the first items aloud is in line with some previous studies (Crowder, 1970; Greene & Crowder, 1984; Kappel et al., 1973).

The strict recall data were analysed with a 2 X 2 X 6 mixed-design ANOVA with list type (mixed or pure) as the between-participants factor and input modality (silent reading or production) and serial position as repeated-measures factors. The ANOVA revealed a main effect of input modality, $F(1,46) = 110.66$, $p < .001$, $\eta_p^2 = .71$, and serial position, $F(5,230) = 53.21$, $p < .001$, $\eta_p^2 = .54$, but the main effect of list type was not significant, $F(1,46) = 1.59$, $p = .21$. All interactions were significant. More precisely, the three two-way interactions were significant with an interaction between list type and input modality, $F(1,46) = 72.61$, $p < .001$, $\eta_p^2 = .61$, list type and serial position, $F(5,230) = 2.75$, $p < .05$, $\eta_p^2 = .06$, and serial position and input modality, $F(5,230) = 9.27$, $p < .001$, $\eta_p^2 = .17$. The three-way interaction between the list type, serial position, and input modality was also significant, $F(5,230) = 16.46$, $p < .001$, $\eta_p^2 = .26$.

The three-way interaction was decomposed by means of two separate ANOVAs. For pure lists, the 2 X 6 repeated-measures ANOVA with input modality (read silently or aloud) and serial position as factors revealed a main effect of input modality, $F(1,23) = 8.79$, $p < .01$, $\eta_p^2 = .28$, of serial position, $F(5,115) = 23.62$, $p < .001$, $\eta_p^2 = .51$, and an interaction between input modality and serial position, $F(5,115) = 17.53$, $p < .001$, $\eta_p^2 = .43$. Post hoc Tukey's HSD tests revealed that produced items were significantly better recalled than items read silently at Position 4, 5, and 6, while the reverse was observed at Position 1, 2 and 3, all *p*s < .05. For mixed lists, the 2 X 6 repeated-measures ANOVA with list composition (PSPSPS vs. SPSPSP) and serial position as factors revealed a significant main effect of serial position, $F(5,115) = 39.53$, $p < .001$, $\eta_p^2 = .63$, as well as significant interaction between serial position and list composition, $F(5,115) = 75.18$, $p < .001$, $\eta_p^2 = .77$. There was no significant effect for list composition, $F < 1$. Tukey's HSD tests revealed that items read aloud were better recalled than those read silently at all serial positions, all *p*s < .001.

Item and order errors. An item error was defined as a missing item or a recalled word that was not part of the list. The proportion of item errors was computed in each condition by dividing the number of item errors by the number of presented items. As can be seen in the top row of Figure 3, results with item errors are analogous to those with strict scoring, albeit taking into consideration that errors are plotted in Figure 3, while correct recall was plotted in Figure 1. Accordingly, the 2 x 2 mixed-design ANOVA with list type (mixed or pure) and input modality (read silently or aloud) as factors revealed a significant main effect of input modality, $F(1,46) = 109.62$, $p < .001$, $\eta_p^2 = .70$, and an interaction between list type and input modality, $F(1,46) = 60.12$, $p < .001$, $\eta_p^2 = .57$. There was no main effect of list type, $F < 1$. Tukey's HSD tests revealed a smaller proportion of item errors for words read aloud than for those read silently in

mixed ($p < .001$) and pure lists ($p < .001$). Furthermore, compared to mixed lists, in pure lists, the proportion of item errors was larger for words read aloud ($p < .001$) and smaller for words read silently ($p < .001$).

An order error was credited when an item was recalled, but in the wrong serial position. The proportion of order errors was computed in each condition by dividing the number of order errors by the number of presented items. The results are summarised in the top row of Figure 4. Overall, the proportion of order errors was much smaller than the proportion of item errors. This finding is typical of performance in immediate serial recall with an open pool of items. In mixed lists, participants made more order errors for produced than for silently read items. This pattern was attenuated with pure lists due to the higher error rate for silently read items. The 2 x 2 mixed-design ANOVA with list type (mixed or pure) and input modality (read silently or aloud) as factors revealed a significant main effect of input modality, $F(1,46) = 12.04$, $p < .01$, $\eta_p^2 = .21$, but neither the main effect of list type, $F = 1.64$, $p = .21$, nor the interaction, $F = 2.78$, $p = .14$, were significant. Tukey's HSD tests revealed a smaller proportion of order errors for words read aloud than for those read silently in mixed ($p < .01$) and pure lists ($p < .05$). Furthermore, compared to mixed lists, in pure lists, the proportion of order errors was larger for silently read words, ($p < .05$), but did not differ for words read aloud ($p = .98$).

Results of Experiment 1A revealed the presence of a large production effect in an STM paradigm with mixed lists which is much reduced, but still present, with pure lists. This pattern of results with a strict serial recall criterion, factoring in item and order information, is in line with the item-order hypothesis (Jonker et al., 2014; McDaniel & Bugg, 2008). According to this hypothesis, relative to silently read items, produced items would benefit from better item-specific processing, but would suffer from poorer relational information. Therefore, as predicted,

participants made fewer item errors for produced than for silently read items and the advantage of produced items was larger in the mixed list condition in which they benefit from the largest item distinctiveness advantage. For order information, as predicted, participants made fewer order errors for silently read items than for produced items in pure lists. However, contrary to the predictions derived from the item-order hypothesis, the order recall advantage of silently read items was larger with mixed lists. With mixed lists, a null effect or a smaller order recall advantage was expected because the presence of produced items within the list would disrupt order processing of silently read items. Given the theoretical importance of order information results for the item-order account, we now turn to the results of Experiment 1B which called upon a reconstruction of order task.

Experiment 1B

In Experiment 1B, we replicated the design of Experiment 1A, except that instead of using an immediate serial recall task, we used an immediate order reconstruction task. According to the item-order account and to previous results with long-term memory tasks, participants should better reproduce the order of silently read words than of words read aloud with pure lists, but not with mixed lists (Jonker et al., 2014; Forrin & MacLeod, 2016a; McDaniel & Bugg, 2008).

Method

Participants. Forty-eight students from the Université de Moncton (8 men, 40 women; mean age 20 years old), took part in this experiment, and received course credits for their participation. Participants were randomly assigned to one of the two list types (pure or mixed), with the restriction that 24 participants were assigned to each group. All participants were native

French speakers with normal or corrected to normal vision. None of the participants took part in the previous experiment.

Materials, Procedure and Design. Materials, procedure, and design were the same as those used in Experiment 1A, except than an order reconstruction task was used. Accordingly, immediately after the presentation of the last word of a list, all six items simultaneously reappeared on the screen. Words were displayed on two lines with three words per line. Words were presented in alphabetical order in black. Participants were asked to click on the items in their presentation order, starting with the first presented word. Once clicked, a word turned blue and could not be clicked again.

Results

As shown in the second row of Figure 1, overall, in mixed lists, produced items show a clear advantage over silently read items. However, this advantage vanished with pure lists. As in Experiment 1A, serial positions shown in Figure 2 (second row) revealed a systematic advantage of produced items in mixed lists producing sawtooth serial position curves with large reversals. With pure lists, the order of silently read items was better reconstructed on the first four serial positions, while the reverse was observed on the last two serial positions.

The 2 x 2 x 6 mixed-design ANOVA with list type (mixed or pure) as the between-participants factor and condition (silent reading or production) and serial position as repeated-measures factors revealed a main effect of condition, $F(1,46) = 62.96$, $p < .001$, $\eta_p^2 = .58$, and of serial position, $F(5,230) = 43.76$, $p < .001$, $\eta_p^2 = .49$, but the effect of list type did not reach significance, $F(1,46) = 1.71$, $p = .20$. All interactions were significant. More precisely, the interaction between list type and condition, $F(1,46) = 78.36$, $p < .001$, $\eta_p^2 = .63$, list type and

serial position, $F(5,230) = 11.30$, $p < .001$, $\eta_p^2 = .20$, and condition and serial position, $F(5,230) = 8.25$, $p < .001$, $\eta_p^2 = .15$s were significant. The three-way interaction between the list type, condition and serial position, $F(5,230) = 10.34$, $p < .001$, $\eta_p^2 = .18$, was also significant.

The three-way interaction was decomposed by means of two separate ANOVAs. For pure lists, the 2 X 6 repeated measures ANOVA with condition (silent reading or production) and serial position as factors revealed a significant main effect of serial position, $F(5,115) = 30.97$, $p < .001$, $\eta_p^2 = .57$, and an interaction between input modality and serial position, $F(5,115) = 32.53$, $p < .001$, $\eta_p^2 = .59$. The main effect of input modality was not significant, $F < 1$. Post hoc Tukey's HSD tests revealed that silently read items were better reordered at the first four positions, and that produced items were better reordered at the last position, all *p*s < .01. The advantage of produced items at Position 5 was not significant, $p = .0521$. For mixed lists, the ANOVA with list composition (PSPSPS vs SPSPSP) and serial position as factors revealed a significant main effect of serial position, $F(5,115) = 20.63$ $p < .001$, $\eta_p^2 = .47$, and an interaction between serial position and list composition, $F(5,115) = 82.01$, $p < .001$, $\eta_p^2 = .78$. There was no main effect of list composition, $F < 1$. Tukey's HSD tests revealed that produced items were significantly better reordered than silently read items at all serial positions, all *p*s < .001.

Discussion (Experiments 1A & 1B)

Experiments 1A and 1B were developed to test the item-order account of the production effect within an STM paradigm (Forrin & MacLeod, 2016a; Jonker at al., 2014; Lambert et al., 2016; McDaniel & Bugg, 2008). As a reminder, according to this account, producing the items enhances item-specific encoding at the expense of order information encoding. In immediate serial recall with pure lists, this was expected to translate into more item errors, but fewer order

errors for silently read items than for produced items. With the order reconstruction task mainly calling upon order information, better performance of silently read items was expected. Furthermore, in the mixed lists condition, order encoding of silently read items would be disrupted by the item-specific processing of produced items. This was expected to abolish or severely reduced the order advantage of silently read items with about the same number of order errors for silently read and produced items and a similar performance at the order reconstruction task.

Overall, results of Experiment 1A, with strict recall scoring, support the predictions of the item-order account and nicely extend previous results with free recall: There was a large production effect with mixed lists that was severely reduced with pure lists (Forrin & MacLeod, 2016a; Jonker et al., 2014; Lambert et al., 2016). However, the pattern of errors was more challenging for the item-order account. As predicted, participants made fewer item errors for items read aloud than for items read silently, and the effect was reduced with pure lists. However, it can be argued that the magnitude of the reduction was larger than predicted. Most importantly, the pattern of order errors diverged from the predictions. As expected, there were fewer order errors for silently read items than for produced items in pure lists. However, the advantage of silently read items was *larger* with mixed lists, while according to the item-order account, it should have vanished. Results of Experiment 1B with the immediate order reconstruction task are also problematic for the item and order account with no benefit of silently read items with pure lists and a large disadvantage with mixed lists.

In addition to only offering partial support to the item-order account, results of Experiment 1 revealed a surprising and remarkable pattern of results when serial positions were considered. While accounting for serial positions is outside the scope of the item-order account,

in light of our results, it would appear essential to consider serial position in order to understand the production effect in immediate memory as well as distinctiveness effects in STM more generally. In both experiments, with mixed lists, where aloud and silent words alternated, we observed a systematic advantage of produced items across the list, creating sawtooth serial position curves. To the best of our knowledge, we are the first to report this pattern in relation to the production effect in STM. Moreover, with pure lists, there was a disadvantage for items read aloud for the first serial positions, while there was also a sizable advantage for the last serial positions. We investigate the nature of the latter recency effect in Experiment 2.

Experiment 2

In short-term ordered recall with pure lists, it is well-established that the last item(s) of a list presented auditorily are better recalled than the corresponding items in a list presented visually, with no detrimental effect on the early positions (Corballis, 1966, Laughery & Pinkus, 1966, Penney, 1975, 1989). This advantage of auditory presentation over visual presentation for the immediate recall of the last serial positions, is known as the modality effect. In the context of the production effect, by reading the items aloud, participants generate an auditory presentation of the items. As mentioned above, with pure lists, the advantage of produced items over silently read items is limited to the last serial positions. Hence, results of Experiment 1 raised an important question: In short-term ordered recall, is the production effect just another name for the modality effect?

Crowder (1970) contrasted a condition in which participants read the items aloud with a control condition in which the experimenter read the items aloud. The recency effect was identical in both cases, but recall was lower on the first serial positions for the items read aloud by the participants themselves. Unfortunately, Crowder did not include a silently read condition

in his experiment. Consequently, it is impossible to know if saying the items aloud was detrimental compared to a visual presentation or if the concomitant auditory presentation provided by the experimenter was beneficial. With mixed lists, in a free recall and an item recognition (LTM tasks), results revealed an advantage of saying the items aloud over hearing them, even when the auditory recording consisted of the participant's previously recorded voice (Forrin & MacLeod, 2018, 2016b; MacLeod, 2011). However, in these studies the serial position curve was not reported. Therefore, it is impossible to know if the effect is systematically distributed across serial positions, and if the effect would generalize to immediate serial recall.

In Experiment 2, we tested the functional equivalence of the modality and the production effects. The design was the same as in Experiment 1A, except that a concomitant auditory presentation was used instead of asking participants to read the items aloud. If the production effect in short lists is just another name for the modality effect, in pure lists, on the last serial positions, items with a concomitant auditory presentation should be better recalled than the silently read items, and they should be less well recalled on the first serial positions. With mixed-lists, items with a concomitant auditory presentation should systematically be better recalled than the silently read items. This would produce a sawtooth serial position curve.

Method

Participants. Forty-eight students from the Université de Moncton (4 men, 44 women; mean age of 22 years old), took part in this experiment, and either received course credits for their participation, or a ticket for the draw of a $ 100 cash prize. Participants were randomly assigned to one of the two list types (pure or mixed), with the restriction that 24 participants were assigned to each group. All participants were native French speakers with normal or corrected to normal vision and none of them took part in the previous experiments.

Materials, Procedure and Design. Materials, procedure, and design were the same as those used in Experiment 1, except for the following changes. The production condition was changed to a concomitant auditory presentation. The auditory stimuli were produced with the French-Canadian male voice (Léo) from Text-to-Speech (http://www.oddcast.com). The auditory stimuli were recorded and edited with SoundTap 3.04 (https://www.nch.com.au) and Audacity 2.1.3 (https://www.audacityteam.org). Audio stimuli were presented through loudspeakers. The auditory presentation of a word lasted between .30 and .90 second. The presentation of the auditory stimuli was synchronized with the visual presentation of the words. The experiment was programmed with OpenSesame 3.1.9 (Mathôt et al., 2012), and all stimuli were presented in black in lowercase letters with a letter size of 30 points.

Results

Results were first analyzed with a strict serial recall criterion. As shown in Figure 1 (third row), compared to visually presented words, words with an additional auditory presentation were better recalled in pure and mixed lists and the advantage was larger with mixed lists. In the pure list condition, an inspection of the serial position curves presented in Figure 2 (third row), revealed an advantage of words with an additional auditory presentation for the last two positions with no disadvantage on the first positions. In the mixed lists condition, words with an additional oral presentation were systematically better recalled, resulting in a sawtooth serial position curve.

A 2 X 2 X 6 mixed-design ANOVA with list type (mixed or pure) as the between-participants factor and input modality (auditory presentation or silent reading) and serial position as repeated-measures factors was computed on the proportion of correct recall. The ANOVA revealed a main effect of input modality, $F(1,46) = 147.48$, $p < .001$, $\eta_p^2 = .76$, and serial

position, $F(5,230) = 33.67$, $p < .001$, $\eta_p^2 = .42$, but the effect of list type was not significant, $F < 1$. The interaction between list type and input modality, $F(1,46) = 54.89$, $p < .001$, $\eta_p^2 = .54$, and between serial position and input modality, $F(5,230) = 40.32$, $p < .001$, $\eta_p^2 = .47$, were significant. In addition, the three-way interaction between list type, input modality and serial position was significant, $F(5,230) = 2.77$, $p < .05$, $\eta_p^2 = .06$. The main effect of the list type, $F < 1$, and the interaction between the list type and input modality, $F(5,230) = 1.56$, $p > .05$, $\eta_p^2 = .03$, were both non-significant.

The three-way interaction was decomposed by means of two separate 2 X 6 repeated measures ANOVAs with input modality (auditory presentation or silent reading) and serial position as factors. For the pure lists, the ANOVA revealed a main effect of input modality, $F(1,23) = 18.36$, $p < .001$, $\eta_p^2 = .44$, of serial position, $F(5,115) = 22.29$, $p < .001$, $\eta_p^2 = .49$, and of the interaction, $F(5,115) = 29.98$, $p < .001$, $\eta_p^2 = .57$. Post hoc Tukey's HSD tests revealed that items presented auditorily were better remembered than items read silently at Position 5 and 6 (both *p*s <.001), while there was no significant difference at the first 4 serial positions (.097 < *p*s < .81). For the mixed lists, the ANOVA with serial position and list composition (ASASAS vs. SASASA) revealed a main effect of serial position, $F(5,115) = 13.45$, $p < .001$, $\eta_p^2 = .37$, and an interaction between input modality and serial position, $F(5,115) = 80.91$, $p < .001$, $\eta_p^2 = .78$, but no main effect of the list composition, $F < 1$. Tukey's HSD tests revealed that words with a concomitant auditory presentation were better recalled than silently read words at all serial positions, all *p*s < .001.

Item and order errors. As shown in Figure 3 (second row), participants made fewer item errors for the words benefiting from a concomitant auditory presentation, and the advantage was larger

in mixed than in pure lists. The 2 x 2 mixed-design ANOVA with list type (mixed or pure) and input modality (read silently or concomitant auditory presentation) as factors revealed a significant main effect of input modality, $F(1,46) = 225.30$, $p < .001$, $\eta_p^2 = .83$, and an interaction between list type and input modality, $F(1,46) = 97.23$ $p < .001$, $\eta_p^2 = .68$. There was no main effect of list type, $F = 1.21$, $p = .28$. Tukey's HSD tests revealed a smaller proportion of item errors for words with an auditory presentation than for word read silently in mixed ($p < .001$) and pure lists ($p < .001$). Furthermore, compared to mixed lists, in pure lists, the proportion of item errors was larger for words with an auditory presentation ($p < .001$) and smaller for words read silently ($p < .001$).

As shown in Figure 4 (second row), there were more order errors for items with an auditory presentation and this effect was larger with mixed lists. The 2 x 2 mixed-design ANOVA with list type (mixed or pure) and input modality (read silently or concomitant auditory presentation) as factors revealed a significant main effect of input modality, $F(1,46) = 61.08$, $p < .001$, $\eta_p^2 = .57$, and a significant interaction, $F(1,46) = 34.68$, $p < .001$, $\eta_p^2 = .43$, but the effect of list type was not significant, $F = 3.08$, $p = .09$. Tukey's HSD tests revealed a smaller proportion of order errors for words with an auditory presentation than for words read silently in mixed ($p < .001$) and pure lists ($p < .05$). Furthermore, compared to mixed lists, in pure lists, the proportion of order errors was larger for silently read words ($p < .01$), and smaller for words with an auditory presentation ($p < .001$).

Discussion

With pure lists, the concomitant auditory and visual presentation reproduced classic modality effect with a large advantage for the dual-modality items limited to the recency portion

of the serial position curve (Penney, 1975). The recency effect observed in Experiment 2 is very similar to the one observed in Experiment 1A. However, contrary to Experiment 1, there was no detrimental effect of the dual-modality presentation on the first serial positions. With mixed lists, as found in Experiment 1, we observed a large effect at all serial positions resulting in sawtooth serial position curves. Furthermore, error analyses revealed the same pattern of results as observed in Experiment 1.

Experiment 3

Experiment 2 was aimed at testing whether in immediate serial recall, a concomitant auditory presentation could generate the same pattern of results as observed with the production effect. The very high similarity between results of Experiment 1 and 2 cast doubts on the added benefit of producing the items ourselves. As a reminder, using only a mixed list condition, MacLeod (2011; Forrin & MacLeod, 2016b, 2018) observed a small but reliable advantage for saying the items aloud over hearing them, even when the auditory recording consisted of the participant's previously recorded voice. With pure lists, in immediate serial recall, Crowder (1970) reported an overall disadvantage of producing the items compared to a control visual condition with a concomitant auditory presentation. More specifically, Crowder observed the same recency effect in the read aloud and silently read condition with a concomitant auditory presentation conditions, but a large disadvantage for the items said aloud on the first serial positions. In Experiment 3, using an immediate serial recall task, we investigated whether there would be a production effect over and above a dual-modality presentation in both pure and mixed lists. This was achieved by contrasting a condition in which items were visually presented and participants were required to say them aloud with a condition in which items were visually presented with a simultaneous aural presentation of the words.

Method

Participants. Forty-eight students from the Université de Moncton (11 men, 37 women; mean age of 19 years old) took part in this experiment and received course credits for their participation. Participants were randomly assigned to one of the two list types (pure or mixed), with the restriction that 24 participants were assigned to each group. All participants were native French speakers with normal or corrected to normal vision and none of them took part in the previous experiments.

Material, Procedure and Design. Materials, procedure, and design were the same as those used in Experiment 1A, with the exception that items that were silently read in Experiment 1A now benefited of a concomitant auditory presentation. The auditory stimuli were the same as those used in Experiment 2, and their presentation was synchronized with the visual presentation of the words. The read aloud condition was identical to Experiment 1: Words were presented visually without a concomitant auditory presentation and participants were required to say them aloud. Therefore, in the pure list condition, for half lists, words were presented visually and auditorily, and for half lists, word were presented visually without a concurrent oral presentation and participants were asked to say them aloud. In the mixed list condition, three of the six words of a list were presented visually and orally, while three words were only presented visually, and participants were requited to say them aloud.

Results

As shown in Figure 1, in pure lists, items with a concomitant auditory presentation were slightly better recalled than produced items, while the reverse was observed in mixed lists. However, in the mixed lists condition, the overall advantage of produced items is much smaller

than what was observed in Experiment 1. Serial position curves in Figure 2 clearly show that the disadvantage of produced items in pure lists was due to the initial serial positions. There was no difference between the two conditions on the last three serial positions. For the mixed lists condition, produced items were better recalled than words read aloud on the first four serial positions and the effect reversed for the last serial position.

The 2 X 2 X 6 mixed-design ANOVA with list type (mixed or pure) as the between-participants factor and input modality (auditory presentation or production) and serial position as repeated-measures factors revealed a main effect of list type, $F(1,46) = 5.24$, $p < .05$, $\eta_p^2 = .10$, and of serial position, $F(5,230) = 67.69$, $p < .001$, $\eta_p^2 = .60$, but not of input modality, $F < 1$. All interactions were significant. More specifically, the interaction between list type and input modality, $F(1,46) = 4.53$, $p < .05$, $\eta_p^2 = .09$, list type and serial position, $F(5,230) = 2.81$, $p < .05$, $\eta_p^2 = .06$, input modality and serial position, $F(5,230) = 9.77$, $p < .001$, $\eta_p^2 = .18$, as well as the three-way interaction between list type, serial position and input modality, $F(5,230) = 15.90$, $p < .001$, $\eta_p^2 = .26$, were significant.

The three-way interaction was decomposed by computing separate repeated-measures ANOVAs for pure and mixed lists. For pure lists, the ANOVA revealed a main effect of serial position, $F(5,115) = 46.14$, $p < .001$, $\eta_p^2 = .67$, and an interaction between input modality and serial position, $F(5,115) = 5.25$, $p < .001$, $\eta_p^2 = .19$. The main effect of input modality was not significant, $F = 2.81$, $p = .10$. Post hoc Tukey's HSD tests revealed that the items with an auditory presentation were better remembered than the items read aloud at Position 1 ($p < .005$). However, there was no significant difference at the other 5 serial positions ($p = .051$, $p = .055$, $p = .23$, $p = .08$, $p = .11$, respectively). For mixed lists, the ANOVA revealed a main effect of input

modality, F(1,23) = 18.86, $p < .001$, $\eta_p^2 = .45$, of serial position, $F(5,115) = 26.93$, $p < .001$, $\eta_p^2 = .54$, and a significant interaction, $F(5,115) = 11.16$, $p < .001$, $\eta_p^2 = .33$. Tukey's HSD tests showed that produced items were better remembered than items presented auditorily at Position 1, 2, and 3 (all *p*s < .01); there was no significant difference at Position 4 and 5 ($p = .12$ and $p = .96$, respectively); a reverse effect was observed at Position 6 ($p < .001$).

Item and order errors. The pattern of item errors (see Figure 3, last row) mimicked overall results with strict serial scoring: fewer item errors for produced than for auditory items in mixed lists and the reverse in pure lists. The 2 x 2 mixed-design ANOVA with list type (mixed or pure) and input modality (produced or concomitant auditory presentation) as factors revealed a significant interaction between list type and input modality, $F(1,46) = 4.68$ $p < .05$, $\eta_p^2 = .09$, but neither the main effect of list type, $F = 3.92$, $p = .054$, nor of input modality, $F < 1$, was significant. Tukey's HSD tests revealed a smaller proportion of item errors for words with an auditory presentation than for the silently read ones in pure lists ($p < .001$), but the reverse trend was not significant in mixed lists ($p = .27$).

Inspection of Figure 4 (last row) reveals that order errors were uniformly low. Accordingly, the 2 x 2 mixed-design ANOVA with list type (mixed or pure) and input modality (produced or concomitant auditory presentation) as factors revealed no main effects or interaction, all *F*s <1.

Discussion

Results of Experiment 3 nicely extend to STM previous findings observed with long mixed lists and item recognition or free recall tasks, as the results show a beneficial effect of saying the items aloud over hearing them (Forrin & MacLeod, 2016b, 2018; MacLeod, 2011).

Furthermore, although attenuated, the sawtooth serial position curves discovered in the first two experiments were observed again. With pure lists, as observed in Experiment 1, there was a disadvantage of saying the items aloud on the early serial positions (see also Crowder, 1970). Moreover, in pure lists, isolating the effect of production, showed that the recency advantage for produced items appears identical to the one obtained in the traditional modality effect; this bring into question the idea that produced items benefit from the presence of additional features, relative to audio-visual items that are not produced.

Experiment 4

Thus far, we have established that, within mixed lists, produced words are far better remembered than words read silently, and are better remembered than words presented both auditorily and visually. Within pure lists, for the first serial positions, words that were read silently and words that benefited from a concomitant auditory and visual presentation were better remembered than produced words; for the last serial positions, there was an advantage for produced words and for words presented auditorily compared to words that were read silently. We argue that the advantage of production for the last serial positions comes from the auditory features that are present during pronunciation, and that this pronunciation affects the recall for the first serial positions. However, it is difficult to make these assumptions based solely on Experiments 1 through 3, because some of the critical contrasts can only be done across experiments. Therefore, in Experiment 4, a production condition, a concomitant auditory and visual condition, and a silent visual condition were used within participants. Based on previous experiments, there should be an advantage in the recall for produced words and words presented auditorily compared to words presented visually for the last serial positions, and a disadvantage for produced words compared to the other two conditions for the first serial positions.

Method

Participants. Twenty-four students (3 men, 21 women; mean age of 19 years old) from Université de Moncton took part in this experiment and received course credits for their participation. Participants were native French speakers and had normal or corrected to normal vision. None of them took part in the previous experiments.

Materials. One-hundred and eighty-six French words containing two phonological and orthographic syllables were added to those used in the first three experiments, creating a word pool of four-hundred and eighty-six words. Overall, words had a frequency ranging from 5.07 to 497.43 occurrences per million (M = 36.11, SD = 59.45), according to Lexique 3 (New et al., 2004). The auditory stimuli were generated following the procedure described in Experiment 2. The auditory stimuli had a duration ranging between .30 and .93 second. On the screen, words were presented in lowercase letters with a size of 24 points. Eighty-one lists of six words were generated by drawing without replacement from this pool 486 words, with the constraint that no words within a list could rhyme. These 81 lists were split into 3 sets of 27 lists each. Three lists of each set were randomly selected for practice trials and the remaining 24 lists were used for experimental trials. The experiment was controlled with the program E-Prime 2.0 (Psychology Software Tools, 2016).

Design and Procedure. Overall, each participant underwent recall with three different presentation modalities. Items were either read silently, read silently with a concomitant auditory presentation, or read aloud. Only pure lists were tested. Each participant underwent 3 blocks of 24 experimental trials. Across all participants, the order of presentation of the conditions was systematically counterbalanced. In addition, each list was presented equally often in each

presentation condition. The order of the lists within each condition was randomized for each participant, but word order within a list was the same for all participants.

Stimulus presentation was the same as in Experiment 1, except that oral instead of written recall was used. Participant were instructed to say «passe» [pass] if they forgot a word at a given serial position. Responses were digitally recorded for subsequent coding.

Results

The proportion of correct responses as a function of serial position is shown in Figure 5. The figure reveals the presence of a classic modality effect with better recall performance for the last two serial positions presented auditorily compared to the items read silently. Items produced out loud had the same advantage as items presented aurally for the last serial positions. However, for the first serial positions, the items produced out loud were not recalled as well as the items read silently and as the items benefiting from a concurrent auditory presentation.

A 3 X 6 repeated-measures ANOVA with input modality (silent reading, auditory presentation, and production) and serial position as factors revealed a main effect of input modality, $F(2,46) = 7.52$, $p < .01$, $\eta_p^2 = .25$, and of serial position, $F(5,115) = 14.92$, $p < .001$, $\eta_p^2 = .39$, as well as an interaction, $F(10,230) = 8.05$, $p < .001$, $\eta_p^2 = .26$. Post hoc Tukey's HSD analyses revealed that items read silently and items presented auditorily were better remembered than produced words at Position 1, 2 and 3, all *p*s < .005. The three input conditions did not significantly differ at Position 4 and 5. At position 6, produced words and word presented auditorily did not differ ($p = .70$), but they were better recalled than silently read words, both *p*s < .001.

Item and order errors. Item and order errors were also computed. There were fewer items errors for words presented auditorily (*M*= .56, *SD* = .20), than for silently read (*M*= .64, *SD* = .19) or produced words (*M*= .62, *SD* = .19). Accordingly, the repeated-measures ANOVA was significant, $F(2,46) = 13.17$, $p < .001$, $\eta_p^2 = .36$, and Tukey's tests revealed a significantly lower proportion of item errors in the auditory than in the produced condition ($p <. 005$) or silent condition ($p < .001$). The latter two conditions did not differ one from the other ($p = .28$). Participants made more order errors for produced items (*M*= .12, *SD* = .14), than for the auditorily presented items (*M*= .11, *SD* = .14) and the silently read items (*M*= .07, *SD* = .12). The repeated-measures ANOVA was significant, $F(2,46) = 9.41$, $p < .001$, $\eta_p^2 = .29$, and Tukey's tests revealed a significantly lower proportion of order errors in the silent condition than in the produced ($p <. 001$) or the auditory condition ($p < .01$). The latter two conditions did not differ ($p = .52$).

Discussion

Results of Experiment 4 nicely confirmed those observed in the first three experiments. Produced items generated a modality-like effect. In fact, recall performance for the production and dual-modality conditions was undistinguishable on the last serial positions. On the primacy portion of the serial position curve, again, produced items suffered compared to silently read items or items benefiting from a dual-modality presentation.

Experiment 5

The first four experiments left us with a question. How could we account for the disadvantage of produced words on the first serial positions in pure lists? One idea is that in the production condition, pronouncing aloud each word interferes with rehearsal (see Routh, 1970,

for a similar idea). The suggestion is that items located at the first serial positions ordinarily are rehearsed more frequently and so they suffer the most from rehearsal interruptions due to overt articulation of the presented items (Bhatarah et al., 2009).

To test this rehearsal suppression hypothesis, Experiment 5 recreated the classical production effect task in pure lists, as in Experiment 1A, and added a component of articulatory suppression for the silent condition. More specifically, in the silent condition, after reading each word, participants were required to say aloud an irrelevant word. Consequently, the advantage for words read silently for the first serial positions should be abolished, with a recall for produced words equal to or greater than the recall for words read silently.

Method

Participants. Twenty-four students (2 men, 22 women; mean age of 21 years old) from Université de Moncton took part in this experiment. They also received course credits for their participation. All participants were native French speakers and had normal or corrected to normal vision, and none of them took part in the previous experiments.

Materials, Procedure and Design. Materials, procedure, and design were the same as those used in Experiment 1.A, except for the following changes. All lists presented were pure lists, with presentation modality (read silently vs. read aloud) as the repeated factor. Within the silent condition, the instruction stipulated that participants had to say the word “mathématiques” [mathematics] simultaneously as the words appear on the screen. Therefore, since there are 6 words in the list, the participants had to repeat “mathématiques” 6 times.

Results

The proportion of correct responses as a function of serial position is shown in Figure 5. The figure reveals two distinct serial position curves, with overall better performance for the produced words than for the words read silently.

A 3 X 6 repeated-measures ANOVA with input modality (silent reading and production) and serial position as factors revealed a main effect of condition, $F(1,23) = 62.50$, $p < .001$, $\eta_p^2 = .73$, of serial position, $F(5,115) = 38.31$, $p < .001$, $\eta_p^2 = .62$, and an interaction, $F(5,115) = 4.01$, $p < .01$, $\eta_p^2 = .15$. Post hoc Tukey's HSD analyses revealed that produced items were better recalled than silently read items at all serial positions, all *p*s < .005.

Item and order errors. Item and order errors were also computed. There were significantly more item errors for silently read words (M= .68, SD = .14) than for produced words ($M = .47$, $SD = .18$), $F(1,23) = 121.68$, $p < .001$, $\eta_p^2 = .84$. The reverse pattern was observed with order errors. There were significantly fewer order errors for silently read words ($M = .08$, $SD = .12$) than for produced words ($M = .13$, $SD = .16$), $F(1,23) = 19.48$, $p < .001$, $\eta_p^2 = .46$.

Discussion

In this experiment, our purpose was to test the presence of a rehearsal suppression effect of produced words on the first serial positions in the serial recall of pure lists of words. Results revealed that the presence of an interfering articulatory suppression task abolishes the advantage for words read silently compared to produced words for the first half of the serial positions. This provides further evidence that the recall disadvantage in the first serial positions present in pure lists for the words read aloud compared to other presentation modalities is linked to the presence of rehearsal suppression, which interferes with performance (see also Routh, 1970). Importantly,

if we think of rehearsal suppression as equating strategic processing across the silent and produced conditions, Experiment 5 suggests that there is a positive effect of production, relative to silently read items, across all serial positions. Below we review the main results so far and consider the mechanisms that can account for the findings. We then introduce a computational model as a means of testing the value of the proposed processes when combined and made explicit.

A Computational Model

In summary, the results above suggest an account that includes the following elements:

1) Relative distinctiveness appears important, in the form of more information being stored for produced and auditorily presented items relative to visually presented / silently read ones. This is most clearly seen from the results of Experiment 5, where produced items are recalled more accurately than silently read ones at all positions, once rehearsal possibilities are better equated in both cases. Local relative distinctiveness springs to mind also when considering the saw tooth pattern produced by mixed lists in Experiments 1A, 1B, and 2.
2) For pure lists, the results also point toward a 'modality effect' related to produced items, arguably caused by relative distinctiveness (element 1 above). This modality effect extends to the final 2-3 items in the list. This is most clearly seen in Experiment 1A, but is also apparent in Experiment 4, where, for the last few items, recall of produced items is indistinguishable from audio-visual items (presented both visually and auditorily), while both are boosted relative to silently read ones.
3) Rehearsal also appears to play a role, improving recall, particularly for the first few items in the list, and it operates less effectively for produced items than for auditorily presented and

silently read ones. This is clearly suggested in Experiment 4, and Experiment 5 shows that this can be reversed by introducing articulatory suppression for the silent reading condition.

While the experimental results are compelling, the elements listed above are verbal stories we can tell about the patterns of data, and as such have limited explanatory or predictive value on their own. Further, each is essentially an explanation for a subset of the experimental results, or in some cases a subset of the data from some of the experiments. It is far from clear that we can combine these elements together in a model which can explain all the data reported above.

We are not aware of any existing model of immediate serial recall which includes these elements and which could plausibly account for our data. Our challenge in what follows is to provide a proof of principle that a model including elements 1-3 above can account for the key experimental results reported here, not just in a qualitative way but matching quantitatively the data in these experiments. Our model may not be the only one possible that could account for the data, in particular it is not the only possible way to instantiate these assumptions, but we hope to convince the reader it is at least plausible.

In constructing a model of immediate serial recall that could account for the effects we have documented, there were two possible approaches that we could take. The first was to try and construct something from scratch, tailor-made to the specific task at hand. The second was to adapt an existing model to include any missing elements. We chose the latter approach, since, as we will see, there is an existing model which in many ways is already close to the account we seek. By being mindful of the way in which we modify this basic model, we can also ensure that it will still account for other well-known effects in serial recall, at least in principle.

The Model

The model we adapted is called the Feature Model (Nairne, 1988b; 1990, Neath & Nairne, 1995; Neath & Surprenant, 2007). Many of the characteristics of this model were maintained in the current implementations, but a few important changes were also included. As a result, we identify the modified version as the Revised Feature Model (RFM). We now turn to a brief description of the main characteristics of the model. When describing elements that are unchanged, we refer to the Feature Model (FM).

In the FM, 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 font size or voice quality. On the other hand, items would also produce modality-independent features, generated by internal processes of categorization and identification. This characteristic of the original FM naturally lends itself to modelling differences related to presentation modalities such as those involved in the experiments reported here.

Furthermore, within the FM, items simultaneously generate traces in primary and secondary memory. In both cases, items are represented by vectors of features, with each (randomly generated) feature typically taking on one of a small number of values. Traces in primary memory are subject to degradation through overwriting by the following item. This retroactive interference process is similarity-based: in the FM if feature 5 of item « n » is identical to feature 5 of item « n-1 », then this feature of item « n-1 » will be overwritten (set to 0). In contrast, representations in secondary memory are assumed to remain intact. After presentation of all items, a final overwriting of modality-independent features only takes place due to continuing internal thought activity in preparation for recall. In the RFM, we retain this idea of overwriting, but we use a modified overwriting process; 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 with probability $e^{-\lambda(m-1)}$. If $\lambda \to \infty$ we recover the overwriting process of the original FM. This change has been implemented to allow retroactive interference to operate further back than just the most recently presented item, because the experimental data shows a modality effect that extends further than the final item in the lists.

If overwriting degrades traces in primary memory, in the RFM we assume 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 exponential comes from previous work suggesting a significant drop in rehearsal for lists longer than four items (Bhatarah et al, 2009).

Item 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 to the default value from Neath and Surprenant (2007).

At the point of recall, the correspondence between each degraded primary memory trace and the set of relevant traces in secondary memory is computed. The secondary memory trace with the highest ***relative*** **similarity** to the primary memory trace being considered has the highest probability of being recalled. In the RFM, we have updated the selection rule. While the original

FM called upon a Luce-style choice rule (cf. Luce, 1963), the RFM uses a more general soft-max rule, as follows:

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

Where the conditional probability that the secondary memory trace SM_j will be sampled, given the primary memory trace PM_i , depends on *s(i,j)*, the computed similarity between PM_i and SM_j and τ is a 'temperature parameter'. The lower the value of τ the more deterministic the choice of item with the highest relative similarity.

We also allow for the possibility that no secondary memory trace matches the primary memory trace well enough to be recalled. Omissions are the most frequent error when items are not repeated across trials, so it is important that the model can produce these sorts of errors (Osth & Dennis, 2015). We do this by including an extra 'null' possibility which has constant similarity between itself and all primary memory traces. If this 'null' response is selected, the model makes an omission. We describe how to compute this 'floor' similarity below.

Similarity is related to the feature-to-feature correspondence between primary and secondary traces. This relation calls upon a calculation of the psychological distance between the two traces, which is based on a function described by Shepard (1987):

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

This distance, d_{ij}, is simply calculated by adding the number of mismatched features, M, and dividing by the number of compared features, N, as in:

$$d_{ij} = \frac{a}{N}\sum b_k M_k$$

Where *a* is a scaling constant, and b_k is an attention bias parameter that is set to 1 for all simulations here (and hence has no influence on results). The 'null' possibility (chiefly responsible for producing omissions) is assigned a similarity equal to that expected between two independently chosen vectors of features, which for the set of features described here is approximately $s(null, j) = 1.7\times10^{-2}$.

Finally, output interference is included in the model, by assuming recovery from the SM set is related to prior recall of the item, in the following manner:

$$P_r = e^{-cr}$$

Where « *c* » is a scaling constant and « *r* » is the number of times a sampled item has already been recalled.

The inception of the FM was motivated by an attempt to account for the modality effect, observed in immediate serial recall (Corballis, 1966). The modality effect refers to the fact that recall is superior when lists are presented auditorily as opposed to visually. In fact, the difference is attributable to a greater recency effect in the case of the auditory modality (Conrad & Hull, 1968; Murray, 1966; Nairne, 1988b). To account for the modality effect, the FM assumes that auditory traces have a greater number of modality dependent features than do traces generated through visual presentation. Nairne (1990) noted that this assumption is consistent with the literature indicating that interference based on visual characteristics is not typically found within immediate serial recall performance. He also noted that the evidence for speech-like coding in primary memory with visual presentation is extensive and suggests that, without auditory cues, we tend to rely on modality-independent (inner voice) features to represent items. Hence, in the

FM, it is the greater number of modality dependent features, associated with auditory presentation, that leads to an auditory advantage for the recency position. This occurs because the model posits that the last item of a list is followed by internally generated activity, which overwrites modality-independent features while leaving modality-dependent features intact. The recall of the last item from an auditorily presented list will benefit from these intact modality-dependent features. The reason for this is that the intact features serve to increase the correspondence between the degraded primary memory trace of the last item and its secondary memory counterpart. Conversely, because visually based traces do not have many modality-dependent features, there is a smaller recency effect for these items.

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 at https://osf.io/tkc4p/. Our general approach was to fit all data from a single experiment at once, which is more challenging for any model than allowing all parameters to vary between 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 were the distance scaling

parameter, a, which controls the overall ease of matching cues to items, λ, which controls how many items can be affected retroactively by overwriting, r, which controls how effectively rehearsal can restore overwritten features, and finally τ, which controls how deterministically the secondary trace with the greatest similarity to the primary cue is chosen. Of these parameters, a and τ are less interesting theoretically, since they are not assumed to vary with modality.

For all fits, we report the results in two different ways. First, we show the posteriors of the model predictions for each serial position in the different conditions, which is the standard notion of the model 'fit'. However, secondly, we take the estimates for the best fitting parameters (technically the medians of the posterior estimates for each parameter) and use these to simulate some data which we can compare with the observed data. This latter approach is conceptually simpler (it shows that there is a single set of parameters that can produce a good match to the data) and it can help us to identify any more subtle issues with model fitting.

Since, as mentioned, we are not aware of any other models which could account for all the effects seen in our experiments, 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 the crucial Experiments 1A, 4, and 5. However, we can increase our confidence in our explanations for the data by also fitting a version of our model where we either fix the number of features to be the same for all item types, thus examining the role of relative distinctiveness in accounting for the data, or by demanding that rehearsal rates be fixed for different list types, thus examining the role of differential rehearsal success in accounting for the data. These reduced versions of the model can be used to confirm our intuition about which aspects of our model are important in accounting for different patterns in the data.

Experiment 1A

Our first aim was to show that our model can successfully account for the various effects that we reported in immediate serial recall. The key experiment in this regard is Experiment 1A, which demonstrates two important effects which we need to capture. Firstly, while there is a small overall advantage for produced over silent items in pure lists, the clear pattern in the serial position curve is of a crossover, where earlier on silent items have an advantage, but for later items this flips to an advantage for produced items. Secondly, for mixed lists, there is a distinctive see-saw pattern, with a large advantage for produced over silent items within the same mixed list. We wish to show that these patterns can be explained by relative distinctiveness and differential rehearsal. We implement relative distinctiveness by assuming that produced items have 20 modality-dependent features, compared with 2 for silently read ones.

To fit these data, we assumed that the values of λ and τ are the same for all conditions, but that the value of a may differ for pure and mixed lists, and the value of r can differ between pure produced, pure silent, and mixed lists. The justification for letting the rehearsal rate vary in this way is essentially that we assume production reduces the ability of participants to rehearse previously presented items. This is potentially an even stronger effect for mixed lists, given the extra complexity of the task.

Our intuition is that the distinctive patterns in the data are the result of relative distinctiveness. We can explore this by also fitting a version of the model where we fix the number of features to be the same for all item types (we do this by assuming that the number of modality-dependent features for produced items is also 2). We expect that this reduced model will fail to capture the distinct qualitative trends in the data.

In Figures 6 and 7, we show the results of the model fit. Overall, the model does a good job of fitting the data, and in particular the two key effects that we sought to reproduce — crossover in the serial position curves for pure lists, and a zigzag pattern for mixed lists— are reproduced well. The only systematic errors seem to come from the final item, particularly in the pure produced list, where the data indicate much better recall for the final item than the model expects. Parameter posteriors are shown in Figure 8: We can clearly see that rehearsal is somewhat less effective in produced lists than silent lists, but that it is essentially absent in mixed lists. In contrast, the reduced model, without the implementation of relative distinctiveness, is totally unable to capture even the qualitative trends.

Overall fits to Experiment 1A give us confidence that our model can capture the essential aspects of the production effect in pure and mixed lists, and that this is largely due to the inclusion of an implementation of relative distinctiveness. This therefore provides support to our proposed explanation for the effect in terms of relative distinctiveness, modality, and rehearsal.

Experiment 4

The logic of Experiment 4 was to compare silently read and produced items with auditorily presented ones. In terms of the model, items that are presented auditorily have almost as many features as produced items, but have the same rate of rehearsal as silently read ones. We implemented this by assuming that auditorily presented items have 15 modality-dependent features (compared with 20 for produced and 2 for silently read ones.) We also allowed the rehearsal parameter to vary between produced and silently read items, as for Experiment 1A, but fixed the rehearsal rate to be equal for silently read and auditorily presented items. We also fixed $a, \lambda,$ and τ to be the same in all conditions, and the three conditions were fit simultaneously.

Experiment 4 tests the key hypothesis that in pure lists the difference in recall performance for produced items over silently read ones can be split into two distinct parts. First, there is an advantage for silent over produced items at the start of a list due to suppression of rehearsal when items are produced. Second, there is an advantage for produced over silently read items at the end of a list due to a modality effect of higher distinctiveness caused by produced items possessing a larger number of features.

We can test this hypothesis by also fitting two different versions of the model. Firstly, there is one where rehearsal rates are fixed for all item types, which ought to remove the disadvantage for produced over silent and auditory items at the start of the list, but leave the modality effect unchanged. Secondly, there is one where the number of features is set equal across item types (implemented by setting the number of modality-dependent features to zero, as for Experiment 1A), which ought to remove the disadvantage for silent over produced and auditory items at the end of the list, but leave the disadvantage for produced over silent and auditory items at the start of the list unchanged.

Results are presented in a similar way as for Experiment 1A. In Figure 9 we compare posterior predictions of the full model with the data in each condition separately, while in Figure 10 we compare the data and simulations based on best fit model parameters (again, medians of the parameter posteriors) for the full and restricted versions of the model side by side so we can compare the qualitative patterns. For the full model, there does not appear to be any systematic misfitting, trends for the individual conditions are reproduced well, and the overall relation between conditions is reproduced, with the auditory condition matching the silent condition well at the start of the list, and the produced condition well toward the end of the list. The restricted

versions of the model behave in the expected way, showing that recall performance at the beginning and end of the list are driven by substantially independent effects.

Parameter posteriors for the full model are shown in Figure 11; as for Experiment 1A, we can clearly see that rehearsal is somewhat less effective for produced lists than for silent or auditorily presented lists. Overall fits and parameter estimates for Experiment 4 give us confidence in our hypothesis that the advantage for silent over produced items at the start of a list can be explained by differential rehearsal rates, and the advantage for produced over silently read items at the end of a list can be explained by a modality effect.

Experiment 5

Experiment 5 sought to test the hypothesis that reduced recall for produced over silently read items early in a list results from reduced ability to rehearse previous items. By introducing a form of articulatory suppression (saying the word "mathématiques" once between item presentation) in the silent list, we were able to extinguish the advantage for silently read items earlier in the list.

We implement this in the model fitting simply by allowing the rehearsal parameters to differ between produced and silent conditions, as we have done for previous experiments. However, our expectation is now that, unlike in the previous experiments, we will not see a higher value of this parameter for the silent condition than for the produced condition. The other three parameters, a, λ, τ, were fixed to be the same in both conditions, and both conditions were fit simultaneously.

In a similar way to the other two experiments, we also fit a restricted version of the model where the rehearsal rates are fixed to be equal. We expect that this restricted model will not be able to reproduce the separation in performance at the start of the list.

Results are presented in a similar way as for Experiments 1A and 4. In Figure 12 we compare posterior predictions of the model with the data in each condition separately, while in Figure 13 we compare the data and simulations based on best fit model parameters (again, medians of the parameter posteriors) for the full and restricted models side by side so we can compare the qualitative patterns. For the full model, there does not appear to be any systematic misfitting, trends for the individual conditions are reproduced well, and the overall relation between conditions is reproduced, with the silent condition now having a lower rate of correct recall at all positions. The restricted version of the model behaves as expected, with identical recall rates for produced and silent items at the start of the list.

Parameter posteriors are shown in Figure 14; as expected, we can clearly see that rehearsal in the silent condition is much less effective than for produced lists, a reversal of previous results. Overall fits and parameter estimates for Experiment 5 give us confidence in our hypothesis that the advantage for silent over produced items early in a list is due to enhanced rates of rehearsal for silent items, and that disrupting this removes the advantage compared to produced items.

General Discussion of the Model Fits

Overall, our model incorporating relative distinctiveness and rehearsal accounts well for the data across the three experiments that are particularly important for probing these hypotheses.

Some misfitting is evident in some places, but the model captures all the qualitative trends, and is quantitively very close to the observed data for most conditions and list positions.

Importantly, by comparing the full model with restricted versions that remove relative distinctiveness or differential rehearsal rates we can clearly see the importance of these two mechanisms on recall at the end and start of the list respectively.

Another reassuring result of the model fitting is that the parameters we are less immediately interested in, the distance scaling parameter for pure lists, a, along with λ and τ, are estimated to be broadly comparable across the three experiments.

The purpose of the model fitting was to prove that our general hypotheses about the mechanisms responsible for the production effect can be turned into precise mathematical proposals and implemented in a concrete model of serial recall. There are likely other possible ways of implementing these ideas, and we make no claim that our approach is unique, or even best (we did not compare our model with an alternative.) Nevertheless, proving that these mechanisms can coexist in a single fully specified model which captures the data well is a significant achievement.

General Discussion

This study initially aimed to systematically explore the production effect in the context of a short-term memory task. Originally, we pursued two aims. First, we wanted to delineate the nature of the production effect in memory over the short term. Second, we wished to test the item-order account of the production effect by using paradigms well-known for calling upon item and order information. As results unfolded, new and striking findings emerged: for both pure and mixed lists, the production effect interacted with serial positions. After exploring these

interactions, we developed a revised version of the Feature Model to account for the full set of data and we tested this new version with the findings from the three most critical experiments. Before we turn to the theoretical implications of our results, let us summarise the main findings.

The five experiments yield a coherent pattern which can be summarised as follows. Firstly, our results suggest that adding a relevant dimension to item processing at the point of study improves performance in mixed alternating lists, leading to sawtooth serial position curves. With an immediate serial recall task, items read aloud and those benefiting from an audio-visual presentation were systematically better recalled than silently read items, and produced items were better recalled than those benefiting from an audio-visual presentation. In addition, the size of the benefit is proportional to the number of dimensions added, with the contrast between produced and silently read items generating a larger effect than the contrast isolating the auditory features (audio-visual vs. visual) or the production-related features (produced vs. audio-visual). The assumption underlying these statements is that for verbal content, a visual presentation accompanied by silent reading generates fewer useful, distinctive features than a visual presentation accompanied by a concomitant auditory one, i.e., auditory features are helpful. Likewise, reading aloud a visually presented item would generate more useful modality related features than the audio-visual presentation – so the unique features associated with overt production are also helpful (Forrin & MacLeod, 2016b, 2018; MacLeod, 2011).

The second new finding was that, in pure lists, immediate recall of produced items gave rise to a recency advantage relative to silently viewed items. This recency advantage did not differ from the modality effect, i.e., the recency advantage that audio-visual items showed relative to silent items. In other words, the recency advantage of produced items was indistinguishable from the recency advantage associated to audio-visual presentation. These

recency effects were not limited to the last serial position; they were observed for the last two or three serial positions.

Thirdly, again in pure lists, producing the items hindered immediate recall for the first serial positions. The cost for the first serial positions was not due to the auditory features, because there was no cost for the audio-visual presentation compared to the visual/silent presentation. We hypothesized that pronouncing the items disrupts rehearsal, which is more important for the first items in the list (Bhatarah et al., 2009). This hypothesis is reminiscent of Routh's (1970) suggestion that the "comparison of the effects of vocal and silent monitoring on the serial recall of a visual stimulus, presented element-by-element, is potentially confounded by differences in the intra-presentation opportunities for selective attention and rehearsal." (p.355). It is worth noting however, that this disadvantage for produced items has not been systematically found in older studies. In support of Routh's suggestion, with a presentation rate of two seconds per word, promoting rehearsal in the silent condition, Kappel et al. (1973) reported a disadvantage on the first serial positions. However, with a presentation rate of two digits per second, limiting rehearsal opportunities, neither Conrad and Hull (1968), nor Nairne and Walters (1983) found a disadvantage. In Experiment 5, when rehearsal was hindered in the silent condition through pronunciation of an irrelevant word, produced items were better recalled than silently read items at all serial positions. These results nicely extend those of Routh (1970) who contrasted a written and an oral production condition by assuming that writing the items would also block rehearsal. Routh's results revealed better recall in the oral than in the written production condition at all serial positions. We are in agreement with Routh's original suggestion and our findings are also in line with the insights from Bhatarah et al. (2009): When presentation time is sufficient to allow for rehearsal, producing items will hinder said rehearsal and have a

detrimental effect on memory for the items most rehearsed. If rehearsal is equated through some means, then produced items will be better remembered.

Importantly, taken together our findings show that the production effect in STM is different in nature from the modality effect (Conrad, 1964). The latter is related to the advantage for the last serial position(s) that is observed when items are presented aurally, relative to visually (silently read). Here, we found that for pure lists, producing the items did indeed generate a modality effect, but it also had a negative impact on the primacy positions. Moreover, in alternating mixed lists, producing the items had an impact that was greater than that of items that were presented in an audio-visual modality. It follows that the production effect in STM cannot be equated with the classic modality effect; other mechanisms are at play. Here we argue that these include the encoding of features that are unique to production as well as an impact on the time or resources available for rehearsal. We now turn to how these findings inform current explanations of the production effect before considering the other insights associated to the present empirical and modelling work.

The item-order account of the production effect in STM

In addition to reporting performance as a function of serial positions, we analysed errors. The error analyses along with the inclusion of an order reconstruction task were aimed at testing the item-order account of the production effect in immediate recall (e.g., Jonker et al., 2014; McDaniel & Bugg, 2008). According to the item-order account, it is assumed that item-specific and order/relational information compete for encoding resources; moreover, it is assumed that recall is guided by order/relational information. Relative to silently read items, produced items would benefit from more item-specific processing due to the presence of additional motor and auditory features. However, the item-specific processing would come at a cost: It would disrupt

order encoding. For pure lists, the implication is that produced items have better item-specific encoding and poorer relational information while the reverse is true for silently read items. In mixed lists, produced items are thought to disrupt the relational encoding of silent items, while silent items support better order encoding for produced items – on balance, this is supposed to lead to approximately equivalent order encoding for both types of items. This would lead to superior recall of produced items as the better item-specific encoding is no longer counteracted by poorer order encoding.

There were a series of predictions derived from the item-order account for the tasks called upon here. More specifically, in the immediate serial recall task, pure lists of produced items were expected to lead to more order errors than pure lists of silently read items. With mixed lists, order errors were expected to be better equated, with an increase for silently read items and a reduction for produced items. In the case of the order reconstruction task, order encoding effects were expected to be highlighted – i.e., pure silently read lists were expected to fare much better than pure produced lists. With mixed lists –relative to pure lists– a drop in performance was expected for silently read items while produced items would see performance improve.

Results with the order reconstruction task provided no support for the item-order account. Contrary to the predictions, in mixed lists, there was a large advantage of produced items and in pure lists, silently read items did not exhibit an overall advantage. This pattern of results is contrary to what has been previously observed (Jonker et al., 2014). However, the analysis of the serial position curves suggests that these inconsistencies may be more apparent than real. In effect, in pure lists, produced items were less well reordered for the first serial positions, but this cost was masked by a large benefit of produced items on the last serial positions, which is

analogous to the modality effect. In Jonker et al.'s experiment, after list presentation, participants engaged in a 30-second distractor task before performing the order reconstruction task. Importantly, Crowder (1970) showed that a simple suffix reduced recency for produced items but left the silently read ones unaffected. Therefore, the lower order reconstruction performance reported by Jonker et al. for pure produced lists could be entirely driven by the first serial positions, attributed here to production interfering with rehearsal (see also, Routh, 1970). The analysis of Experiments 1, 3, and 4, which used an immediate serial recall task without interference, revealed the same pattern of order errors: There were fewer order errors for silently read items than for produced items and the same pattern was observed for pure and mixed lists. For item errors, in Experiment 1, there were more errors for silently read than for produced items in pure and mixed lists, but there was no difference in pure lists in Experiment 4. As can be seen, the error analysis results only offer partial support to the item-order account. While previous error analyses in free recall offered better support to the item-order account (Forrin & MacLeod, 2016a; Jonker et al., 2014), other studies did not confirm the predictions of the view (Lambert et al., 2016). Given the limited support found for the item-order account and considering that it cannot deal with the sawtooth serial position curves for mixed lists reported here, it is difficult to see how this view can account for the production effect as studied here.

Limitations and further research – STM, LTM, and the Production Effect

Our results with short-term ordered recall tasks extend previous production effect findings reported with long-term memory tasks. Some of the patterns reported here mirror those obtained with tasks classically associated with long-term memory (e.g., free recall), while others are new and may only apply to STM settings. With pure lists, we observed a small beneficial effect of produced items over silently read items in Experiment 1A, with an immediate serial

recall task, but not in Experiment 1B with an order reconstruction task, nor in Experiment 3 and 4 with an immediate serial recall task. This small and evanescent production effect is also the norm in free recall (Forrin & MacLeod, 2016a; Jones & Pyc, 2014; Jonker et al., 2014; MacLeod & Bodner, 2017). Here, however, it was revealed that there is a crossover interaction between serial position and the production effect, such that production was associated with a positive modality effect for the last serial positions, and a rehearsal disadvantage for the first serial positions. It remains to be established whether similar patterns would also be found with free recall, and whether the same explanations we called upon here would be useful in an LTM context. For instance, it remains to be determined that the interaction of production with serial positions and rehearsal can be established for tasks such as free recall. Moreover, it is not clear whether the dramatic sawtooth patterns observed here would also be observed with LTM recall. Finally, should these effects be reproduced, more research would be needed to establish if the same types of processes are at play and if the RFM can be extended to account for data from tasks such as free recall.

The similarities between our overall findings and those previously reported with free recall were sufficient however to lead us to consider theoretical explanations of the production effect that are associated with long-term memory tasks. The most prominent hypothesis in the area is the distinctiveness hypothesis (MacLeod & Bodner, 2017); below we argue that this view accords well with the findings reported here.

The Distinctiveness account

MacLeod and Bodner (2017) summarised 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, p. 392, 2017). Essentially, MacLeod and Bodner proposed, as we have here, that producing items involves distinctive processing that supports 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, through enhanced distinctiveness processing, rather than a cost to silent ones. That said, if one adopts the view that there could be a cost to the distinctive processing, then the finding of a cost to the silent items is not, as such, problematic. The pattern that seems less compatible with a general distinctiveness view is the absence of benefit for produced items coupled with a cost for silent items, when going from pure to mixed lists.

Here we observed both a cost for silently read words and a benefit for produced items when going from pure to mixed lists. The RFM easily handles both the effects because of the similarity-based retroactive interference central to encoding in the model. We return to this when reviewing the contribution of the RFM below. Here, we wish to point out that our formal definition of distinctiveness leads to the prediction of a cost for silent items as well as a benefit for produced words, with alternating lists and immediate recall.

The Revised Feature Model

One of the assumptions of the view reported here, as in other accounts, is that the amount of useful information generated by producing items is greater than what is obtained with audio-

visual and silent items (Forrin & MacLeod, 2016b, 2018; MacLeod, 2011). Likewise, audio-visual presentation is thought to generate more useful information than silent presentation (Nairne, 1988b, Nairne, 1990). This was operationalised in the RFM by increasing the number of modality-dependent features that were associated with audio-visual items relative to silently read items; produced items were associated with the largest number of modality-dependent features.

Although the number of features that represents an item has an influence on performance, the large sawtooth pattern is the result of the *local distinctiveness* that is created through feature overwriting and the influence this has on the similarity-based retrieval mechanism. In effect, the retroactive interference embodied in the overwriting mechanism eliminates redundant information in successive items, leaving behind the features that make an item unique relative to the few items that come after. This creates a form of local, context-dependent, distinctiveness and naturally leads to recency effects (which are ubiquitous in memory research; e.g., Sederberg et al., 2008). At the point of retrieval, the cues created by the overwriting process are used to identify an item to retrieve. The retrieval mechanism selects said cues in their order of presentation (albeit with a perturbation process producing order errors). Produced items have features that cannot be overwritten by non-produced items, as the overwriting process is feature-to-feature—i.e. feature 5 in item 1 can overwrite feature 5 in item n-1, and/or item n-2 with some probability. It follows that any item that has a higher number of features than following items will be less prone to overwriting, which will increase its similarity to the intact representation of that item in memory.

In the case of the mixed lists studied here, the benefit of having more relevant features comes at the expense of the items with fewer relevant features. As illustrated in Figure 1, compared to pure lists, in mixed lists, produced items were better recalled and silently read items

suffered. In the RFM, this happens mainly because of three mechanism. The first is the retroactive interference mentioned above. The second relates to the effect of production on rehearsal, and the third is the influence that these two processes have on the relative distinctiveness of retrieval cues at the point of recall. Item *n* interferes with item *n-1* the most, somewhat less with item *n-2,* and so on. Because silent items have fewer features, if they follow a produced item they will not lead to as much overwriting as a produced item would. Conversely, produced items can overwrite the features of a silent item just as well as another silent item can. Finally, produced items interfere with rehearsal; hence, in an alternating list, produced items will interfere with the rehearsal of silent items, and silent items will have a beneficial effect on the probability that produced items will be rehearsed. We assumed this is what was happening. However, for simplicity, the model fitting estimated a single rehearsal rate for each type of list. Interestingly, fits showed a lower rehearsal rate for mixed lists than for pure lists, with the highest rehearsal rate for pure silent, followed by pure produced, and finally mixed lists. These differences are possibly capturing the impact of having to alternate between silent reading and aloud reading and vice versa in mixed lists - i.e., a form of task-switching cost may be at work and captured by the model.

Regarding the impact of production on rehearsal, it is worth noting that Forrin et al. (2019) proposed a similar idea. They argued that mixed lists generate a next-in-line effect (Brenner, 1973). Simply put, reading the items aloud in the presence of the experimenter would create self-presentation concerns. Therefore, in a mixed list, there would be performance anticipation while processing the item immediately preceding the item to be read aloud, that is the silently read item. Performance anticipation for aloud items would weaken the encoding of silently read items, and this could occur by disrupting their rehearsal.

Another novel finding reported here was the reliable production disadvantage when recalling the first few items from pure lists. Based on prior findings and suggestions by Routh (1970) and Bhatarah et al. (2009), the RFM included a rehearsal mechanism. In the RFM, rehearsal carries a small probability that each feature will be reinstated after having been overwritten. Because the probability of rehearsal decreases with the number of items to rehearse, rehearsal has a more positive impact on the first few items presented in a list. Since production has a negative impact on rehearsal, this will be particularly clear for the first few items.

The last new finding described relates to the highly similar impact of production and auditory presentation on recency; in effect, the beneficial impact of these two presentation modalities on the last few items could not be separated. The original feature model provided a feature-based / distinctiveness-based interpretation of the modality effect that is included in the RFM. The model assumes that the number of relevant modality-independent features relating to items read silently and aloud is similar – i.e., categorisation and meaning related processing does not vary with presentation modality. However, the original model and the RFM both assume that when going from silently read items to audio-visual items, the number of modality-dependent features increases. Visually presented items have fewer (encoded) modality-dependent features; as there are no further items after the presentation of the list is ended, these features provide a small recency advantage. In the case of audio-visual and produced items, a larger number of modality dependent features means that for the last few items, given no further items are presented, there is a much larger recency effect. One might expect that produced items would generate an even stronger recency effect, as further modality-dependent features are added for those items. However, bear in mind that producing an item has a negative impact on the probability of rehearsal. Overall, in the RFM, the probability of rehearsal reduces as the number

of items to rehearse increases, implying that rehearsal will have a diminishing impact across the serial positions within the list. Nevertheless, the negative impact of production on rehearsal of the last few items was enough to offset the advantage that extra features would provide for produced items.

Together, as implemented in the RFM, a few simple processes generated very good fits to the complex pattern of observed data. Basically, to explain the sawtooth findings and the relative changes in performance going from pure to mixed lists, the following were needed: a) the assumption that produced items generate more modality-dependent features; b) a mechanism that reduces redundancy in the encoding of successive items and involves loss of information–that is feature-to-feature overwriting; c) an implementation of rehearsal where production has a negative impact and, d) a retrieval mechanism that relies on relative distinctiveness. Taken each on their own, these ideas have appeared in other proposals or models (e.g., rehearsal in Baddeley & Hitch's 1974 working memory model). However, their simultaneous implementation in a formal model is novel – and in particular, we would argue that two mechanisms merit highlighting. The first is the idea that encoding more useful information has a *cost*; the second is that context-dependent *local distinctiveness* is central to being able to account for the complex patterns reported here. By context-dependent, we mean that what surrounds an item has a considerable impact on that item's memorability. Both these processes relate to encoding rather than retrieval – although of course they interact with the retrieval mechanism – whereas most models in the field have very heavily relied on retrieval processes to account for empirical findings (Brown et al, 2007; Henson, 1998; Page & Norris, 1998; see Lewandowsky & Farrell, 2008, for an exception and a discussion of this point). Overall, our modelling and account

emphasise the importance of encoding operations in memory performance; encoding matters and distinctive processing at encoding matters even more (Hunt, 2006, 2013).

Conclusion and future directions

What are the main take-home messages of the work presented here? Some of the points under this heading are related to new empirical findings. The production effect, and its complex interaction with list composition, were clearly replicated within a classic short-term memory paradigm. In alternating lists, serial recall data showed a large sawtooth effect. Moreover, relative to pure lists, with mixed alternating presentation, produced items benefit while silent items are less well recalled. Finally, when pure lists of silent and produced items are compared, production hurts the strict serial recall of the first few items in a list while benefiting the recall of the last few items.

These phenomena were uncovered through requiring ordered recall of the studied items and through an examination of recall as a function of presentation position. Further research is necessary to establish whether these patterns can also be found within paradigms that are more typical of episodic memory research. Our expectation is that if ordered recall is required, and if mixed lists are alternating, similar results will be obtained. If free recall is used and aloud and silent items are randomly distributed across the list, then what is expected changes and complexities are introduced – i.e., adjacent aloud and silent items are likely to still involve a *local distinctiveness* effect. However, any negative impact of production on rehearsal is likely to be complex because of how rehearsal interacts with list length and recall order (Grenfell-Essam & Ward, 2012).

Another question of interest is whether the proposed negative impact of production on rehearsal is a specific effect, in the sense that production might hinder rehearsal as it calls upon processes / resources that are necessary for rehearsal (e.g. motor planning for pronunciation). The other alternative is that the negative impact is a more general phenomenon. In the latter case, any processing necessary to generate extra distinctive features would be expected to hinder other encoding mechanisms, such as rehearsal, or to hurt the encoding of other dimensions; in a nutshell, distinctive processing comes at a cost. Embedded in this proposal is an assumption of limited resources. Although more research is necessary to clarify these issues, there are indications that a limited resources perspective is likely to be useful (see Popov & Reder, 2020, for related theoretical ideas).

In conclusion, the production effect is as an example of a wider category of phenomena known as encoding effects, which interact with list composition. Other examples include distinctive encoding effects such as generation, enactment, perceptual interference, bizarreness (Engelkamp & Dehn, 2000; McDaniel & Einstein, 1986; Nairne, 1988a; Serra & Nairne, 1993). The findings reported here can be viewed as a prototypical demonstration of the impact of encoding operations that generate relative distinctiveness. The most striking pattern – a sawtooth serial position curve—was observed when alternating different types of items. We argued that the latter set up a local contrast in a way that produces distinctive features for the items that have more retrieval-relevant information. Importantly, the account presented here highlights a *cost-benefit trade-off*, between encoding retrieval-relevant information and encoding adjacent items or rehearsing the encoded items. Although we did not retain the item-order hypothesis as an adequate explanation of the data patterns reported here, both our verbal hypotheses and the implementation of the RFM involve something embodied in the item-order account: A trade-off

between the encoding of relevant item-specific dimensions and other important operations – in our case rehearsal. Ultimately, it may be the case that rehearsal supports the encoding of relational information as items tend to be rehearsed in sub-groups (see, e.g., Bhatarah et al., 2009). All the encoding effects mentioned earlier (i.e., generation, etc) as well as the production effect appear to require an explanation that proposes a trade-off. This may prove to be a more important feature of the findings than has been readily recognised until now.

Exploring how obligatory and widespread the trade-off is will require further consideration and further research. If we were to venture a prediction, we would expect that better distinctive processing will require a trade-off in many if not most typical circumstances (e.g., circumstances where there is not unlimited time and resources for encoding). Moreover, it may well impact behaviour in a way that has not yet been systematically assessed. To illustrate with an example at a more complex level, many students have intuited that the learning strategies suggested by memory research actually do produce better performance, but with the caveat that the encoding operations will require more effort; this might go some way to explaining why very successful study strategies are often not implemented (e.g., spacing instead of massed study, retrieval practice, etc).

The experiments presented here and the RFM highlight the sizeable impact that local distinctiveness can have on retrieval over the short-term. The findings underline the fact that while item-specific information matters, what immediately surrounds the item also has a very significant impact on memory performance. According to the RFM, this is not about relational processing – i.e., creating a relation between previously unrelated items – the sawtooth effect underscores how our memory system reacts to a contrast based on task-relevant dimensions.

References

Baddeley, A. D. (1968). How does acoustic similarity influence short-term memory? *The Quarterly Journal of Experimental Psychology, 20*, 249–264. https://doi.org/10.1080/14640746808400159

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G.H. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory (Vol. 8, pp. 47–89). New York: Academic Press.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning & Verbal Behavior, 14(6)*, 575–589. https://doi.org/10.1016/S0022-5371(75)80045-4

Beaman, C. P., & Jones, D. M. (1998). Irrelevant sound disrupts order information in free recall as in serial recall. T*he Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 51A(3)*, 615–636. https://doi.org/10.1080/027249898391558

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

Brenner, M. (1973). The next-in-line effect. *Journal of Verbal Learning and Verbal Behavior, 12,* 320-323.

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

Conrad, R. (1964). Acoustic confusions in immediate memory. *British Journal of Psychology, 55(1)*, 75–84. https://doi.org/10.1111/j.2044-8295.1964.tb00899.x

Conrad, R., & Hull, A. J. (1968). Input modality and the serial position curve in short-term memory. *Psychonomic Science, 10(4)*, 135–136. https://doi.org/10.3758/BF03331446

Cowan, N., Baddeley, A. D., Elliott, E. M., & Norris, J. (2003). List composition and the word length effect in immediate recall: A comparison of localist and globalist assumptions. *Psychonomic Bulletin & Review, 10(1)*, 74–79. https://doi.org/10.3758/BF03196469

Cowan, N., Day, L., Saults, J. S., Keller, T. A., Johnson, T., & Flores, L. (1992). The role of verbal output time in the effects of word length on immediate memory. *Journal of Memory and Language, 31(1)*, 1–17. https://doi.org/10.1016/0749-596X(92)90002-F

Corballis, M. C. (1966). Rehearsal and decay in immediate recall of visually and aurally presented items. *Canadian Journal of Psychology, 20*, 43-51. doi:10.1037/h0082923

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

Engelkamp, J., & Dehn, D. M. (2000). Item and order information in subject-performed tasks and experimenter-performed tasks. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26(3), 671–682. https://doi.org/10.1037/0278-7393.26.3.671

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. (2016a). 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)

Forrin, N. D., & MacLeod, C. M. (2016b). Auditory presentation at test does not diminish the production effect in recognition. *Canadian Journal of Experimental Psychology, 70(2)*, 116–124. https://doi.org/10.1037/cep0000092

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

Forrin, N. D., & MacLeod, C. M. (2018). This time it's personal: The memory benefit of hearing oneself. *Memory, 26*, 574–579. https://doi.org/10.1080/09658211.2017.1383434

Forrin, N. D., Ralph, B. C. W., Dhaliwal, N. K., Smilek, D., & MacLeod, C. M. (2019). Wait for it…performance anticipation reduces recognition memory. *Journal of Memory and Language, 109.* https://doi.org/10.1016/j.jml.2019.104050

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

Greene, R. L. (1989). Immediate serial recall of mixed-modality lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 266–274. https://doi.org/10.1037/0278-7393.15.2.266

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

Greene, R. L., & Pearlman, I. (1996). The Effects of Vocalisation on Situational Frequency Estimation. *Memory, 4(4)*, 453–460. https://doi.org/10.1080/096582196388924

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*, 1909–1933. https://doi.org/10.1037/xlm0000430.supp (Supplemental)

Grohe, A.-K., & Weber, A. (2018). Memory advantage for produced words and familiar native accents. *Journal of Cognitive Psychology, 30(5–6)*, 570–587. https://doi.org/10.1080/20445911.2018.1499659

Guérard, K., & Saint-Aubin, J. (2012). Assessing the effect of lexical variables in backward recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 312–324. https://doi.org/10.1037/a0025481

Guitard, D., Saint-Aubin, J., & Cowan, N. (2020). Asymmetrical interference between item and order information in short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000956.supp (Supplemental)

Henson, R. N. A. (1998). Short-term memory for serial order: the start-end model. *Cognitive Psychology*, 36, 73–137. https:// https://doi.org/10.1006/cogp

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

Hulme, C., Stuart, G., Brown, G. D. A., & Morin, C. (2003). High- and low-frequency words are recalled equally well in alternating lists: Evidence for associative effects in serial recall. *Journal of Memory and Language, 49*, 500–518. https://doi.org/10.1016/S0749-596X(03)00096-2

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

Hunt, R. R. (2013). Precision in memory through distinctive processing. *Current Directions in Psychological Science, 22(1)*, 10–15. https://doi.org/10.1177/0963721412463228

Jamieson, R. K., & Spear, J. (2014). The offline production effect. *Canadian Journal of Experimental Psychology, 68*, 20–28. doi:10.1037/cep0000009

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(1)*, 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

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

Laughery, K. R., & Pinkus, A. L. (1966). Short-term memory: Effects of acoustic similarity, presentation rate and presentation mode. *Psychonomic Science, 6,* 285-286

Lewandowsky, S., & Farrell, S. (2008). Phonological similarity in serial recall: Constraints on theories of memory. *Journal of Memory and Language, 58(2)*, 429–448. https://doi.org/10.1016/j.jml.2007.01.005

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.). *Handbook of mathematical psychology* (pp. 103–189). New York: Wiley.

MacLeod, C. M. (2011). I said, you said: The production effect gets personal. *Psychonomic Bulletin & Review, 18*, 1197–1202. https://doi.org/10.3758/s13423-011-0168-8

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. S*tatistics and Computing. 22*, 1167-1180. doi: 10.1007/s11222-011-9288-2

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44(2)*, 314–324. https://doi.org/10.3758/s13428-011-0168-7

McDaniel, M. A., & Bugg, J. M. (2008). Instability in memory phenomena: A common puzzle and a unifying explanation. *Psychonomic Bulletin & Review, 15*, 237–255. https://doi.org/10.3758/PBR.15.2.237

McDaniel, M. A., & Einstein, G. O. (1986). Bizarre imagery as an effective mnemonic aid: The importance of distinctiveness. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 54–65. doi:10.1037/0278-7393.12.1.54

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

Murray, D. J. (1965). Vocalization-at-presentation and immediate recall, with varying presentation-rates. *The Quarterly Journal of Experimental Psychology, 17*, 47–56. https://doi.org/10.1080/17470216508416407

Murray, D.J. (1966). Vocalization-at-presentation and immediate recall, with varying recall methods. *The Quarterly Journal of Experimental Psychology,18A*, 9-18. doi:10.1080/14640746608400002

Nairne, J. S. (1988a). The mnemonic value of perceptual identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition,14*, 248–255. doi:10.1037/0278-7393.14.2.248

Nairne, J. S. (1988b). A framework for interpreting recency effects in immediate serial recall. *Memory & Cognition*, *16*, 343-352.

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

Nairne, J. S., & Walters, V. L. (1983). Silent mouthing produces modality- and suffix-like effects. *Journal of Verbal Learning & Verbal Behavior, 22*, 475–483. https://doi.org/10.1016/S0022-5371(83)90300-6

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

Neath, I., & Surprenant, A. M. (2007). Accounting for age-related differences in working memory using the feature model. In N. Osaka, R. H. Logie, & M. D'Esposito (Eds.), *The cognitive neuroscience of working memory: Behavioural and neural correlates* (pp. 165-179). Oxford, UK: Oxford University Press.

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

Osth, A. F., & Dennis, S. (2015). The fill-in effect in serial recall can be obscured by omission errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41(5)*, 1447–1455. https://doi.org/10.1037/xlm0000113

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

Page, M. P. A., & Norris, D. (1998). The primacy model: a new model of immediate serial recall. *Psychological Review, 105*, 761–781. https://doi.org/10.1037/0033-295x.105.4.761-781

Penney, C. G. (1975). Modality effects in short-term verbal memory. *Psychological Bulletin, 82(1)*, 68–84. https://doi.org/10.1037/h0076166

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

Poirier, M., Schweickert, R., & Oliver, J. (2005). Silent reading rate and memory span. *Memory, 13*, 380–387. https://doi.org/10.1080/09658210344000440

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 and Cognition, 47(4)*, 603–618. doi:10.3758/s13421-018-0882-9.

Pollack, I. (1963). Interference, rehearsal, and short-term retention of digits. *Canadian Journal of Psychology, 17(4)*, 380–392. https://doi.org/10.1037/h0083279

Popov, V., & Reder, L. M. (2020). Frequency effects on memory: A resource-limited theory. *Psychological Review, 127(1)*, 1–46. https://doi.org/10.1037/rev0000161

Quinlan, C. K., & Taylor, T. L. (2013). Enhancing the production effect in memory. *Memory, 21*, 904–915. https://doi.org/10.1080/09658211.2013.766754

Quinlan, C. K., & Taylor, T. L. (2019). Mechanisms underlying the production effect for singing. *Canadian Journal of Experimental Psychology*. https://doi.org/10.1037/cep0000179

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

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review, 115(4)*, 893–912. https://doi.org/10.1037/a0013396

Serra, M., & Nairne, J. S. (1993). Design controversies and the generation effect: Support for an item-order hypothesis. *Memory & Cognition, 21(1)*, 34–40. https://doi.org/10.3758/BF03211162

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317-1323. doi: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. doi: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.

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

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

Watkins, M. J. (1977). The intricacy of memory span. *Memory & Cognition, 5,* 529–534. https://doi.org/10.3758/BF03197396

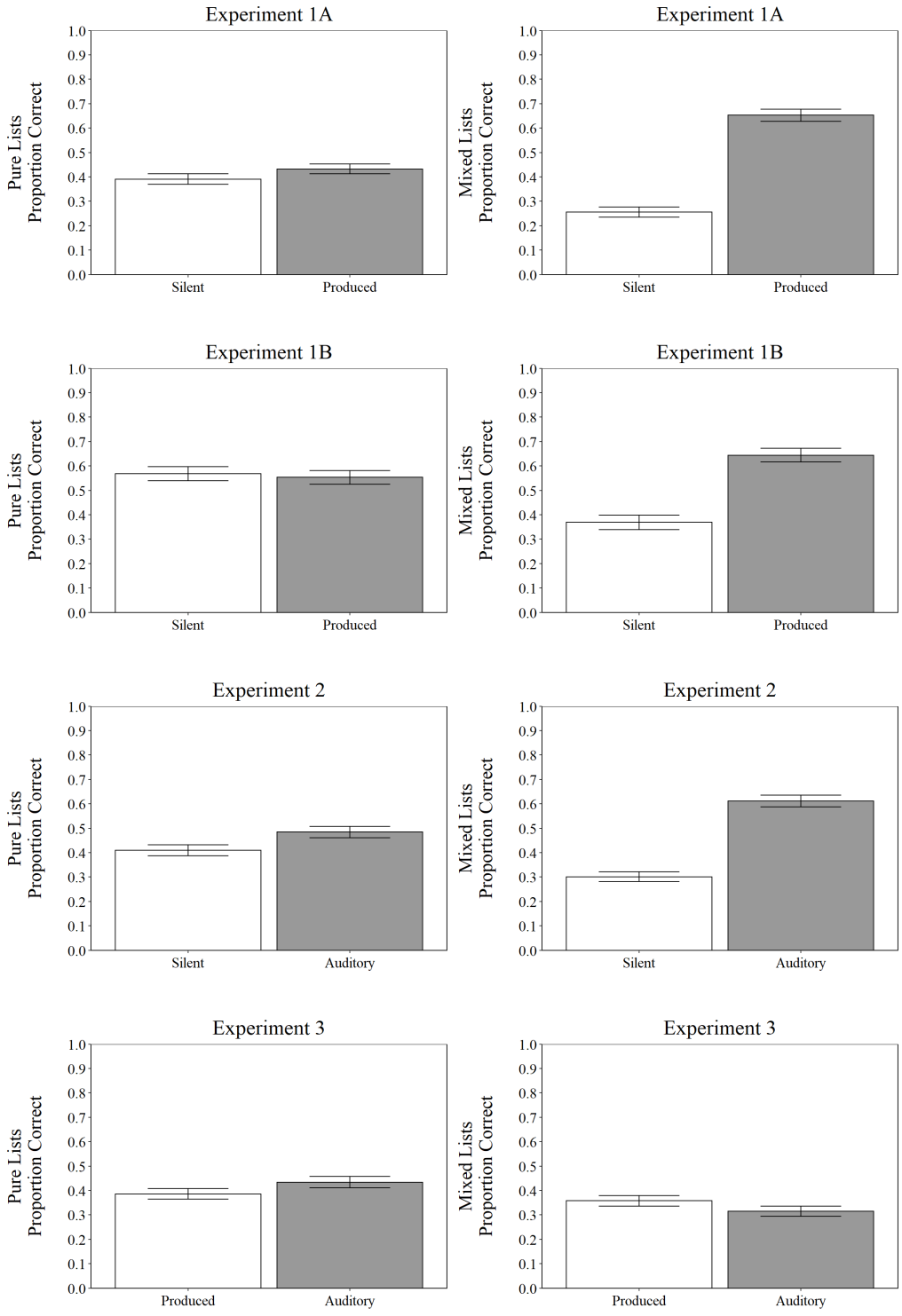

Figure 1
Proportion of correct recall as a function of experiment, condition (silent or produced) and list type (pure or mixed). Error bars represent confidence intervals at 95% for the repeated measures factor, computed with Morey's (2008) method.

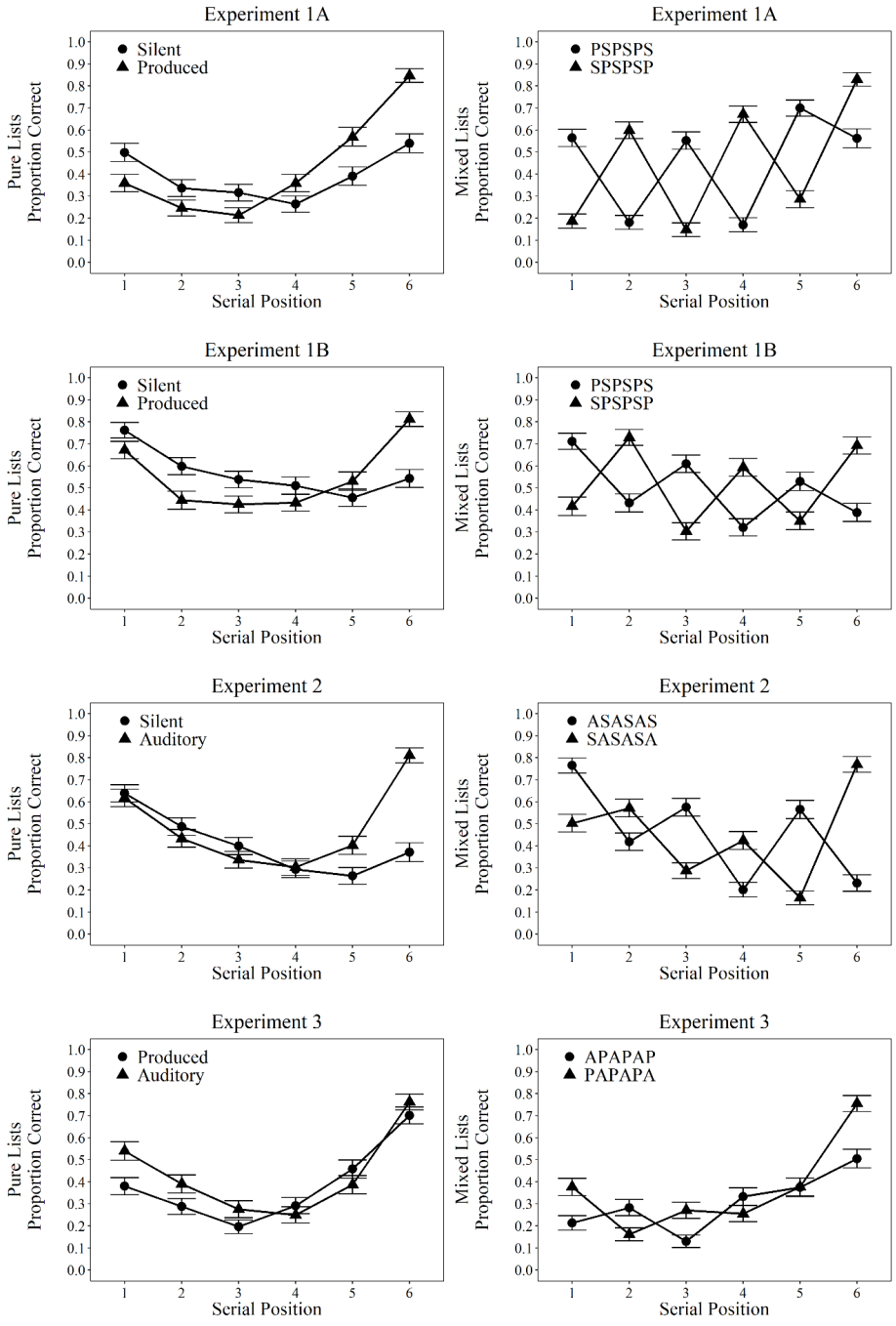

Figure 2

Proportion of correct recall as a function of experiment, serial position, input modality (silent or produced) and list type (pure or mixed). The letters S, A, and P indicate that the word holding that position was read silently, with a concomitant auditory presentation, or was read aloud, respectively. Error bars represent confidence intervals at 95% for the repeated measures factor, computed with Morey's (2008) method.

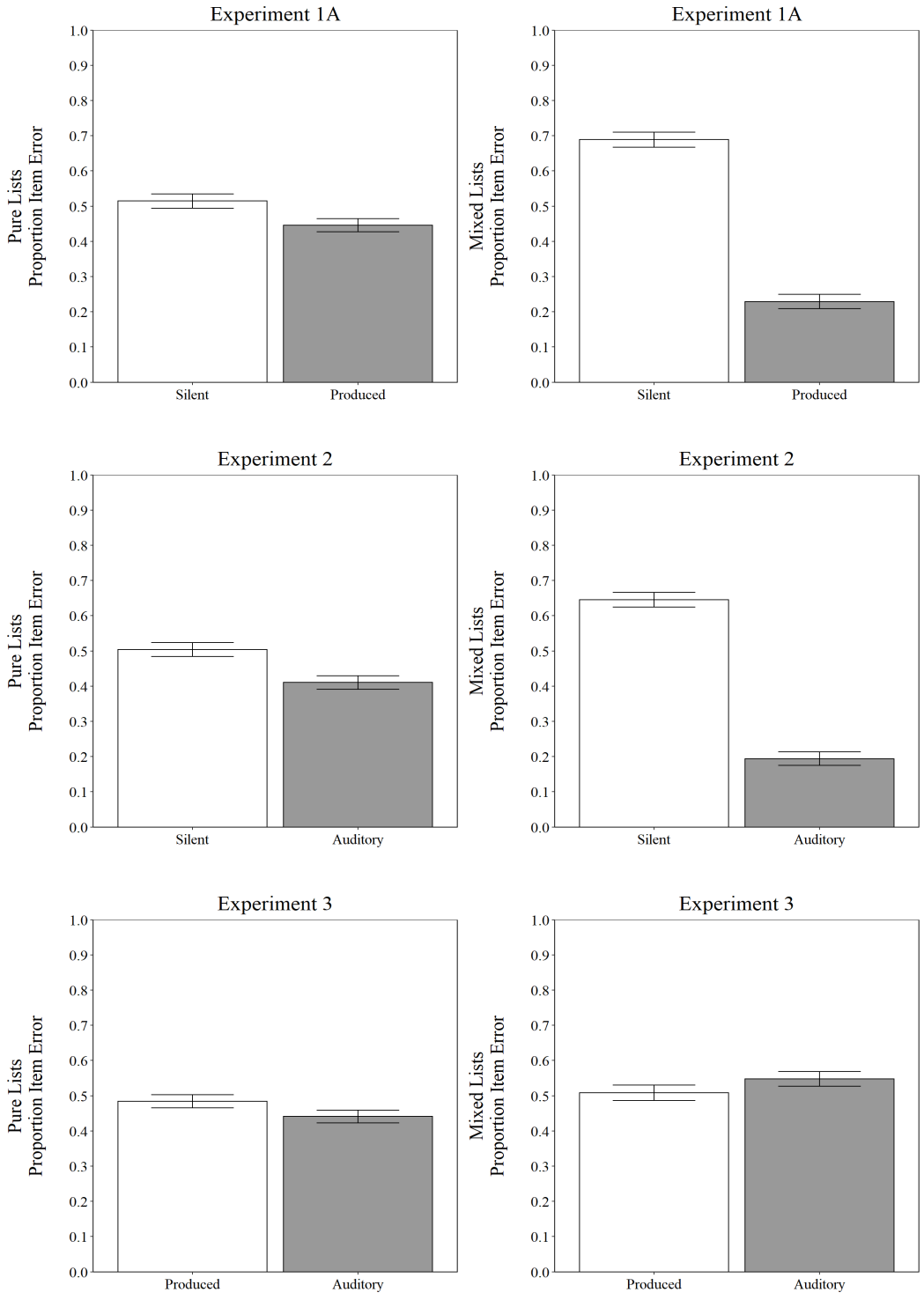

Figure 3

Proportion of item errors as a function of experiment, condition (silent or produced) and list type (pure or mixed). Error bars represent confidence intervals at 95% for the repeated measures factor, computed with Morey's (2008) method.

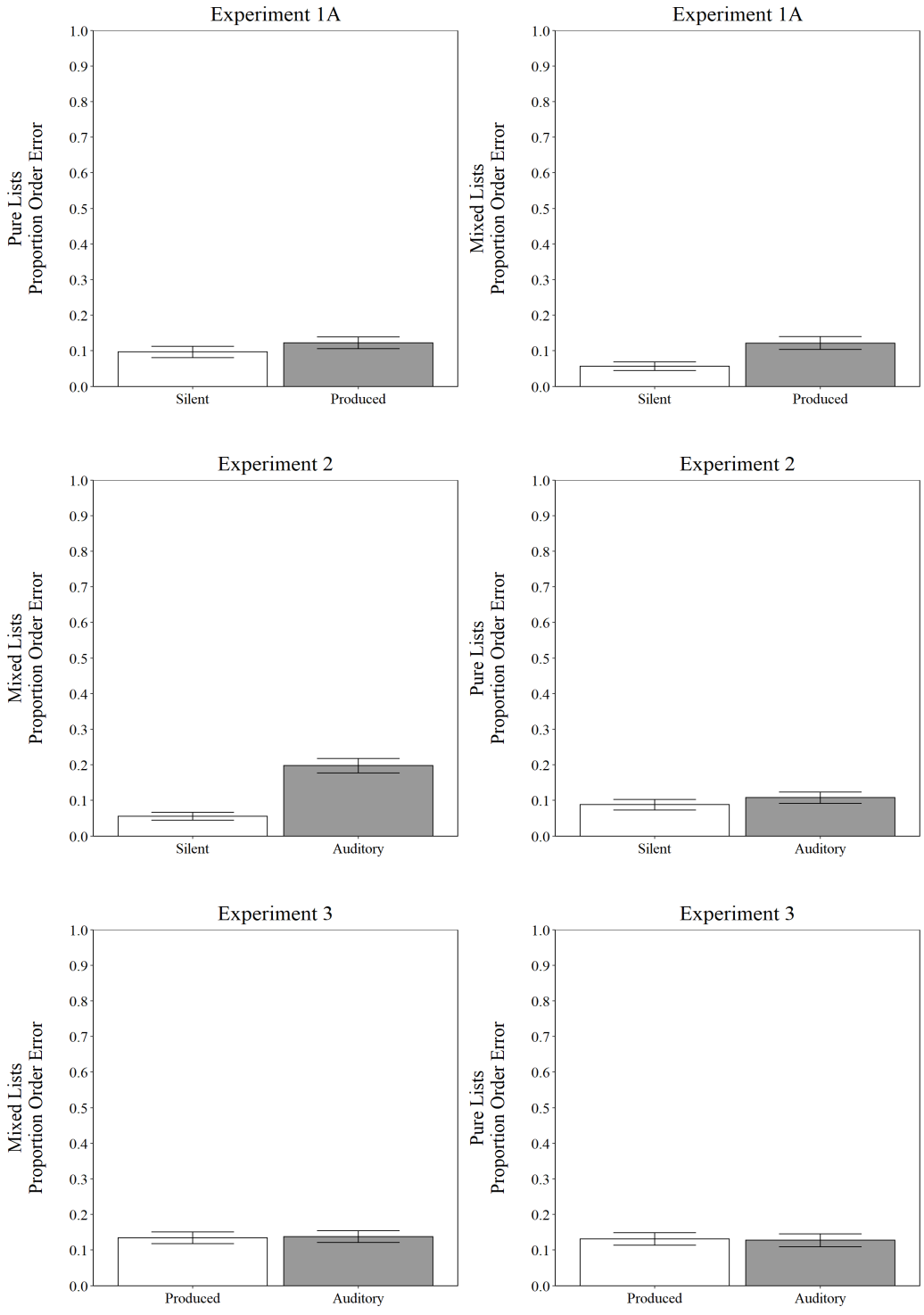

Figure 4

Proportion of order errors as a function of experiment, condition (silent or produced) and list type (pure or mixed). Error bars represent confidence intervals at 95% for the repeated measures factor, computed with Morey's (2008) method.

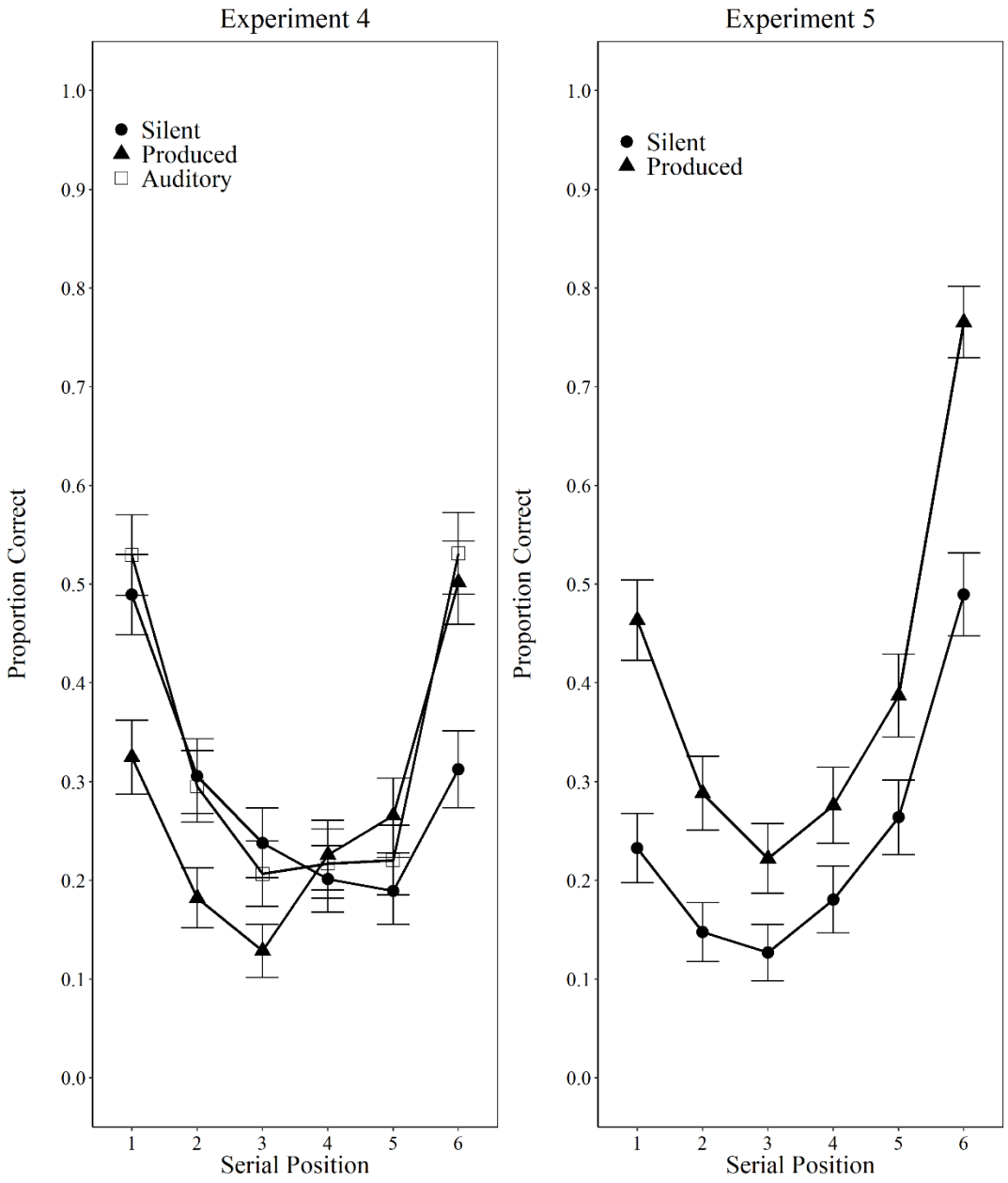

Figure 5

Proportion of correct recall as a function of experiment, serial position, and input modality. Error bars represent confidence intervals at 95% for the repeated measures factor, computed with Morey's (2008) method.

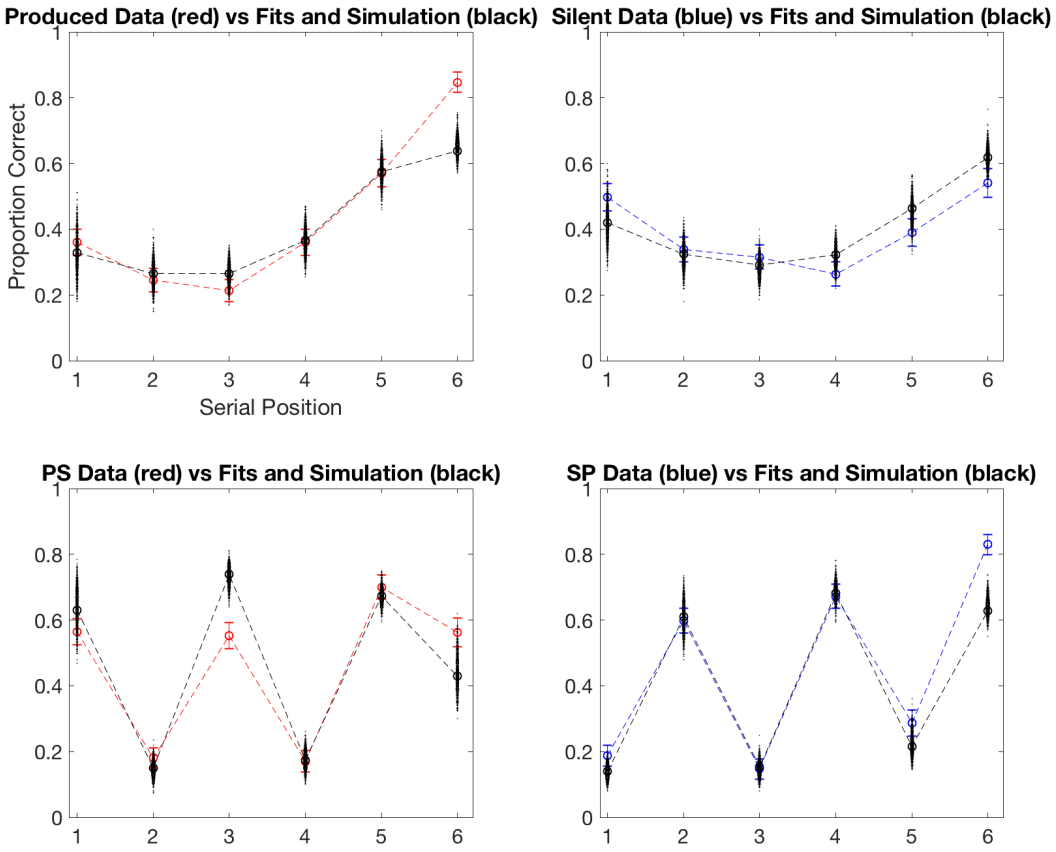

Figure 6

Data vs. model fit for each of the four conditions in Experiment 1A. On the whole the model matches the data well, although some missfitting is evident for the final item in each list. Error bars on the data are 95% CIs.

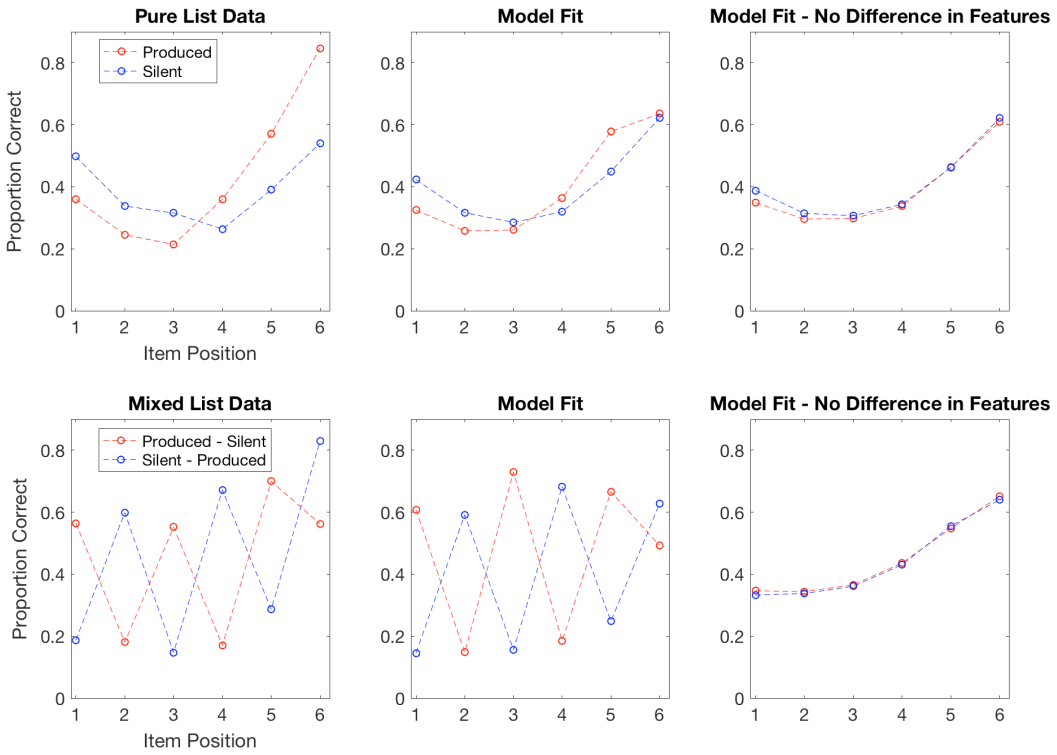

Figure 7

Data and Model Simulations for Experiment 1A. The top panels refer to pure lists, and the bottom one to mixed. The left column shows data, the middle column simulations based on best fitting parameters for the full model, and the right column simulations based on best fitting parameters for a reduced model where the number of features is assumed equal for the two item types. This figure clearly shows that the full model is reproducing the qualitative patterns in the data, in particular the cross-over for pure lists and the zigzag pattern for mixed ones. In contrast fixing the number of features, ie removing relative distinctiveness, removes the ability of the model to capture the trends in the data, particularly for mixed lists.

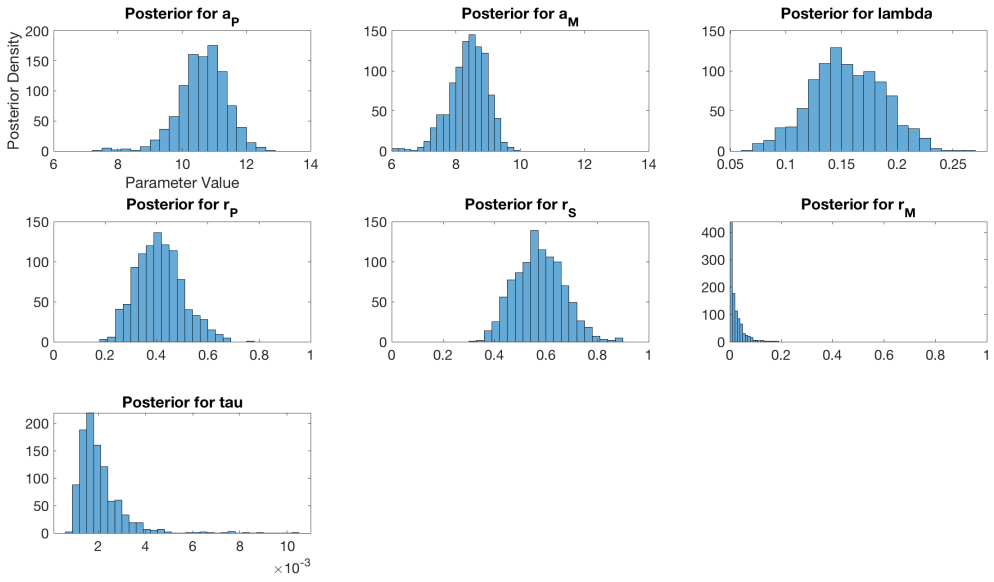

Figure 8

Posteriors for the parameters from fitting to Experiment 1A. All parameters look well behaved. Crucially the distributions of the rehearsal parameters suggest that while rehearsal is somewhat less effective for produced over silent items, it is much more substantially reduced in mixed lists, likely as a result of the extra demands of the task.

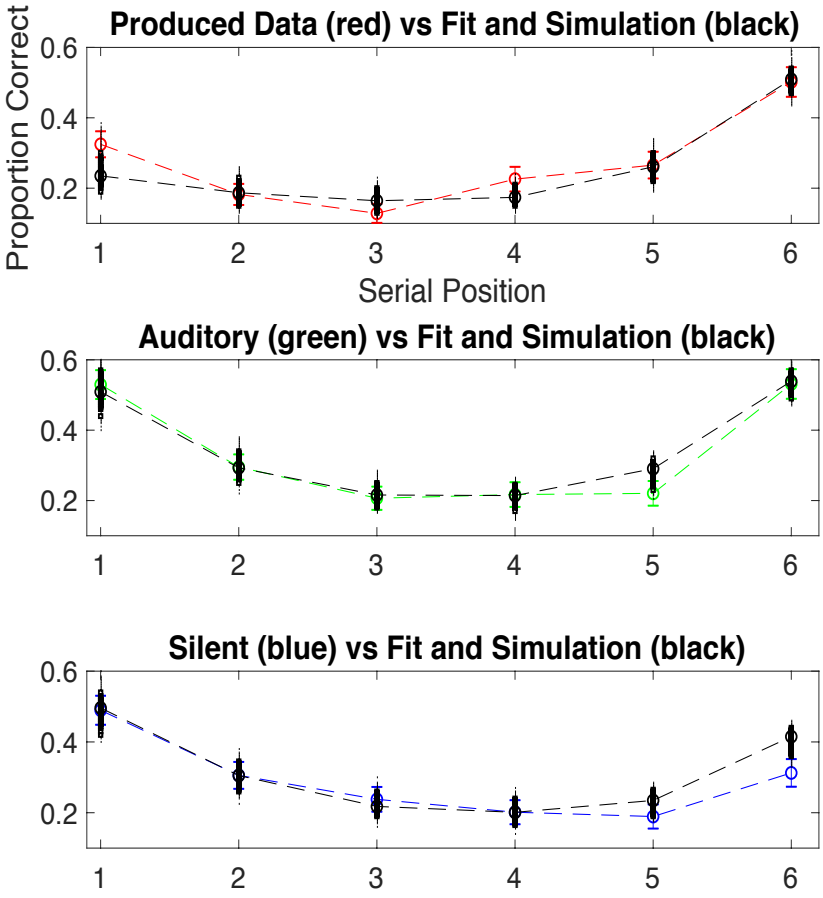

Figure 9

Data vs. model fits for each of the three conditions in Experiment 4. Coloured lines represent data, black squares are histograms of model posterior predictions. Error bars on the data are 95% CIs. There does not appear to be any systematic misfitting.

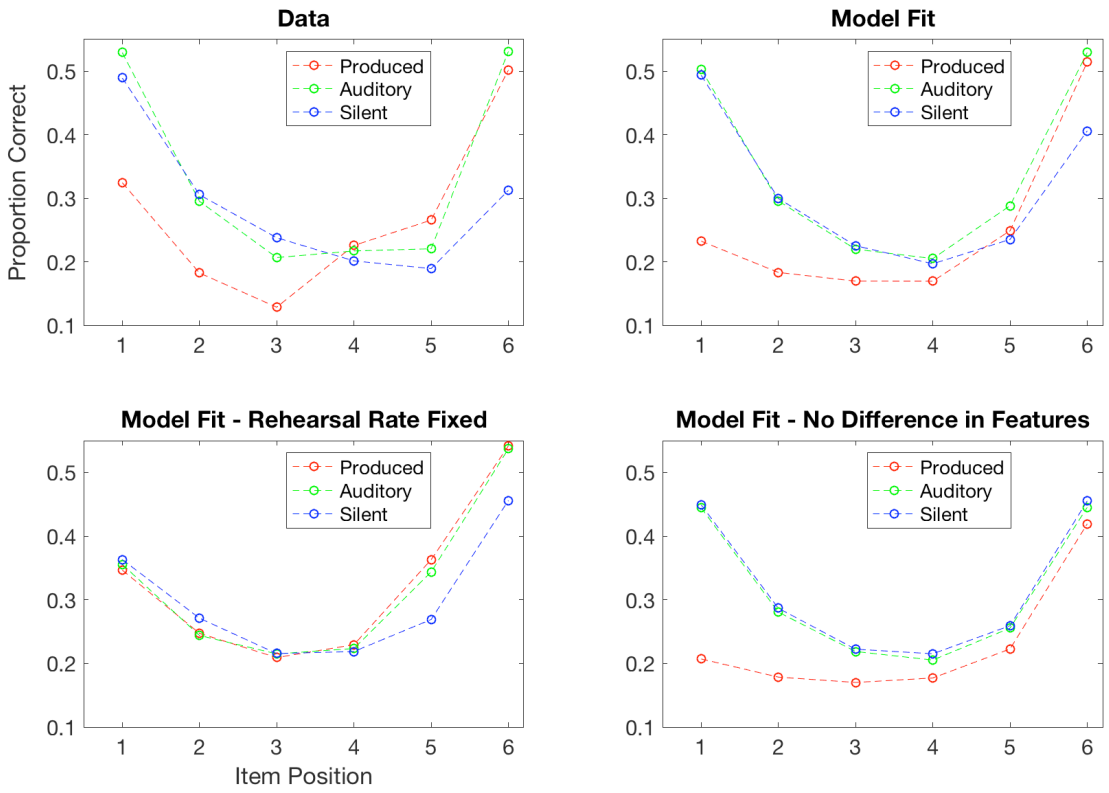

Figure 10

Data and Model Simulations for Experiment 4. The top left panel shows the data for the three conditions. The top right panel shows simulations based on the best fitting model parameters. The bottom panels for simulations based on best fitting parameters for versions of the model where the rehearsal rate is fixed (bottom left) or the number of features is set equal for different item types (bottom right.) The overall pattern of a match between auditory and silent early on, and between auditory and produced later in the list is reproduced well by the full model. Fixing the rehearsal rate removes the advantage for auditory and silent items early on, and demanding the number of features be the same for different item types removes the modality effect at the end of the list.

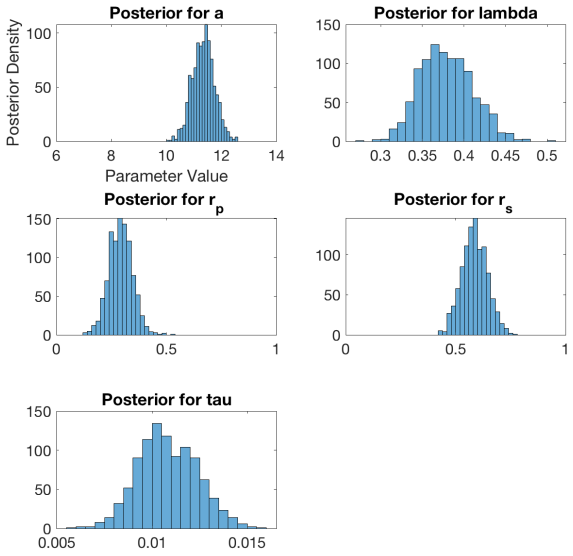

Figure 11

Posterior estimates of the model parameters for Experiment 5. All parameters appear to be well behaved. As for Experiment 1A, the rate of rehearsal is somewhat higher in the silent and auditory conditions, than in the produced condition,

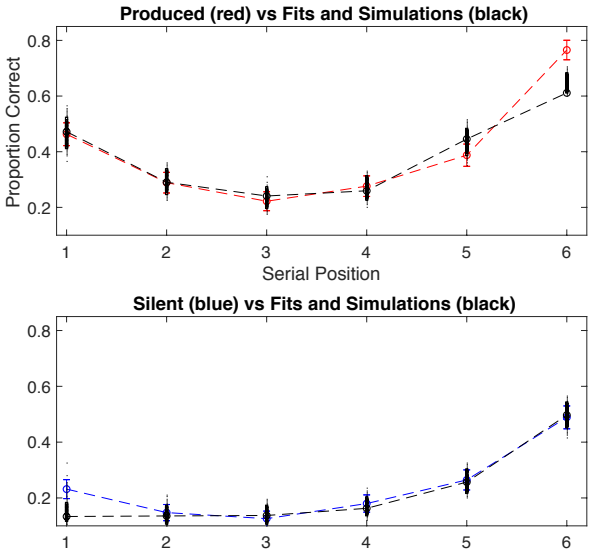

Figure 12

Data vs. model fits for the two conditions in Experiment 5. Coloured lines represent data, black squares are histograms of model posterior predictions. Error bars on the data are 95% CIs. As for Experiment 1A, the model slightly underestimates recall at the final position of the produced list, however generally the fits are good.

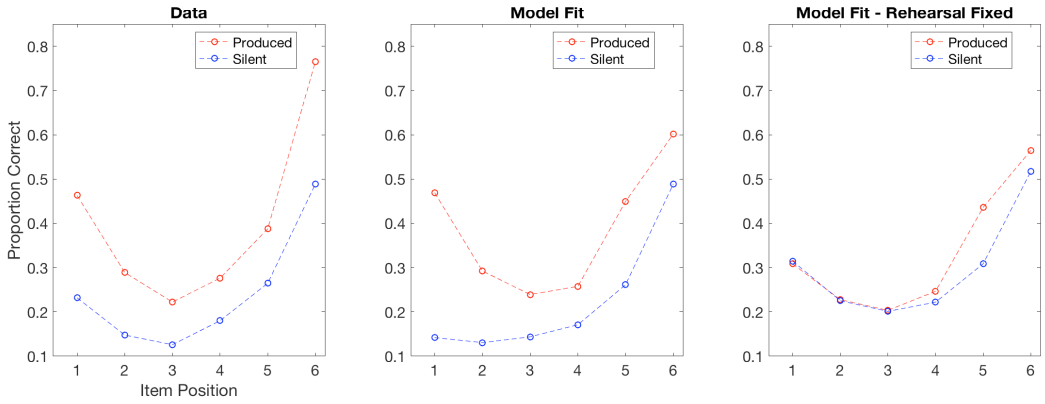

Figure 13

Data and model fits for Experiment 5. The left panel shows the data, the middle panel shows model simulations for the full model based on best fitting parameters, and the right panel shows model simulations for a version of the model with rehearsal rates fixed for produced and silent lists. The overall pattern of lower recall performance across list positions for silently read items is reproduced well by the full model (middle panel), and fixing rehearsal rates eliminates the advantage for produced lists, particularly for early items (right panel).

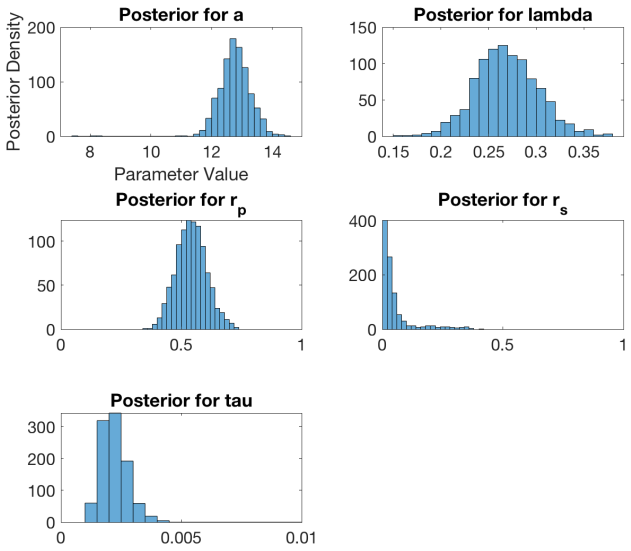

Figure 14

Posterior estimates of the model parameters for Experiment 5. All parameters appear to be well behaved. In contrast to Experiments 1A and 4, but as hypothesised, the rate of rehearsal is now much lower in the silent condition, than in the produced condition,

Appendix 1 – 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).

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 position 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.

Details of the model and fitting parameters for each data set are given in Table S1 below. Given the model complexity it is doubtful that we have good parameter identifiability, so it is more useful to focus on parameter comparisons within an experiment.

Table S1.

Parameter Values for all model fits reported in the main text: the majority of these parameter values are identical to those used in the simulations in Neath and Nairne (1995).

		Experiments		
	Parameter values	Experiment 1A	Experiment 4	Experiment 5
Fixed Model Parameters	Forgetting Probability	100		
	Recovery constant	2		
	# of items	6		
	# independent features	***20***		
	# dependent features	20 (produced) 2 (silent)	20 (produced) 15 (auditory) 2 (silent)	20 (produced) 2 (silent)
	Type Features	3		
Estimated Parameters	Distance Scaling Parameters Median [95% HDI]	$a_{Pure} = 10.69$ [9.11, 11.96] $a_{Mixed} = 8.45$ [7.13, 9.35]	$a_{Pure} = 11.36$ [10.53, 12.15]	$a_{Pure} = 12.77$ [11.88, 13.71]
	Overwriting Parameter Median [95% HDI]	$\lambda = .15$ [.09, .22]	$\lambda = .38$ [.32, .44]	$\lambda = .27$ [.21, .34]
	Choice Consistency Median [95% HDI]	$\tau = .0018$ [.0011, .0043]	$\tau = .011$ [.0079, .014]	$\tau = .0022$ [.0013, .0035]

Fitting Parameters

Rehearsal Median [95% HDI]	$r_{Produced} = .41$ [.25, .61] $r_{Silent} = .57$ [.40, .76] $r_{Mixed} = .012$ [.000, .105]	$r_{Produced} = .29$ [.18, .40] $r_{Silent/Auditory} = .59$ [.47, .70]	$r_{Produced} = .54$ [.42, .67] $r_{Suppressed} = .026$ [.0012, .2853]
Number of Particles	1000	1000	1000
Simulations per step	500	500	500
Generations	24	10	9
ABC parameters	Minimum ϵ = .45 Maximum Proposal λ = 80 Minimum Proposal SD = .4	Minimum ϵ = .2 Maximum Proposal λ = 50 Minimum Proposal SD = .5	Minimum ϵ = .1 Maximum Proposal λ = 50 Minimum Proposal SD = .5