Distinctiveness and Interference in Free Recall: A Test with the Production Effect

Sébastien Gionet[1], Dominic Guitard[2], Marie Poirier[3], James M. Yearsley[3], and Jean Saint-Aubin[1]

[1]École de Psychologie, Université de Moncton

[2]School of Psychology, Cardiff University

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

Authors' Note

We have no conflicts of interest to disclose. This research was supported by Discovery grant RGPIN-2023-05943 from the Natural Sciences and Engineering Research Council of Canada (NSERC) to JSA. While working on the manuscript, SG was supported by a scholarship from NSERC. We would also like to thank Ian Dauphinee, Mathis Roy, Isabelle Bernard, Maude Emilie Thériault, Mylène Miner, and Sophie Plouffe for assisting with the data collection.

Correspondence regarding this article should be addressed to Sébastien Gionet at esg8807@umoncton.ca or to Jean Saint-Aubin, Université de Moncton, 18 ave Antonine-Maillet, Moncton, New Brunswick, E1A 3E9, Canada, Email: jean.saint-aubin@umoncton.ca.

CRediT Author Contribution Statement

SG: Conceptualization, Formal analysis, Investigation, Software, Visualization, Writing-original draft. **DG**: Conceptualization, Formal analysis, Supervision, Writing-review and editing. **MP**: Conceptualization, Writing-review and editing. **JMY**: Conceptualization, Formal analysis, Visualization, Writing-original draft. **JSA**: Conceptualization, Funding acquisition, Resources, Supervision, Writing-review and editing.

Abstract

It is well established that when some words within a list are read aloud, or produced, and others are read silently, produced items are better recalled. According to the Revised Feature Model (RFM), this benefit stems from additional features and enhanced distinctiveness. Since the model assigns forgetting to similarity-based retroactive interference, produced items should be better recalled when followed by a silent item than by another produced item. However, this mechanism had never been directly tested. In addition, Forrin et al. (2019) suggested a memory cost for silent items in mixed lists. Based on their Production Anticipation Hypothesis (PAH), this cost derived from social anxiety and performance anticipation emerging within the experimental setting. Here, we tested these two competing accounts in 4 experiments in which 240 participants completed an immediate or delayed free recall task with 10-word lists. Produced and silent items were presented in two blocks of various lengths, and the occurrence of produced items within the list was predictable. We also manipulated the presence of the experimenter in the room during the task, so that the experimenter was present for Experiments 1A and 2A, but absent for Experiments 1B and 2B. Overall, results offer full support for the RFM, but very limited support for the PAH. The main trends in the data were also modeled by the RFM. In sum, with the production effect, we provide the first comprehensive test of the retroactive interference mechanism described within the RFM and further inform our understanding of memory.

Keywords. production effect, serial positions, free recall, revised feature model, production anticipation hypothesis

Distinctiveness and Interference in Free Recall: A Test with the Production Effect

It is well-established that if an ITEM stands out relative to its surrounding context, it is more likely to be remembered. For instance, if tested for memory of the above sentence, the word *item* should be better recalled. This example illustrates the importance of distinctiveness, one of the few core memory principles most agree on (Brown et al., 2007; Hunt, 2013; Surprenant & Neath, 2009). Some of the most striking distinctiveness effects are known as local distinctiveness effects, observed when there is a salient contrast between two adjacent to-be-remembered items (see, Hunt, 1995; Neath et al., 2006; Saint-Aubin et al., 2021; von Restorff, 1933).

Arguably, the most dramatic demonstration of such an effect in terms of sheer effect size is seen when participants must memorize a mixed list in which words are alternatively read aloud and read silently (Cyr et al., 2022; Saint-Aubin et al., 2021). Under these conditions, aloud items (hereafter referred to as produced items) are better remembered than silent items, a phenomenon known as the production effect (MacLeod & Bodner, 2017; MacLeod et al., 2010). Importantly, when all items are read aloud or silently (pure lists), the recall advantage of produced items is very much reduced or even absent (see Fawcett, 2013). Therefore, one of the hallmarks of this effect is that produced items in mixed lists are better remembered than those in pure lists (Cyr et al., 2022; Jonker et al., 2014; Lambert et al., 2016). However, it is also true that there is typically a decrease in memory for silent items in mixed lists relative to pure lists – an empirical finding that does not figure as prominently in discussions of the production effect (e.g., Bodner et al., 2014; Jones & Pyc, 2014; MacLeod & Bodner, 2017). Figure 1 below illustrates what happens when going from pure to mixed lists: Produced items benefit whereas silently read items suffer.

Theoretical and computational accounts of the production effect typically focus on the benefits than can be derived from producing the items, maintaining the idea that produced items are likely to be more distinctive due to the additional features generated by production (Caplan &

Guitard, 2024a; Jamieson et al., 2016; Kelly et al., 2022; Wakeham-Lewis et al., 2022). However, Forrin et al. (2019) recently drew attention to the cost, for silent items, of going from pure lists to mixed lists. This is important, as distinctive encoding can generate costs of various types across tasks and settings (e.g., Huff et al., 2021; Saint-Aubin et al., 2021). Moreover, in an interesting departure from previous explanations of the cost for silent items in mixed lists (see, Bodner et al., 2014; Jonker et al., 2014), Forrin et al. attributed the cost to social rather than cognitive factors, the former often being neglected in cognitive science.

The influence of social factors on memory is well illustrated by the common experience of waiting to give a talk at a conference or an oral presentation in school and not remembering the content of the presentation just before ours. This real-life experience has been systematically investigated in the laboratory (e.g., Bond & Kirkpatrick, 1982; Bond & Omar, 1990; Brenner, 1973; Brown & Oxman, 1978). Typically, in those studies, a group of around 22 participants were tested together. Each participant had to read one word aloud, listening as the other words were read aloud by other participants, before trying to recall all items at the end. Results for each participant showed better recall performance for the item that was read aloud, but poorer recall performance for the item presented just before the item that was read aloud. The benefit of the aloud item was attributed to a Von Restorff effect (Hunt, 1995; von Restorff, 1933), whereas the disadvantage for the preceding item was assigned to performance anticipation. Building on those older studies from experimental social psychology, Forrin et al. (2019) developed the Production Anticipation Hypothesis (PAH) to account for performance in an item recognition memory task. According to the PAH, anticipating having to read words aloud in front of an experimenter would induce a form of performance anxiety that takes resources away from encoding. Importantly, this distraction would be absent in pure lists composed only of silent items, which would explain the memory cost observed for silent items in mixed lists.

Contrary to other computational accounts of the production effect that focus on the benefit for produced items (Caplan & Guitard, 2024a; Jamieson et al., 2016; Kelly et al., 2022; Spear et al., 2024; Wakeham-Lewis et al., 2022), the Revised Feature Model (RFM) handles both the benefit for produced items and the cost for silent items when going from pure lists to mixed lists (Cyr et al., 2022; Saint-Aubin et al., 2021). According to the RFM, the benefit for produced items in mixed lists is due to the enhanced distinctiveness provided by being followed by silent items. Conversely, the cost for silent items in mixed lists reflects reduced rehearsal opportunities – i.e., production interferes with the covert rehearsal of all list items (e.g., Murray, 1967).

Interestingly, for certain compositions of mixed lists (i.e., lists where produced and silent items are presented in two blocks), the RFM and PAH make predictions that are sort of a mirror image of each other. Specifically, from the PAH's perspective, the most *detrimental* situation is one where participants know that a silent item will be followed by a produced item. However, the RFM argues that the most *beneficial* situation is the one where a produced item is followed by a silent item. This is because the RFM includes a retroactive interference mechanism based on feature similarity (Cyr et al., 2022; Saint-Aubin et al., 2021, 2024). As silent items do not have the extra features generated by production, they cannot interfere with produced items to the same degree as other produced items interfere, leading to a boost in distinctiveness for any produced item that is followed by a silent item. In this article, we sought to systematically investigate both sides of this theoretical coin, focusing on the conditions that test the critical predictions of each view and, where possible, contrasting conflicting predictions.

The Production Anticipation Hypothesis (PAH)

Although most researchers do not often dwell on this fact, memory experiments require participants to behave in a specific way governed by a series of unwritten rules, thus making the experimental situation a unique social setting (e.g., Klein, 2014; Klein & Marghetis, 2017). For

instance, participants are made to perform as “ideal subjects”, with the experimenter observing and judging their responses. However, how this setting influences participants’ performance in memory experiments remains unclear. Forrin et al.’s (2019) study of recognition memory is an important exception. As mentioned above, they suggested that the well-established production effect could be driven—at least in part—by social rather than cognitive factors. Specifically, they suggested that there was a memory cost for silent items preceding produced items in mixed lists. This cost would be due to the discomfort elicited by anticipating having to read aloud in front of an experimenter, as when someone is next in line to speak in public. During the presentation of items to be read aloud or silently, awareness of the evaluative stance of the experimenter would interfere with memory for the content presented immediately before having to read aloud; that is, preoccupation with performance would interfere with current processing and encoding.

This view was previously illustrated by Brenner (1973) in a seminal article. Twenty-two participants sat in pairs around a large table with cards displayed in front of each pair. Going around the circle, one by one, a predesignated pair member was required to turn a card to display a word and read it out loud while the other pair member listened. Participants were then asked to recall all the words in a free recall task. Results showed a typical production effect with produced items being better recalled relative to those that were heard (see also, Forrin & MacLeod, 2018). Most importantly, pair members tasked with reading aloud showed reduced memory (compared to those tasked with listening) for the item that preceded the one that they read aloud. Recently, Forrin et al. (2019) extended Brenner’s findings by using a typical contemporary design and a production effect experiment in which participants were individually tested and had to read some words aloud and others silently. Horizontal lines on the screen informed participants in advance of how they were to encode each word. Results showed a cost for silent items preceding produced items, but this cost was only observed when the experimenter was in the room.

Distinctiveness / The Revised Feature Model

Since its delineation by MacLeod et al. (2010), the production effect has typically been explained by suggesting that produced items benefit from enhanced distinctiveness, derived from the additional features that production entails (Cyr et al., 2022; Ozubko & MacLeod, 2010; Saint-Aubin et al., 2021). This idea has also been implemented in formal models of memory, including MINERVA 2 (Jamieson et al., 2016; Spear et al., 2024), the REM framework (Kelly et al., 2022; Wakeham-Lewis et al., 2022), and the Attentional Subsetting Theory (Caplan & Guitard, 2024a). Moreover, feature-based distinctiveness is also at the heart of the Revised Feature Model (RFM; Saint-Aubin et al., 2021), which has been applied to the production effect in various recall and order reconstruction tasks (Cyr et al., 2022; Dauphinee et al., 2024; Saint-Aubin et al., 2021, 2024). In the RFM, it is assumed that the enhanced distinctiveness derived from reading words aloud—which requires extra processing—also comes at a cost. In recall, production would interfere with maintenance processes, more specifically with rehearsal. Here, we test these core assumptions about the costs and benefits of production by testing a series of new predictions.

As mentioned above and shown in Figure 1, when comparing overall recall performance in pure lists and mixed lists, memory for produced items is better in mixed lists than in pure lists. However, the opposite is observed for silent items, which are better remembered in pure lists than in mixed lists. This pattern of results can be explained by the RFM (Cyr et al., 2022; Saint-Aubin et al., 2021). In mixed lists, relative to silent items, produced items benefit more from their additional features. This happens because the RFM includes a retroactive interference mechanism that is similarity driven. Specifically, this mechanism allows an item's features to overwrite the similar features of the preceding items with a probability that decreases as the distance between the items increases. It follows then that any produced item followed by a silent item will have its production-specific features relatively well protected from interference, compared to the situation

where a produced item is followed by another produced item (see Figure 2). Moreover, the RFM assumes that the act of production, similarly to articulatory suppression (Murray, 1967), impedes covert rehearsal (for a similar idea, see Routh, 1970). This implies that, relative to those in a pure list, less rehearsal is possible for silent items in a mixed list, thus leading to a memory cost.

The Current Study

Here, we contrasted predictions from the PAH with those of the RFM, a model based on relative distinctiveness. The RFM offers a different interpretation of the production effect and makes different predictions to those of the PAH. Therefore, we built a design that allowed us to simultaneously test opposing predictions from both accounts. To do so, we selected the free recall task, because it is within the scope of the RFM which applies to recall and order reconstruction tasks. Furthermore, Forrin et al. (2019) were inspired by Brenner's (1973) work on the next-in-line effect. Importantly, using an item recognition task with long lists of items and multiple switches between aloud and silent items, Forrin et al. observed the same pattern of results as reported by Brenner with a free recall task, much shorter lists, and only two switches.

Specifically, we asked participants to memorize lists of 10 words and manipulated production within each list by grouping produced and silent items into two blocks. Then, immediately after encoding (Experiment 1) or after a 30-second filled interval (Experiment 2), they were asked to complete a free recall task. This blocked design was similar to Forrin et al.'s (2019) second experiment, where produced and silent items alternated throughout a mixed list of 80 items by blocks of 5 items. We used shorter lists of 10 items, which are better suited to free recall, because it usually results in poorer performance than recognition tasks. Finally, echoing the procedure used by Forrin et al. (2019) in their third experiment with random mixed lists, we made the occurrence of produced and silent items predictable by adding a conveyor of colored horizontal lines on the screen that indicated the condition of the upcoming words (see Figure 3).

An important issue when assessing free recall as a function of an item's position within a block or a list is that recall performance varies across the serial position curve (e.g., Ward, 2002). Therefore, to test predictions from the PAH and the RFM while also controlling for any serial position effects, we used blocks of 4, 5, and 6 items and systematically varied the order of the produced and silent blocks within each list. As shown in Table 1, this 6-condition design created key intersections or transitions at serial positions 4, 5, and 6 which allowed us to contrast the predictions derived from the PAH with those of the RFM.

Consider a basic comparison involving both blocks. For illustrative purposes, we will focus on the key contrasts at position 4. In condition D, the PAH predicts that participants will experience the highest level of disruption as they are aware that the next item will be read aloud in front of the experimenter (Forrin et al., 2019). Therefore, the recall of this silent item should be lower than that of other silent items at position 4 which are not followed by a produced item; this happens in conditions E and F. Alternatively, according to the RFM, recall of the 4^{th} item in condition A should be considerably improved, relative to conditions B and C, as said item is the only produced item in this position that is followed by a silent item. In this situation, the upcoming silent item would be unable to interfere with the produced item's features as much as other produced items would, leaving the produced item with more intact features and a higher probability of recall (Cyr et al., 2022; Saint-Aubin et al., 2021).

Finally, our experimental design also pits the PAH and the RFM head-to-head. For instance, when conditions E and F are compared in position 5, the PAH would predict lower recall for Condition E due to the following item in this condition being produced. However, the RFM would not expect any difference in recall between conditions E and F. First, the efficiency of rehearsal drops significantly after the first few items of a list, a behavior that is integrated in the RFM (see Bhatarah et al., 2009; Rundus, 1971). Therefore, this far into the list, the

introduction of a produced item would have minimal impact on rehearsal; the expectation is hence that rehearsal will be very limited—and equivalent—for the items compared in conditions E and F (Saint-Aubin et al., 2021). Second, based on the retroactive interference mechanism included in the RFM, item 6 should interfere with the prior encoding of item 5. However, this process is based on feature similarity, and since the features shared between a produced or silent item in position 6 and the silent item in position 5 should be roughly equivalent on average, the RFM predicts that the similar number of intact features for conditions E and F would lead to no difference in subsequent recall performance.

Predictions from both accounts for all key contrasts are summarized in Table 2. As can be seen at a given serial position, when produced items are presented in the first block, the RFM always predicts better recall of a produced item that is followed by a silent item. When silent items are presented in the first block, the RFM predicts a similar level of performance for silent items whether or not they are followed by a produced item. Contrary to the RFM, the PAH makes no predictions for produced items but does make clear predictions for silent items: the latter will not be recalled as well when they are followed by a produced rather than a silent item if, and only if, the experimenter is present in the room. Therefore, following Forrin et al.'s (2019) procedure, we removed the experimenter from the room in Experiment 1B to suppress any performance anticipation. Indeed, Forrin et al. showed that without the experimenter, participants' memory for silent items was the same whether the upcoming word ought to be read aloud or silently. In sum, according to the PAH, all contrasts for which a difference was predicted in Experiment 1A with the experimenter present are expected to show no difference in Experiment 1B. However, the RFM's predictions remain the same whether or not the experimenter is present in the room. Finally, in the following experiments, fits of the RFM will be presented alongside the results to facilitate comparison, although details of the model will be provided later in the paper.

Experiment 1A: Immediate Recall with Experimenter Present

Method

Transparency and Openness. We reported all manipulations and measures and have made all stimuli, data, program codes and R Markdowns publicly available on the Open Science Framework repository (OSF). Study designs and analyses were not preregistered. Experiments were approved by the Ethics Board for Research Involving Humans of Université de Moncton.

Sample Size Calculation. Our target sample size was based on the effect size from Forrin et al.'s (2019) third experiment for the interaction between the status of the target word (produced vs. silent) and the status of the next word (produced vs. silent). Using G*Power 3.1.9.7 (Faul et al., 2007), we estimated the size of the interaction as being $f = 0.25$. Based on this effect size, we conducted an a priori repeated measures analysis with an alpha of .05, power of .95 and default parameters for the correlation between repeated measures and the non-sphericity correction. The analysis revealed that 36 participants would be enough to uncover the interaction.[1] However, due to uncertainty associated with the changes in memory task, we chose to overpower our design to maximize the likelihood of finding decisive evidence in favor of or against this interaction.

Importantly, since the frequentist approach is unable to estimate the likelihood of finding evidence for the absence of a difference (H0), we turned to a Bayesian approach. We used the Bayes Factors Design Analysis (BFDA; Schönbrodt & Wagenmakers, 2018), default parameters, 10,000 simulations for a non-directional Bayesian paired *t*-test (using BF > 3 as our decision criteria), and a null effect size ($d = 0.00$). The simulation for a sample of 50 participants revealed that 1.4% of samples showed evidence for H1 (BF > 3), 19.3% were inconclusive (0.3333 < BF <

[1] It is important to note that this a priori power analysis was conducted before we became aware of a limitation in G*Power that prevents accurate power calculations for designs involving more than one within-subject variable, such as the design used in the present study. Nevertheless, we report the original calculations to maintain transparency in our research process, while cautioning readers about this issue when using G*Power for similar analyses.

3), and 79.3% showed evidence for the null (BF < 0.3333). Finally, after recruiting the targeted sample of 50 participants per experiment, we used a sequential testing approach with Bayesian statistical tests (see, Schönbrodt et al., 2017) and increased the sample size for each experiment by 10 so that the majority of the Bayes Factors (BFs) reached the criterion (BF > 6 for the null or the alternative hypothesis). Therefore, the final sample size was 60 participants per experiment.[2]

Participants. We recruited 60 students (47 women, 13 men, *M* age = 20.53 years, *SD* = 2.25 years) from Université de Moncton, who received course credits or an entry in a $100 monthly draw as compensation. Inclusion criteria for all experiments were: (1) being a native French speaker; (2) being aged between 18 and 30 years; (3) having normal or corrected-to-normal vision; and (4) having never participated in a study on the production effect. To ensure that all participants were completely naive to the task, we asked participants if they remembered taking part in a similar experiment and we looked through our database where all participants are listed along with the experiments that they had completed. All participants gave their free and informed consent prior to the beginning of the study. However, one participant who did not follow the instructions during the task was removed and replaced.

Stimuli. We sampled from the *Lexique* database (New et al., 2004) to create a pool of 620 disyllabic French nouns with 8 letters, between 3 and 7 phonemes, and a frequency ranging from 0 to 583.45 occurrences per million (*M* = 7.49, *SD* = 31.36). For each participant, we sampled without replacement from the word pool to create 62 lists of 10 words, from which 2 lists served as practice trials and 60 as experimental trials. Therefore, each participant was presented with 62 different lists of words. In addition, 10 of the 60 experimental lists were randomly assigned to

[2] To facilitate comparison between the frequentist and Bayesian approaches, we used G*Power and computed a sensitivity power analysis for a paired samples two-tailed *t*-test. Based on a sample size of 60 participants, an alpha of .05 and power of .95, the analysis revealed that the smallest effect size our design could detect was *dz* = .47.

each of the 6 conditions illustrated in Table 1. For each condition, silent and produced items were blocked: Produced items were presented first in three conditions; silent items were presented first in the other three conditions. Finally, all blocks had between 4 and 6 items, which allowed us to test predictions derived from the PAH and the RFM at positions 4, 5, and 6.

Design and Procedure. We used a 10 X 6 repeated measures design with serial position (1-10) and list condition (A-F) as factors. As mentioned earlier, the words in each list were randomized for each participant. While an equal number of lists were assigned to each list condition, the order in which those lists were presented was also randomized. Each participant was individually tested in a quiet room and in a single session lasting approximately 45 minutes. Following Forrin et al.'s (2019) procedure, participants sat at about 60 cm from the monitor while the experimenter sat at a table approximately 3-4 feet behind them, out of their line of sight. All research assistants who conducted the experimentation were instructed to be friendly and to remain as constant as possible in their interactions with participants. They presented the same instructions to each participant and answered all questions prior to the beginning of the task. Then, after receiving all instructions, participants who consented to take part in the study initiated the first practice trial by pressing the space bar of the keyboard.

The experiment was controlled with PsyToolkit (Stoet, 2010, 2017) and the stimuli were displayed in lowercase blue or white 20-point Times New Roman on a black background (see Figure 3 for an illustration). Participants were instructed to read the blue words aloud and to read the white words silently, without mouthing or whispering the words. They were also instructed to try to remember all the words, regardless of their color. Each word was presented for 2 seconds (2000 ms on, 0 ms off). After each list was presented, participants had to recall as many words as possible from that list, regardless of their presentation order. Participants typed the words on the keyboard and pressed the enter key after each word to register their answer. Recalled words

remained on the screen after being typed. When they were done recalling the items, participants were instructed to press the enter key to skip the remaining items and to move on to the next trial. This procedure was then repeated for each of the 62 trials (including two practice trials and 60 experimental trials).

As shown in Figure 3, during word presentation, a horizontal line was displayed under each word, its color matching the color of the word above it. Participants were instructed that the color of the line was a further indication of whether the word had to be read aloud (blue) or read silently (white), and that it would always match the color of the word.[3] To allow participants to anticipate how to process the upcoming words in the list, we added four horizontal lines on each side of the central line. Participants were told that the lines on the right indicated the color of the upcoming words and that the lines on the left indicated the color of the previous words. For example, if the first line to the right of the central line was blue, participants knew they would have to read the next word aloud. They were also warned that after each word was presented, the lines would all move one spot to the left, like a conveyor belt, and that the next word would appear over the new central line, matching its color.

Data Analysis. Participants' answers were considered correct if a word was recalled, independently of its output serial position. Correct responses were then analyzed as a function of their input serial position (1 – 10) and list condition (A – F). Prior to conducting any analyses, we checked participants' answers for misspelling and corrected all responses that could be identified without any ambiguity (e.g., letter repetitions: moountain instead of mountain; letter omissions: montain instead of mountain; or letter substitutions: mountein instead of mountain). All reported

[3] It is worth noting that previous studies have shown that manipulating the color of the produced/silent items at encoding or the match between the color of the words at encoding and recall had no effect on recall performance (see, e.g., Cyr et al., 2022; Saint-Aubin et al., 2021).

analyses were based on corrected data but, importantly, except for marginally reduced overall performance, results with uncorrected spelling were the same. Importantly, although no formal pre-registration was done for this study, all statistical analyses reported below (except the combined analyses that were requested during the review process) were planned.

Bayesian inferences conducted here were driven by Bayes Factor (BF) paired *t*-tests and ANOVA analyses, which were conducted in the *R* software (*R* Core Team, 2023) by using the "BayesFactor" package and the default parameters (Version 0.9.12-4.2; Morey & Rouder, 2018; Rouder et al., 2009, 2012). After introducing participants as a random factor, main effects and interactions from our BF ANOVA were tested by comparing models without each effect to the full model (Condition + Serial position + Condition: Serial position + Participant). Proportional errors were lower than 5% for all BFs, which were estimated through Monte Carlo simulations using 100,000 iterations. Finally, predictions from the RFM and the PAH were also tested with BF *t*-tests comparing mean recall between list conditions at positions 4, 5, and 6.

Corresponding *F* ratios and partial eta squares, computed with the "ez" package (Version 4.4-0; Lawrence, 2016) and the "lsr" package (Version 0.5; Navarro, 2015), were also reported as complementary descriptive information. Finally, the results from our BF ANOVA and *t*-tests are reported based on a nomenclature in which BF_{10} represents evidence in favor of an effect or a difference (H1) and BF_{01} ($1/BF_{10}$) represents evidence against an effect or a difference (H0). We interpreted our findings using guidelines by Goss-Sampson (2020), where BF_{10} or BF_{01} values between 1 and 3 indicate anecdotal evidence, whereas values between 3 and 10, 10 – 30, 30 – 100, and over 100 respectively indicate moderate, strong, very strong, and decisive evidence. Finally, we followed the same data analysis procedure for all experiments.

Results

Overall Free Recall Performance. Proportions of correct free recall as a function of list condition and serial position are presented in Figure 4. Across all conditions, there was a large recency effect but no primacy effect, which is typical of immediate free recall with 10-word lists (Grenfell-Essam & Ward, 2012). Also, as predicted by the RFM, for list conditions starting with a produced block (A, B, and C), we found a systematic peak in free recall performance for the last produced item, which had no equivalent for the first produced item in conditions starting with a silent block (D, E, and F). The 6 X 10 repeated measures ANOVA also supported these trends by showing decisive evidence against a main effect of list condition, $F(5, 295) = 5.97$, $\eta^2_p = .09$, $BF_{01} = 300.70$, in favor of a main effect of serial position, $F(9, 531) = 320.05$, $\eta^2_p = .84$, $BF_{10} > 10{,}000$, and in favor of an interaction between serial position and list condition, $F(45, 2655) = 30.58$, $\eta^2_p = .34$, $BF_{10} > 10{,}000$. Finally, we further investigated the interaction by testing the 12 contrasts shown in Table 2 with a series of BF *t*-tests (see Figure 5 for the means).

Production Anticipation Hypothesis. First, we tested predictions from the PAH with 6 contrasts involving silent items. Only 3 of the 6 contrasts showed lower recall when the silent item was the last of its block and was followed by a produced item. Specifically, at each of the three critical positions (4, 5, and 6), only one of the two contrasts revealed lower recall of a silent item followed by a produced item rather than by another silent item (Position 4 : D < F, $BF_{10} = 4.34$; D = E, $BF_{01} = 5.16$; Position 5 : E < A, $BF_{10} = 968.31$; E = F, $BF_{01} = 5.65$; Position 6 : F ? A, $BF_{10} = 1.91$; F < B, $BF_{10} = 1{,}483.69$).

Revised Feature Model. We tested critical predictions from the RFM with 6 contrasts involving produced items. Here, all contrasts showed better recall when the produced item was the last of its block and was followed by a silent item. More specifically, at positions 4, 5, and 6, both contrasts revealed better recall of the list condition where a produced item was followed by a silent item rather than by another produced item (Position 4: A > B, $BF_{10} = 236.95$; A > C, BF_{10}

> 10,000; Position 5: B > C, BF_{10} > 10,000; B > D, BF_{10} > 10,000; Position 6, C > D, BF_{10} > 10,000; C > E, BF_{10} = 3.13).

Discussion

Overall, the results of Experiment 1A with mixed lists, an immediate free recall task, and the experimenter in the room showed a typical production effect and a large interaction between list condition and serial position. After decomposing said interaction, we found near-perfect support for the predictions derived from the RFM. Specifically, all contrasts involving produced items at positions 4, 5, and 6 revealed better recall for conditions in which a produced item was followed by a silent item (last of its block). In addition, of the 3 critical contrasts for silent items pitting predictions from the two accounts against each other, 2 favored the RFM by showing no difference between list conditions. This fits well with the RFM's retroactive interference and rehearsal mechanisms, showing that the last produced item of the block benefited from more intact features and higher distinctiveness, whereas all silent items at positions 4 and 5 had similar amounts of intact features and rehearsal (see Cyr et al., 2022; Gionet et al., 2022; Saint-Aubin et al., 2021).

Conversely, our results revealed only partial support for the PAH's predictions, with 3 of the 6 contrasts involving silent items showing lower recall for list conditions in which a silent item was followed by a produced item (last of its block). This recall cost was predicted due to the experimenter's presence and to participants being able to predict exactly when they would have to read words aloud or silently, which should have resulted in enhanced performance anxiety and disrupted encoding of the silent item preceding a produced item (Forrin et al., 2019). However, of the remaining 3 contrasts, 2 showed enough evidence to support the null hypothesis (H0) and 1 showed only anecdotal evidence. In sum, results from Experiment 1A strongly supported the predictions derived from the RFM but offered more ambiguous support for the PAH.

Experiment 1B: Immediate Recall Without Experimenter

Experiment 1B was identical to Experiment 1A, except for the experimenter's presence during the task, and served as a critical test of the predictions derived from the PAH. Specifically, based on the results from their fourth experiment, Forrin et al. (2019) suggested that removing the experimenter from the room during the task should eliminate any cost inflicted on silent items by performance anticipation or social discomfort. In the present study, erasing this cost should lead to no differences for silent items between list conditions at positions 4, 5, and 6. In contrast, based on the RFM's predictions, this manipulation should not have any effect on the results for produced or for silent items (see Table 2 for the updated predictions).

Method

Participants. A novel sample of sixty students from Université de Moncton (41 women, 17 men, 2 non-binary individuals, *M* age = 21.81 years, *SD* = 2.96 years) were recruited based on the same inclusion criteria. Once again, participants received course credits or an entry in a $100 monthly draw as compensation. One participant was removed and replaced because it was an outlier with an exceedingly low recall performance (see data on OSF).

Stimuli, Design, and Procedure. The stimuli, design, and procedure were identical to those used in Experiment 1A, except for the experimenter's presence in the room during the task. Echoing Forrin et al.'s (2019) procedure from their fourth experiment, the experimenter brought the participants to the laboratory room, provided instructions for the task, and answered all their questions. Then, after supervising the practice trials to assess participants' compliance with the instructions, the experimenter left the room and remained absent for the entire session.

Results

Overall Free Recall Performance. Proportions of correct free recall as a function of list condition and serial position are shown in Figure 6. Overall, results are nearly identical to those

of Experiment 1A, with a large recency effect and no primacy effect. Also, although there are no trends across conditions for silent items, recall performance still reaches a systematic peak for the last produced item in lists beginning with the produced block. Finally, the 6 X 10 repeated measures ANOVA showed decisive evidence against an effect of list condition, $F(5, 295) = 4.45$, $\eta^2_p = .07$, $BF_{01} = 846.10$, in favor of an effect of serial position, $F(9, 531) = 307.03$, $\eta^2_p = .84$, $BF_{10} > 10{,}000$, and in favor of the interaction, $F(45, 2655) = 37.68$, $\eta^2_p = .39$, $BF_{10} > 10{,}000$.

Production Anticipation Hypothesis. Once again, we tested predictions from the PAH with a series of BF *t*-tests (see Figure 7). Critically, only 1 of the 6 contrasts involving silent items supported the PAH's predictions by showing no difference between conditions. At position 4, both contrasts showed lower recall of the silent item that was followed by a produced item rather than by another silent item (D < E, $BF_{10} = 4.74$; D < F, $BF_{10} = 7.61$). At position 5, one of the contrasts showed lower recall of the silent item followed by a produced item, but the other one showed no difference (E < A, $BF_{10} > 10{,}000$; E = F, $BF_{01} = 3.02$). Finally, at position 6, both contrasts showed lower recall of the silent item that was followed by a produced item (F < A, $BF_{10} = 152.62$; F < B, $BF_{10} > 10{,}000$)

Revised Feature Model. As in Experiment 1A, all 6 contrasts involving produced items supported the RFM's predictions, revealing better recall of the produced item that was followed by a silent item than by another produced item. More specifically, at positions 4, 5, and 6, both contrasts showed better recall of the condition where a produced item was the last of its block (Position 4: A > B, $BF_{10} = 213.06$; A > C, $BF_{10} > 10{,}000$; Position 5: B > C, $BF_{10} = 265.71$; B > D, $BF_{10} = 47.55$; Position 6: C > D, $BF_{10} > 10{,}000$; C > E, $BF_{10} = 6.46$).

Combined Analyses

Per request from the reviewers, we further explored the effect of the experimenter's presence during the task by combining and analyzing the full data from Experiments 1A and 1B.

We conducted a 2 X 6 X 10 mixed ANOVA with Experimenter presence (present or absent) as a between-subjects factor, and List Condition (A – F) and Serial Position (1 – 10) as within-subject factors. For main effects, we found moderate evidence in favor of an effect of list condition, $F(5, 590) = 9.83$, $\eta^2_p = .08$, $BF_{10} = 4.42$, and decisive evidence in favor of an effect of serial position, $F(9, 1062) = 626.52$, $\eta^2_p = .84$, $BF_{10} > 10{,}000$. However, contrary to predictions from the PAH, we found moderate evidence against an effect of experimenter presence, $F(1, 118) = 2.25$, $\eta^2_p = .02$, $BF_{01} = 9.38$. For two-way interactions, we found decisive evidence in favor of an interaction between list condition and serial position, $F(45, 5310) = 67.11$, $\eta^2_p = .36$, $BF_{10} > 10{,}000$, but also decisive evidence against the interaction between experimenter presence and list condition, $F(5, 590) = 0.41$, $\eta^2_p < .01$, $BF_{01} > 10{,}000$, and decisive evidence against the interaction between experimenter presence and serial position, $F(9, 1062) = 0.87$, $\eta^2_p = .01$, $BF_{01} > 10{,}000$. Finally, we found decisive evidence against the three-way interaction between experimenter presence, list condition, and serial position, $F(45, 5310) = 1.33$, $\eta^2_p = .01$, $BF_{01} > 10{,}000$. Detailed results for the individual contrasts and all *R* Markdowns are available on OSF. Overall, these results provide critical evidence against the PAH, showing that manipulating the experimenter's presence during the task had no effect on the magnitude of the production effect.

Discussion (Experiments 1A & 1B)

Despite removing the experimenter from the room in Experiment 1B, we successfully replicated the findings of Experiment 1A. Overall, all 12 contrasts involving produced items in Experiment 1 revealed higher recall of the list conditions where a produced item was followed by a silent item. This fits well with the RFM, supporting the idea that combining the additional features of produced items with similarity-based retroactive interference allows the last produced item to benefit from enhanced distinctiveness (Cyr et al., 2022; Dauphinee et al., 2024; Saint-Aubin et al., 2021). In addition, of the 3 contrasts involving silent items for which the RFM and

PAH were head-to-head in Experiment 1A, 2 favored the RFM by showing evidence against a difference between conditions. As a reminder, the RFM assumes that silent items at positions 4 and 5 should benefit from similar amounts of features and rehearsal, regardless of whether they are followed by a produced item or a silent item (Cyr et al., 2022; Saint-Aubin et al., 2021).

Furthermore, our results provided very limited support for the PAH's predictions. In fact, even after making the occurrence of produced and silent items predictable and manipulating the experimenter's presence in the room, only 4 of the 12 contrasts involving silent items were in the expected direction, i.e., a drop in silent item recall when they are followed by produced items if, and only if, the experimenter is present in the room (Forrin et al., 2019). Furthermore, the joint analysis revealed no main effect of the presence of the experimenter nor any interactions with key factors. These results contradicting the predictions of the PAH are in line with those of Wakeham-Lewis et al. (2022) who also manipulated the presence of the experimenter and found no evidence for the role of performance anxiety in the production effect. In summary, our results strongly favor the RFM over the PAH by showing that an explanation of the production effect based on basic memory processes such as interference and rehearsal fits the data better than an explanation based on social factors.

Experiment 2

Although the results from our first experiment with an immediate free recall task strongly support the RFM's predictions (Cyr et al., 2022; Saint-Aubin et al., 2021), they only offer partial support to those derived from the PAH (Forrin et al., 2019). However, despite their clarity, it can be argued that these results are not the most critical due to the production effect being typically studied in long-term memory (LTM) tasks. Therefore, it is possible that the limited support found for the PAH's predictions is due to fundamental differences between short-term (STM) and long-term memory. In support of this view, many factors influencing performance in STM tasks have

no effect or even opposite effects in LTM tasks (e.g., McCabe, 2008). For instance, concurrent articulation of irrelevant items—known as articulatory suppression—reduces immediate recall but not delayed recall (see, Camos & Portrat, 2015).

To further test predictions derived from the PAH and the RFM in Experiment 2, we used a delayed free recall task during which the experimenter was either present (Experiment 2A) or absent (Experiment 2B). We also shifted from a STM task to an LTM task by adding a 30-second filled retention interval between item presentation and recall. This duration was selected because it exceeds the known duration of information in STM (Cowan, 2017; Prisko, 1963). In addition, the results observed after 30 seconds or after a longer interval (e.g., 2 minutes) are the same (Cyr et al., 2022). During this interval, participants performed a parity judgment task, which has been shown to block working memory processing (Jonker et al., 2014). Importantly, since both the PAH and the RFM assume that the processes—cognitive or social—operating in immediate and delayed free recall should be identical, the predictions for Experiment 2 remain the same.

Experiment 2A: Delayed Recall with Experimenter

Method

Participants. Another sample of sixty students (43 women, 17 men, *M* age = 21.06 years, *SD* = 2.59 years) from Université de Moncton who met all inclusion criteria and who had never taken part in a study on the production effect (including our first two experiments) were recruited and received course credits or an entry in a $100 draw as compensation.

Stimuli, Design, and Procedure. The stimuli were created from the same word pool as in Experiment 1. However, given the addition of a 30-second retention interval, we slightly reduced the number of lists to 50 lists of 10 words to keep the duration of the experiment within reasonable limits. The design and procedure were the same as in Experiment 1A, except for the following changes. First, participants completed 2 practice trials and 48 experimental trials (8 per

list condition) in a single session lasting approximately 90 minutes. Second, a 30-second parity judgment task was added between item encoding and recall (see Cyr et al., 2022). Here, single integers ranging from 0 to 9 were sequentially displayed at the center of the screen. Participants were instructed to answer as quickly and as accurately as possible by pressing the "M" key of their keyboard if the stimulus was an even number and the "Z" key if the stimulus was an odd number. This task was self-paced but lasted 30 seconds for all participants.

Results

Overall Free Recall Performance. Proportions of correct recall as a function of list condition and serial position are shown in Figure 8. Overall, there is still a large recency effect and a more modest primacy effect, and across serial positions, performance is generally better for produced items. As in Experiment 1, while there are no trends across list conditions for silently read items, there is still a systematic peak for the last produced item in lists beginning with the produced block (A, B, and C). To support this, the 6 X 10 repeated measures ANOVA revealed decisive evidence against an effect of list condition, $F(5, 295) = 1.98$, $\eta^2_p = .03$, $BF_{01} > 10{,}000$, but decisive evidence in favor of an effect of serial position, $F(9, 531) = 79.14$, $\eta^2_p = .57$, $BF_{10} > 10{,}000$, and in favor of the interaction, $F(45, 2655) = 26.90$, $\eta^2_p = .31$, $BF_{10} > 10{,}000$. We once again investigated the large interaction by testing the contrasts shown in Table 2 (see Figure 9).

Production Anticipation Hypothesis. As in Experiment 1A, 3 of the 6 contrasts involving silent items showed lower recall of the condition where a silent item was the last of its block and was followed by a produced item. At position 4, neither of the 2 contrasts showed lower recall of the silent item that was followed by a produced item rather than by another silent item (D = E, $BF_{01} = 3.88$; D ? F, $BF_{10} = 1.33$). At position 5, only one of the 2 contrasts showed lower recall of the silent item that was followed by a produced item (E < A, $BF_{10} = 8.70$; E = F,

BF_{01} = 5.97). Finally, at position 6, both contrasts showed lower recall of the silent item that was followed by a produced item (F < A, BF_{10} = 319.59; F < B, BF_{10} = 25.01).

Revised Feature Model. Once again, 2 of the 3 critical contrasts involving silent items and opposing predictions from the two accounts favored the RFM by showing evidence against a difference between conditions (see Table 2). In addition, all contrasts involving produced items at position 4, 5, and 6 showed better recall of the condition where a produced item was the last of its block and followed by a silent item rather than by another produced item (Position 4: A > B, BF_{10} = 47.22; A > C, BF_{10} = 4,988.32; Position 5: B > C, BF_{10} = 403.97; B > D, BF_{10} > 10,000; Position 6: C > D, BF_{10} > 10,000; C > E, BF_{10} = 46.87).

Experiment 2B: Delayed Recall without Experimenter

Experiment 2B was identical to Experiment 2A, except that, as in Experiment 1B, the experimenter was absent from the room during the task. Based on the PAH, no difference was expected between list conditions due to the additional social cost on silent items being erased. However, based on the RFM, we expected the same pattern of results as in Experiment 2A.

Method

Participants. Sixty additional students (44 women, 15 men, 1 non-binary individual, *M* age = 21.53 years, *SD* = 2.97 years) who met all inclusion criteria and who had never taken part in a production effect experiment (including the previous three experiments) were recruited and received course credits or an entry in a $100 draw as compensation.

Stimuli, Design, and Procedure. The stimuli, design, and procedure were almost identical to those of Experiment 2A. However, as in Experiment 1B, after giving the instructions for the task, answering all questions, and supervising the practice trials to assess task compliance, the experimenter left the laboratory room and remained absent for the entire session.

Results

Overall Free Recall Performance. Proportions of correct recall as a function of list condition and serial position are shown in Figure 10. Overall, results were nearly identical to those from Experiment 2A, with a large recency effect and a more modest primacy effect. Performance was still generally better for produced items across all serial positions, and there was still a peak in free recall performance (although smaller) for the last produced item in lists starting with the produced block. As expected, the 6 X 10 ANOVA showed decisive evidence against an effect of list condition, $F(5, 295) = 3.17$, $\eta^2_p = .05$, $BF_{01} > 10{,}000$, in favor of an effect of serial position, $F(9, 531) = 69.52$, $\eta^2_p = .54$, $BF_{10} > 10{,}000$, and in favor of the interaction, $F(45, 2655) = 19.50$, $\eta^2_p = .25$, $BF_{10} > 10{,}000$. Finally, we tested this interaction through the contrasts shown in Table 2 (see Figure 11 for the means).

Production Anticipation Hypothesis. Only 2 of the 6 contrasts involving silent items supported the PAH by showing evidence against a difference between conditions. At position 4 and position 5, one of the two contrasts showed no difference between the silent items followed by a produced item and those followed by another silent item (Position 4: D = E, $BF_{01} = 3.78$; D ? F, $BF_{01} = 2.61$; Position 5: E ? A, $BF_{10} = 1.53$; E = F, $BF_{01} = 7.08$). However, at position 6, both contrasts showed lower recall of the condition where a silent item was followed by a produced item (F < A, $BF_{10} = 59.77$; F < B, $BF_{10} = 236.70$).

Revised Feature Model. Of the 6 contrasts involving produced items, 5 showed better recall of the list condition where a produced item was the last of its block and was followed by a silent item. At position 4, both contrasts showed better recall of the produced item followed by a silent item (A > B, $BF_{10} = 16.33$; A > C, $BF_{10} = 25.77$). At position 5, only one of the two contrasts showed better recall of the produced item followed by a silent item (B ? C, $BF_{10} = 2.09$;

B > D, BF_{10} = 15.18). Finally, at position 6, both contrasts showed better recall of the produced item followed by a silent item (C > D, BF_{10} = 3,128.25; C > E, BF_{10} = 5.56).

Combined Analyses

Once again, per request from the reviewers, we explored the effect of the experimenter's presence during the task by combining and analyzing the full data from Experiments 2A and 2B. Results from the main analysis are reported below, but detailed results for the individual contrasts can be found on OSF along with the *R* markdowns. As for Experiment 1, we conducted a 2 X 6 X 10 mixed ANOVA with Experimenter presence (present or absent), List Condition (A – F), and Serial Position (1 – 10) as factors. First, we found decisive evidence in favor of an effect of serial position, $F(9, 1062) = 147.35$, $\eta^2_p = .56$, $BF_{10} > 10{,}000$, but decisive evidence against an effect of list condition, $F(5, 590) = 4.85$, $\eta^2_p = .04$, $BF_{01} = 2{,}229.05$. Importantly, contrary to predictions from the PAH, we found strong evidence against an effect of experimenter presence, $F(1, 118) = 1.43$, $\eta^2_p = .01$, $BF_{01} = 12.31$. For two-way interactions, we found decisive evidence in favor of an interaction between list condition and serial position, $F(45, 5310) = 45.06$, $\eta^2_p = .28$, $BF_{10} > 10{,}000$, but decisive evidence against both the interaction between experimenter presence and list condition, $F(5, 590) = 0.28$, $\eta^2_p < .01$, $BF_{01} > 10{,}000$, and the interaction between experimenter presence and serial position, $F(9, 1062) = 1.16$, $\eta^2_p = .01$, $BF_{01} > 10{,}000$. Finally, we also found decisive evidence against the three-way interaction between list condition, serial position, and experimenter presence, $F(45, 5310) = 1.05$, $\eta^2_p = .01$, $BF_{01} > 10{,}000$.

Discussion (Experiments 2A & 2B)

Despite the methodological change (i.e., going from an STM task to an LTM task), the results of Experiment 2 replicated those of Experiment 1. First, the overall results from Experiment 2 revealed a clear (but slightly smaller) production effect along with an interaction between list condition and serial position. In addition, as we typically observe after adding a

filled retention interval, the recency effect was smaller than in Experiment 1, while the primacy effect remained similar (e.g., Tan & Ward, 2000). Importantly, the presence of a large recency effect despite the presence of a 30 second distraction interval aligns with data from the continual distractor paradigm in free recall in which a large recency effect is observed despite the inclusion of a distractor task after the last item (see Neath, 1993; Watkins et al., 1989).

Results from Experiment 2 also extended the support for the RFM, with 11 of the 12 key contrasts involving produced items showing better recall of the list condition where a produced item was followed by a silent item. This not only fits with the idea of retroactive interference and enhanced local distinctiveness benefiting recall of the last produced item, but also suggests that those mechanisms, as implemented in the RFM, are operating in the same fashion across STM and LTM recall tasks (see Cyr et al., 2022; Saint-Aubin et al., 2021). Finally, as in Experiment 1, we were unable to reliably replicate the memory cost for silent items preceding a produced item that was observed by Forrin et al. (2019). Specifically, only 5 of the 12 contrasts involving silent items across Experiments 2A and 2B were consistent with the PAH's predictions. In addition, 2 of the 3 critical contrasts testing the opposing predictions derived from the two accounts in Experiment 2A favored the RFM by showing evidence against a difference between conditions (see Table 2 for a summary of all results). In sum, results from the current study offer clear support for the RFM's rehearsal and retroactive interference mechanisms across a series of four large-scale experiments with different methodologies.

Computational Modeling

Detailed Description of the RFM

The Revised Feature Model (RFM) is an adaptation of the Feature Model (Nairne, 1988, 1990; Neath & Nairne, 1995; Neath & Surprenant, 2007). The key differences between the original Feature Model and the RFM are the inclusion of a rehearsal process and a change in the

way overwriting works. The RFM was originally seen as a model of immediate serial recall but was easily adapted to deal with immediate and delayed free recall (see Cyr et al., 2022). We will begin by explaining how information about the items is encoded, which is common across all recall tasks. Then, we will describe the retrieval process in free recall.

In the RFM, as in the original Feature Model, items are represented by two types of features: (1) modality-dependent features, related to physical presentation conditions such as font size or voice quality, and (2) modality-independent features, generated by internal processes of categorization and identification. During presentation, items simultaneously generate traces in primary and secondary memory. In both cases, items are represented by vectors of features, with each (randomly generated) feature taking values 1-3. Traces in primary memory are then subject to similarity-based retroactive interference: If feature i of item n is identical to feature i of item n-m, then this feature of item n-m will be overwritten (set to 0) with probability $e^{-\lambda(m-1)}$. This change from the original Feature Model allows retroactive interference to operate further back than just the most recent item. This can be thought of as the limiting case where $\lambda \to \infty$. This change was introduced to deal with the fact that the modality effect extends further back than the final item in the lists (Saint-Aubin et al., 2021). However, until now, this retroactive interference mechanism had never been directly tested.

After presentation of all items, a final overwriting of modality-independent features takes place due to continuing internal thought activity in preparation for recall. For delayed recall, there is an additional overwriting step affecting both modality-dependent and modality-independent features, representing the additional impact on trace veracity.

If overwriting degrades traces in primary memory, within the RFM, we also assume that a process of rehearsal can act to restore some overwritten features. Specifically, after each item is

presented, there is an attempt to rehearse all previous items. This rehearsal cycle, which runs after presentation of item n, will successfully restore any overwritten features 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). Importantly, the value of r is assumed to depend on what happens at or shortly after item presentation. For example, we assume that producing the items hinders rehearsal and that having to perform a filler task, such as saying an unrelated word, can also reduce rehearsal (Saint-Aubin et al., 2021).

The rationale for the exponential suppression of rehearsal as a function of list position is that rehearsal has been shown to happen less frequently as list length increases (Bhatarah et al., 2009; Rundus, 1971). However, in our experiments, produced and silent items naturally fall into two blocks. Given this, it is possible that participants are prompted to restart rehearsal attempts when the block (produced / silent) changes. Preliminary fitting suggested that such a mechanism does indeed capture behavior better than a model without it. Thus, we assume that the tendency to rehearse items ‘resets’ at the start of a new block. However, there are multiple ways in which this could happen. In what follows, we assume that each block is, for the purposes of rehearsal, treated as a separate list so that a change of modality triggers a restarting of rehearsal without the previous block being rehearsed again. In Appendix B, we consider the alternate possibility that restarting rehearsal includes all previously presented items.

Finally, order information is encoded in the same way in the RFM as in the original Feature Model. More specifically, each presented item is tagged with its position in the list. In the

RFM, 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).

The RFM for Free Recall

The RFM was developed by Saint-Aubin et al. (2021) to account for the production effect in immediate serial recall and order reconstruction. Shortly after, Cyr et al. (2022) described how to adapt the model to deal with delayed free recall. Importantly, a key characteristic of the RFM is that the encoding and retrieval processes are completely separate, which means that it can be readily adapted to different retrieval methods.

In the RFM, retrieval is similarity based. The similarities between cues and items stored in secondary memory are computed via Shepard's Law (1987). Specifically, the similarity between a cue i and an item in secondary memory j is given by,

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

The distance, d_{ij}, between cue and item is related to the proportion of mismatching features,

$$d_{ij} = a \times M_{ij}$$

where *a* is a scaling constant. For free recall, the activation of an item is given as the sum of the similarities between that item and all cues. Specifically, each item i in secondary memory gets an activation, $p(i) \sim e^{\frac{s(i)}{\tau}}$, where τ is a temperature parameter that controls how deterministically the item with the highest similarity is chosen. We also allow for the possibility that no secondary memory trace matches the primary memory trace well enough to be recalled. We do this by including an extra 'null' possibility which has constant similarity between itself and all primary memory traces. The model also includes a step where multiple recalls of the same item are suppressed and instead result in an omission. Full details are provided in Cyr et al. (2022).

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, 2009) hereafter referred to as ABC-PRC. Full details are given in Appendix A and code to fit the model can be found on OSF. We fit all six conditions from a given experiment at once, assuming that all parameters not directly set by the trial design were fixed between conditions. This is more challenging for the model than fitting a single curve at a time.

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. The parameters that were allowed to vary fall into two groups; first, we have the distance scaling parameter a, the 'floor' similarity which controls omissions, and the temperature parameter τ. These influence overall accuracy but are less theoretically significant. Second, we have rehearsal parameters for the silently read, r_S, and produced items, r_A, and the value of λ which controls how far back overwriting occurs. These are more theoretically interesting to examine. For each experiment, there are therefore six parameters that were allowed to vary, against 60 data points that are being fit (values of nonvarying parameters are given in Table 3). For all model fits, we report the results by showing the means and 95% HDIs of the posteriors of the model predictions for each serial position in the different conditions (see Figures 4, 6, 8, and 10). We also report medians and 95% HDIs for the posteriors of the parameters allowed to vary in Table 3.

General Discussion

In this article, we simultaneously tested two distinct accounts of the production effect with the aim of contributing to our understanding of the basic processes — social and/or cognitive — that underlie short-term and long-term memory. First, we examined a new set of predictions derived a priori from the RFM, which has recently been called upon to account for performance in a variety of tasks, including immediate serial recall, free recall, and verbal / visuo-spatial order reconstruction (Cyr et al., 2022; Dauphinee et al., 2024; Poirier et al., 2019; Saint-Aubin et al., 2021, 2023, 2024). Specifically, we focused on the RFM's retroactive interference mechanism, which had not been directly tested before. Second, we tested predictions from the Production Anticipation Hypothesis (Forrin et al., 2019), an alternative explanation of the production effect based on social factors and participant-experimenter interactions within the experimental setting. Across four experiments, we investigated the production effect in immediate (Experiments 1A and 1B) and delayed (Experiments 2A and 2B) free recall. As in recent tests of the PAH, we made the occurrence of produced and silent items predictable and manipulated the experimenter's presence in the room. Finally, with produced and silent items presented in blocks within each list, we used a design that allowed us to repeatedly test both explanations.

Taken together, the results across experiments are very consistent. First, each experiment revealed a significant production effect along with the expected interaction between list condition and serial position. This aligns well with previous work showing the same interaction in free recall (Cyr et al., 2022; Gionet et al., 2022, 2024). In addition, the experiments were designed to create key contrasts for testing predictions derived from the two hypotheses. Some contrasts were specific to a given hypothesis, while others opposed both. As shown in Table 2, in general, the RFM was better supported than the PAH. A summary of the support received by each hypothesis

along with a detailed discussion of the implications of our results for other theories and models of the production effect are presented in the following sections.

The Production Anticipation Hypothesis

Adopting a novel and interesting perspective, the PAH was mainly developed to account for the drop in performance seen for silent items in going from pure silent lists to mixed lists that contain both silent and produced items (see, e.g., Bodner et al., 2014; Cyr et al., 2022; Forrin et al., 2019; MacLeod & Bodner, 2017). In theory, one could argue for a view where a mixed list boosts the relative (local) distinctiveness of all the items, due to the greater variability in the presentation modalities. Paradoxically, we know that such lists favor the produced items while depressing memory of the silently encoded ones (see Cyr et al., 2022; Forrin et al., 2016; Jonker et al., 2014; Lambert et al., 2016). According to the PAH, this cost for silent items is driven by social or higher-order mechanisms. The idea is that when the experimenter is in the room, there is some level of performance anxiety and therefore, knowing that the upcoming item is to be said aloud perturbs encoding of the current silent item.

This hypothesis innovates by acknowledging the role of social factors in experimental settings used to investigate memory (e.g., Klein, 2014; Klein & Marghetis, 2017). However, it should be noted that it cannot account for results from all studies on the production effect. In fact, in most studies, participants do not know in advance whether the next item will be produced. To apply to situations in which participants do not have foreknowledge of the status of the forthcoming item, it must be assumed that in mixed lists, anxiety was present for all silent items. However, even with this additional assumption, the PAH cannot account for the production effect observed when the experimenter was far away from the participant. For instance, using the crowdsourcing platforms Prolific and MTurk, researchers repeatedly observed the typical production effect, even when participants were not monitored during the task (e.g., Kelly et al.,

2024; Roberts et al., 2024). Under these conditions, it is hard to imagine that alone in the comfort of their home, participants would feel the social pressure of an experimenter that could be located thousands of kilometers away.

That being said, with the current design, the PAH leads to a set of predictions focused on the fate of silently read items that precede produced items. Across Experiments 1A (immediate free recall) and 2A (delayed free recall) during which the experimenter was present in the room, there were 12 critical contrasts involving silent items. In all cases, the silent block was presented first and recall of the last silent item was compared to recall of other silent items at the same serial position which were followed by another silent item instead of a produced item. Results supported the PAH for 6 contrasts, showing lower recall of the silent item that was the last of its block and followed by a produced item. However, we also found enough evidence to reject the PAH for 4 contrasts, along with inconclusive evidence for the remaining 2 contrasts.

Importantly, the PAH predicts that when the experimenter is absent, social interference and performance anticipation will no longer play any significant role. In other words, the 12 contrasts for which a difference was predicted in Experiments 1A and 2A (experimenter present) should have shown no difference in Experiments 1B and 2B (experimenter absent). However, we found enough evidence to support the PAH's predictions for only 3 contrasts and to reject the PAH for 7 contrasts, along with inconclusive evidence for 2 contrasts. Furthermore, we computed combined analyses between Experiment 1A and 1B, and between Experiment 2A and 2B. Results of both combined analyses showed evidence against an effect of the experimenter's presence and against all interactions involving the experimenter's presence. These results clearly show that, contrary to the PAH, manipulating the presence of the experimenter in the room had no effect on the magnitude of the production effect. The same pattern of results was also observed in a recent study by Wakeham-Lewis et al. (2022). Using an item recognition task and a more typical mixed-

list paradigm in which participants do not have foreknowledge of the status of the forthcoming word, they found that the experimenter's presence in the room during the task had no significant effect on the magnitude of the production effect. In addition, echoing the procedure that was used by Bond and Omar (1990), they measured participants' anxiety and self-consciousness but found no associations between these variables and the production effect. These results do not fit with the PAH's predictions, according to which variations in the cost for silent items and in the magnitude of the production effect should arise due to participants' increased performance anxiety in the presence of an experimenter (Forrin et al., 2019).

Finally, one could also argue that differences in the number of lists that were presented to participants during the task might explain the differences between the present results and those of Forrin et al. (2019). Indeed, whereas Forrin et al. only exposed their participants to a single list of words, Brenner (1973) presented 4 lists of items, and we presented either 50 or 62 lists depending on the experiment. However, based on results from past studies, it is unlikely that this difference can account for differences between the outcome patterns reported here and those of Forrin et al. For instance, in a short-term recall task with a large sample of 256 participants, Saint-Aubin et al. (2005) reported that the effects of semantic similarity on recall were the same across all 14 trials. Also, in a recent study with a sample of 550 participants, Guitard et al. (2025) showed the same pattern of errors (false recalls) with a single-trial protocol as previously reported with multiple-trial protocols. Therefore, these results suggest that the number of trials used in this study cannot, on its own, account for the lack of support observed for the PAH's predictions.

Free Recall and the PAH

It is worth noting that the predictions of the PAH and of the RFM could both have been borne out, at least to some degree, as they do not focus on the same elements of performance. That said, there were directly opposing predictions that could be identified and tested, and those

specific contrasts favored the RFM over the PAH. Overall, for the immediate and delayed free recall tasks used here, the predictions of the RFM were much better supported.

On the face of it, this makes the current study the odd one out. Prior work mentioned in the introduction used free recall tasks and reliably obtained a recall deficit consistent with the PAH's predictions: Silent items immediately preceding the one that each participant had to read aloud were relatively poorly remembered (Bond & Kirkpatrick, 1982; Bond & Omar, 1990; Brenner, 1973; Brown & Oxman, 1978). However, it is important to mention that in these studies, the set-up differed from standard free recall paradigms in important ways. As an example, consider the fact that the task typically involved a group of more than 20 participants simultaneously taking part in the study, in contrast with the more typical situation of a sole participant and an observer experimenter. In Brenner's (1973) study, one member of each of the 11 pairs read a word aloud, while the other pair members listened silently. At test, all participants attempted to remember all studied words. Results showed a form of Von Restorff effect for the one word read aloud (Hunt, 1995; von Restorff, 1933), in that whatever word that had been read aloud by a given participant was well recalled. Moreover, whereas each participant showed a recall deficit for the words immediately preceding and following the word that they read aloud, this deficit did not extend to the other participants who were listening (see also Bond & Kirkpatrick, 1982; Bond & Omar, 1990).

This design nicely illustrates the main differences between these classic studies and the typical free recall paradigm. First, only one word per list is read aloud by any given participant, as opposed to typical production effect paradigms where multiple items are read aloud by a sole participant. Second, any stumbling in production can potentially influence the performance of a sizeable group of peers, relative to the situation in a classic free recall task where mispronouncing a word will have no effect on an observing experimenter. Interestingly, Bond and Omar (1990)

reinforced the idea that performance anxiety might be involved in those "next-in-line" effects. To do this, they called upon a similar design and assessed participants' base level of social anxiety. They found that the effect was observed for participants with higher social anxiety scores, but that it disappeared for those with lower anxiety scores (see, Bond & Omar, 1990). To summarize, while there does seem to be a reliable performance anticipation / anxiety effect when production is used in the context of a social presentation paradigm, the results reported here suggest that this type of effect is not reliably observed in the situation where a single experimenter is present, who, notably, is not actively involved in or directly affected by task performance.

Implications for Other Theories and Models

Before discussing the implications of current results for the RFM, it is worth relating our findings to memory models accounting for the production effect. Interestingly, our findings suggest that the pattern of costs and benefits observed with the mixed-list production effect varies across serial positions as a function of list composition. These results provide a critical test for distinctiveness-based views of recall and recognition (Caplan & Guitard, 2024a; Jamieson et al., 2016; Kelly et al., 2022; Wakeham-Lewis et al., 2022). While the benefit for produced items has been accounted for by previous models and theories, explaining the cost for silent items has been more problematic. First, Jamieson et al. (2016; Spear et al., 2024) used MINERVA 2 to model the production effect by assuming that produced items would benefit from additional features generated through sensory feedback. Then, Kelly et al. (2022) and Wakeham-Lewis et al. (2022) relied on the Retrieving Effectively from Memory framework (REM; Shiffrin & Steyvers, 1997). Like the RFM, both of these models assume that producing items at encoding generates additional features that can boost their retrieval. However, they make no predictions on how features from adjacent produced and silent items might interact with each other in mixed lists, thus preventing them from explaining the patterns of costs and benefits reported here.

More recently, Caplan and Guitard (2024a) accounted for the production effect by using the Attentional Subsetting Theory (Caplan, 2023; Caplan & Guitard, 2024b). This theory distinguishes between shallow features (e.g., orthographic and phonological attributes) that emerge early in a densely packed space and deep features (e.g., semantic attributes) that emerge later in a sparser space. Specifically, they posit that production would boost the produced items' distinctiveness by (1) influencing the encoding of specific features (e.g., phonological features) and by (2) increasing the number of features that are attended to in memory tasks. Interestingly, contrary to past studies using MINERVA 2 or REM, Caplan and Guitard (2024a) also considered the possibility that producing items might divert attention from the encoding of orthographic or semantic features, thus potentially leading to a trade-off cost in memory. However, their hypothesis would predict a consistent cost across serial positions. Therefore, as for the other models described in this section, the Attentional Subsetting Theory in its current version is unable to explain how organizing produced and silent items in different configurations could yield different results.

Finally, although most studies on the production effect have focused on describing the benefit for produced items, two accounts attempted to explain the cost for silent items in mixed lists. First, Bodner et al. (2014) extended the lazy reading account of the generation effect (Begg & Snider, 1987), suggesting that participants would devote less effort or attention to processing items that they saw as relatively less important (i.e., silent items). Then, Jonker et al. (2014) called upon the item-order account (Forrin & MacLeod, 2016; McDaniel & Bugg, 2008; Nairne et al., 1991) suggesting that producing items in mixed lists would enhance the encoding of item information at the expense of order information for all items. This hypothesis would apply to free recall, because it has been shown that order information affects item recall in this context (e.g., Beaman & Jones, 1998). However, although these accounts can explain the extra cost on silent

items in going from pure to mixed lists, they would also predict the cost for silent items to be consistent across serial positions, which does not fit with our findings showing an interaction between list condition and serial position.

To summarize, whereas other current accounts of the production effect are able to explain the overall memory benefit for produced items and sometimes the cost for silent items, they all share one common shortcoming: They are unable to account for any serial position effects in the mixed list production effect. As our findings clearly demonstrate, the production effect in free recall is an item-level phenomenon. Produced and silent items are encoded with different amounts of item-specific features, but most critically, an item's features will interact with those of the adjacent list items in complex ways to determine subsequent recall performance. We now turn to the RFM to provide a single comprehensive account of the mixed list production effect.

The Production Effect and the RFM

The production effect is an interesting empirical phenomenon for several reasons. On one hand, it has been suggested that there are some applied uses to the knowledge developed around the effect – in the context of silently processed items, produced items have a clear mnemonic advantage in many circumstances, making production an interesting addition to the arsenal of mnemonic techniques available (for a review, Putnam, 2015). Moreover, contrary to many of the well-known mnemonic strategies (e.g., method of loci, pegword method, acrostics), the cognitive effort involved in production is reduced as it usually relies on very well-established responses.

More aligned with the preoccupations of this article, the effect sizes seen with alternating silent and produced items are substantial; the case can be made that they are even larger than classic levels-of-processing manipulations (e.g., Watkins et al., 2000; Lewandowsky & Farrell, 2008; Monsell et al., 1992). Conversely, the patterns observed with pure lists are more subtle and complex (Dauphinee et al., 2024; Fawcett et al., 2023; Gionet et al., 2022, 2024; Whitridge et al.,

2024). Taken together, these patterns provide a challenging opportunity to test a distinctiveness view such as that embodied in the RFM.

In a departure from the original Feature Model (see Nairne, 1988, 1990; Neath & Nairne, 1995; Neath & Surprenant, 2007), Saint-Aubin et al. (2021) developed the RFM by revising the overwriting process and adding a rehearsal mechanism. Since then, the RFM has been successful in accounting for patterns observed in a variety of paradigms (immediate serial recall, free recall, and both verbal and visuo-spatial reconstruction of order). Furthermore, Dauphinee et al. (2024) recently tested the RFM's rehearsal mechanism by investigating the effect of presentation speed on the pure list production effect in immediate serial recall. However, the current experiments are the first to directly and systematically test the predictions that can be derived from the retroactive interference mechanism described within the RFM.

Said predictions are mainly focused on the expectation that a produced item followed by a silent item will benefit from a sizeable local distinctiveness effect. To reiterate, the retroactive interference is thought to be similarity-based, as opposed to being a general type of interference – in essence, for this mechanism to exert its influence, feature n in item x has to be of the same nature as feature n in item x-1. It follows that the RFM has to predict that a feature-rich produced item will have a whole set of features that will remain intact if it is followed by a silent item. This will boost the produced item's distinctiveness, in the sense that as a retrieval cue, the produced item will support better discrimination among the competing candidates. Importantly, contrary to the PAH, the RFM also assumes that all memory processes in play are unaffected by the presence of the experimenter in the room. Therefore, the prediction is simple: A produced item followed by a silent item should be better recalled than a produced item followed by another produced item. In the present study, 23 of the 24 critical contrasts involving produced items across all experiments showed support for the RFM by showing better recall of the produced items that

were followed by silent items rather than by other produced items. The only contrast for which there was not enough evidence ($BF_{10} = 2.08$) was in the predicted direction and the lack of evidence is probably due to the sample size.

In addition to allowing specific tests of the PAH and the RFM, our experimental design also allowed 6 critical tests of the two hypotheses in Experiments 1A and 2A, during which the experimenter was present in the room. These contrasts involved silent items and directly opposed predictions from the two accounts. For each of the 6 contrasts, the PAH predicted lower recall of the silent items that were followed by a produced item rather than by another silent item. On the contrary, the RFM predicted no difference between conditions due to silent items benefiting from similar amounts of features and rehearsal regardless of what items (produced or silent) followed. Importantly, results provided support for the RFM on 4 of the 6 contrasts, for the PAH on one of the 6 contrasts in Experiment 1A, and for neither hypothesis (anecdotal evidence) on one of the 6 contrasts in Experiment 2A. Overall, although slight support was found for the PAH, the majority of the evidence supported the RFM. In light of the evidence supporting the RFM, the sole contrast showing evidence to the contrary can likely be attributed to random noise rather than significant deviations. Given the RFM's success in capturing the key predictions across conditions, this overall support for the model clearly outweighs any minor discrepancies.

Finally, fits of the RFM across all experiments nicely replicated the most important trends in the data, highlighting the recall advantage for the last produced item in conditions starting with a produced block as well as the lack of differences between recall for silent items in conditions starting with a silent block. Importantly, this is the first time that a model of the production effect was able to simultaneously account for the costs and benefits of the mixed-list production effect across the serial position curve. This sets the RFM apart from its competitors by showing that an explanation based on local distinctiveness, retroactive interference, and rehearsal can account for

the production effect regardless of list composition. Finally, these findings also highlight the robustness of this pattern when going from a short-term to a long-term memory task.

Conclusion

In this study, we relied on the production effect to conduct the first comprehensive test of the retroactive interference mechanism described within the RFM (Cyr et al., 2022; Saint-Aubin et al., 2021). Over all four experiments, we found extensive support for the RFM by showing that produced items in mixed lists are better remembered when followed by a silent item rather than by another produced item. This fits well with the similarity-based aspect of the RFM's retroactive interference mechanism, suggesting that silent items lack the features needed to interfere with the preceding produced item's features (see, Saint-Aubin et al., 2021). In addition, we found that this process operates similarly in both immediate and delayed free recall tasks. Finally, we were also able to test the RFM in contrast with a competing account of the memory cost for silent items in mixed lists: the Production Anticipation Hypothesis (Forrin et al., 2019). Even after making the occurrence of produced items predictable and manipulating the presence of the experimenter in the room, we still found limited support for the PAH. We suggest that the production effect in free recall is better explained by the joint effects of basic memory processes as described in the RFM than by social factors within the experimental setting.

References

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

Begg, I., & Snider, A. (1987). The generation effect: Evidence for generalized inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(4), 553–563. https://doi.org/10.1037/0278-7393.13.4.553

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*(5), 689-713. https://doi.org/10.3758/MC.37.5.689

Bodner, G. E., Taikh, A., & Fawcett, J. M. (2014). Assessing the costs and benefits of production in recognition. *Psychonomic Bulletin & Review, 21*(1), 149-154. https://doi.org/10.3758/s13423-013-0485-1

Bond, C. F., & Kirkpatrick, K. (1982). Distraction, amnesia, and the next-in-line effect. *Journal of Experimental Social Psychology, 18*(4), 307-323. https://doi.org/10.1016/0022-1031(82)90056-7

Bond, C. F., & Omar, A. S. (1990). Social anxiety, state dependence, and the next-in-line effect. *Journal of Experimental Social Psychology, 26*(3), 185–198. https://doi.org/10.1016/0022-1031(90)90034-J

Brenner, M. (1973). The next-in-line effect. *Journal of Verbal Learning and Verbal Behavior, 12*(3), 320-323. https://doi.org/10.1016/S0022-5371(73)80076-3

Brown, G. D., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review, 114(3),* 539-576. https://doi.org/10.1037/0033-295X.114.3.539

Brown, A. S., & Oxman, M. (1978). Learning through participation: Effects of involvement and anticipation of involvement. *The American Journal of Psychology*, *91*(3), 461–472. https://doi.org/10.2307/1421692

Camos, V., & Portrat, S. (2015). The impact of cognitive load on delayed recall. *Psychonomic Bulletin & Review, 22(4),* 1029–1034. https://doi.org/10.3758/s13423-014-0772-5

Caplan, J. B. (2023). Sparse attentional subsetting of item features and list-composition effects on recognition memory. *Journal of Mathematical Psychology*, *116*, 102802. https://doi.org/10.1016/j.jmp.2023.102802

Caplan, J. B., & Guitard, D. (2024a). A feature-space theory of the production effect in recognition. *Experimental Psychology, 71(1),* 64-82. https://doi.org/10.1027/1618-3169/a000611

Caplan, J. B., & Guitard, D. (2024b). Stimulus duration and recognition memory: An attentional subsetting account. *Journal of Memory and Language, 139,* 104556. https://doi.org/10.1016/j.jml.2024.104556

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

Cyr, V., Poirier, M., Yearsley, J. M., Guitard, D., Harrigan, I., & Saint-Aubin, J. (2022). The production effect over the long term: Modeling distinctiveness using serial positions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*(12), 1797–1820. https://doi.org/10.1037/xlm0001093

Dauphinee, I., Roy, M., Guitard, D., Yearsley, J. M., Poirier, M., & Saint-Aubin, J. (2024). Give me enough time to rehearse: Presentation rate modulates the production effect. *Psychonomic Bulletin & Review, 31,* 1603-1614. https://doi.org/10.3758/s13423-023-02437-5

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

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

Fawcett, J. M., Baldwin, M. M., Whitridge, J. W., Swab, M., Malayang, K., Hiscock, B., Drakes, D. H., & Willoughby, H. V. (2023). Production improves recognition and reduces intrusions in between-subject designs: An updated meta-analysis. *Canadian Journal of Experimental Psychology, 77*(1), 35–44. https://doi.org/10.1037/cep0000302

Forrin, N. D., & MacLeod, C. M. (2016). Order information is used to guide recall of long lists: Further evidence for the item-order account. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale, 70*(2), 125–138. https://doi.org/10.1037/cep0000088

Forrin, N. D., & MacLeod, C. M. (2018). This time it's personal: The memory benefit of hearing oneself. *Memory, 26*(4), 574-579. https://doi.org/10.1080/09658211.2017.1383434

Forrin, N. D., Groot, B., & MacLeod, C. M. (2016). The d-Prime directive: Assessing costs and benefits in recognition by dissociating mixed-list false alarm rates. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(7), 1090-1111. http://dx.doi.org/10.1037/xlm0000214

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*, 104050. https://doi.org/10.1016/j.jml.2019.104050

Gionet, S., Guitard, D., Poirier, M., Yearsley, J. M., & Saint-Aubin, J. (2025). *Distinctiveness and interference in free recall: A test with the production effect* [Data, Materials]. https://osf.io/3xjuf/?view_only=bbc5d58454e3421ea1628668d5114380

Gionet, S., Guitard, D., & Saint-Aubin, J. (2024). The interaction between the production effect and serial position in recognition and recall. *Experimental Psychology, 71(5),* 259-277. https://doi.org/10.1027/1618-3169/a000623

Gionet, S., Guitard, D., & Saint-Aubin, J. (2022). The production effect interacts with serial positions: Further evidence from a between-subjects manipulation. *Experimental Psychology*, *69*(1), 12–22. https://doi.org/10.1027/1618-3169/a000540

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

Gross-Sampson, M. (2020). *New Bayesian Guide*. https://doi.org/10.17605/OSF.IO/CKNXM

Guitard, D., Saint-Aubin, J., Reid, J. N., & Jamieson, R. K. (2025). An embedded computational framework of memory: Accounting for the influence of semantic information in verbal short-term memory. *Journal of Memory and Language, 140,* 104573. https://doi.org/10.1016/j.jml.2024.104573

Huff, M. J., Bodner, G. E., & Gretz, M. R. (2021). Distinctive encoding of a subset of DRM lists yields not only benefits, but also costs and spillovers. *Psychological Research, 85,* 280-290. https://doi.org/10.1007/s00426-019-01241-y

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

Hunt, R. R. (1995). The subtlety of distinctiveness: What von Restorff really did. *Psychonomic Bulletin & Review, 2*(1), 105–112. https://doi.org/10.3758/BF03214414

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

Jones, A. C., & Pyc, M. A. (2014). The production effect: Costs and benefits in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(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(2), 441–448. https://doi.org/10.1037/a0034977

Kelly, M. O., Ensor, T. M., Lu, X., MacLeod, C. M., & Risko, E. F. (2022). Reducing retrieval time modulates the production effect: Empirical evidence and computational accounts. *Journal of Memory and Language, 123*, 104299. https://doi.org/10.1016/j.jml.2021.104299

Kelly, M. O., Ensor, T. M., MacLeod, C. M., & Risko, E. F. (2024). The prod eff: Partially producing items moderates the production effect. *Psychonomic Bulletin & Review, 31*, 373-379. https://doi.org/10.3758/s13423-023-02360-9

Klein, S.A. (2014). Don't blink : Performance experimental time in the brain laboratory. *Performance Research, 19*(3), 88-92. https://doi.org/10.1080/13528165.2014.935181

Klein, S. A., & Marghetis, T. (2017). Shaping experiment from the inside out: Performance-collaboration in the cognitive science lab. *Performance Matters, 3*(2), 16-40.

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 / Revue canadienne de psychologie expérimentale, 70*(2), 165-176. https://doi.org/10.1037/cep0000086

Lawrence, M. A. (2016). *ez: Easy analysis and visualization of factorial experiments*. R package version 4.4-0. https://CRAN.R-project.org/package=ez

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

MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science, 26*(4), 390-395. https://doi.org/10.1177/0963721417691356

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

Marin, J.-M., Pudlo, P., Robert, C. P., & Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing, 22*(6), 1167-1180. https://doi.org/10.1007/s11222-011-9288-2

McCabe, D. P. (2008). The role of covert retrieval in working memory span tasks: Evidence from delayed recall tests. *Journal of Memory and Language, 58*(2), 480-494. https://doi.org/10.1016/j.jml.2007.04.004

McDaniel, M. A., & Bugg, J. M. (2008). Instability in memory phenomena: A common puzzle and a unifying explanation. *Psychonomic Bulletin & Review, 15*(1), 237-255. https://doi.org/10.3758/PBR.15.2.237

Monsell, S., Patterson, K. E., Graham, A., Hughes, C. H., & Milroy, R. (1992). Lexical and sublexical translation of spelling to sound: Strategic anticipation of lexical status. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(3), 452–467. https://doi.org/10.1037/0278-7393.18.3.452

Morey, R. D. & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for common designs.* R package version 0.9.12-4.2. https://CRAN.R-project.org/package=BayesFactor

Murray, D. J. (1967). The role of speech responses in short-term memory. *Canadian Journal of Psychology*, *21*(3), 263–276. https://doi.org/10.1037/h0082978

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

Nairne, J. S. (1988). The mnemonic value of perceptual identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(2), 248-255. https://doi.org/10.1037/0278-7393.14.2.248

Nairne, J. S., Riegler, G. L., & Serra, M. (1991). Dissociative effects of generation on item and order retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(4), 702–709. https://doi.org/10.1037/0278-7393.17.4.702

Navarro, D. J. (2015). *Learning statistics with R: A tutorial for psychology students and other beginners (Version 0.5).* University of Adelaide, Adelaide, Australia.

Neath, I. (1993). Contextual and distinctive processes and the serial position function. *Journal of Memory and Language*, *32*(6), 820-840. https://doi.org/10.1006/jmla.1993.1041

Neath, I., Brown, G. D. A., McCormack, T., Chater, N., & Freeman, R. (2006). Distinctiveness models of memory and absolute identification: Evidence for local, not global, effects. *Quarterly Journal of Experimental Psychology, 59*(1), 121-135. https://doi.org/10.1080/17470210500162086

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

Neath, I., & 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 University Press.

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

Ozubko, J. D., & MacLeod, C. M. (2010). The production effect in memory: Evidence that distinctiveness underlies the benefit. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(6), 1543–1547. https://doi.org/10.1037/a0020604

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

Prisko, L. H. (1963). *Short-term memory in focal cerebral damage*. McGill University, Canada. http://digitool.library.mcgill.ca/webclient/DeliveryManager?pid=115219&custom_att_2=direct

Putnam, A. L. (2015). Mnemonics in education: Current research and applications. *Translational Issues in Psychological Science, 1*(2), 130–139. https://doi.org/10.1037/tps0000023

R Core Team (2023). R*: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

Roberts, B. R. T., Hu, Z. S., Curtis, E., Bodner, G. E., McLean, D., & MacLeod, C. M. (2024). Reading text aloud benefits memory but not comprehension. *Memory & Cognition, 52*, 57-72. https://doi.org/10.3758/s13421-023-01442-2

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*(1), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Routh, D. A. (1970). 'Trace strength' modality, and the serial position curve in immediate memory. *Psychonomic Science, 18*(1), 355-357. https://doi.org/10.3758/BF03332397

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

Saint-Aubin, J., Ouellette, D., & Poirier, M. (2005). Semantic similarity and immediate serial recall: Is there an effect on all trials? *Psychonomic Bulletin & Review, 12*(1), 171–177. https://doi.org/10.3758/BF03196364

Saint-Aubin, J., Poirier, M., Yearsley, J. M., & Guitard, D. (2024). The production effect becomes spatial. *Experimental Psychology, 71*(1). https://doi.org/10.1027/1618-3169/a000609

Saint-Aubin, J., Poirier, M., Yearsley, J. M., Robichaud, J.-M., & Guitard, D. (2023). Modeling verbal short-term memory: A walk around the neighborhood. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 49*(2), 198–215. https://doi.org/10.1037/xlm0001226

Saint-Aubin, J., Yearsley, J.M., Poirier, M., Cyr, V., Guitard, D. (2021). A model of the production effect over the short-term: The cost of relative distinctiveness. *Journal of Memory and Language, 118,* 104219. https://doi.org/10.1016/j.jml.2021.104219

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review, 25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22*(2), 322–339. https://doi.org/10.1037/met0000061

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4280), 1317-1323. https://doi.org/10.1126/science.3629243

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review, 4(2),* 145–166. https://doi.org/10.3758/BF03209391

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

Sisson, S. A., Fan, Y., & Tanaka, M. M. (2009). Correction for Sisson et al., Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America, 106*(39), Article 16889. https://doi.org/10.1073/pnas.0908847106

Spear, J., Reid, N., Guitard, D., & Jamieson, R. K. (2024). Directed Forgetting and the Production Effect: Assessing Strength and Distinctiveness. *Experimental Psychology, 71(5),* 278–297. https://doi.org/10.1027/1618-3169/a000630

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

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

Surprenant, A. M., & Neath, I. (2009). *Principles of memory (1st ed.)*. Psychology Press. https://doi.org/10.4324/9780203848760

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

Turner, B. M., & Van Zandt, T. (2012). A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology, 56*(2), 69-85. https://doi.org/10.1016/j.jmp.2012.02.005

von Restorff, H. (1933). Über die Wirkung von Bereichsbildungen im Spurenfeld. *Psychologische Forschung, 18*, 299–342. https://doi.org/10.1007/BF02409636

Wakeham-Lewis, R. M., Ozubko, J., & Fawcett, J. M. (2022). Characterizing production: The production effect is eliminated for unusual voices unless they are frequent at study. *Memory, 30(10)*, 1319-1333, https://doi.org/10.1080/09658211.2022.2115075

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

Watkins, M. J., LeCompte, D. C., & Kim, K. (2000). Role of study strategy in recall of mixed lists of common and rare words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(1), 239–245. https://doi.org/10.1037/0278-7393.26.1.239

Watkins, M. J., Neath, I., & Sechler, E. S. (1989). Recency effect in recall of a word list when an immediate memory task is performed after each word presentation. *The American Journal of Psychology, 102*(2), 265-270. https://doi.org/10.2307/1422957

Whitridge, J. W., Huff, M. J., Ozubko, J. D., Bürkner, P. C., Lahey, C. D., & Fawcett, J. M. (2024). Singing does not necessarily improve memory more than reading aloud: An empirical and meta-analytic investigation. *Experimental Psychology, 71*(1), 33-50. https://doi.org/10.1027/1618-3169/a000614

Table 1

List Composition Conditions Used in All Experiments

	Input Serial Position									
List Condition	1	2	3	**4**	**5**	**6**	7	8	9	10
Condition A	P	P	P	**P**	**S**	**S**	S	S	S	S
Condition B	P	P	P	**P**	**P**	**S**	S	S	S	S
Condition C	P	P	P	**P**	**P**	**P**	S	S	S	S
Condition D	S	S	S	**S**	**P**	**P**	P	P	P	P
Condition E	S	S	S	**S**	**S**	**P**	P	P	P	P
Condition F	S	S	S	**S**	**S**	**S**	P	P	P	P

Note. Red (P) = Produced, Blue (S) = Read silently; framed positions / conditions are the critical items for which the RFM and the PAH make specific predictions – please see Table 2

Table 2

Summary of Results for Predictions Derived from RFM and PAH as a Function of Production Condition and Serial Position

		Serial Position								
		4			**5**			**6**		
Condition		Predicted	Exp1	Exp2	Predicted	Exp1	Exp2	Predicted	Exp1	Exp2
Experimenter PRESENT (Experiments 1A & 2A)										
Produced		**A > C**	YES	YES	**B > D**	YES	YES	**C > D**	YES	YES
		A > B	YES	YES	**B > C**	YES	YES	**C > E**	YES	YES
Silent	PAH	**D < E**	NO	NO	**E < F**	NO	NO	**F < B**	YES	YES
	RFM	**D = E**	YES	YES	**E = F**	YES	YES			
	PAH	**D < F**	YES	???	**E < A**	YES	YES	**F < A**	???	YES
	RFM	**D = F**	NO	???						
Experimenter ABSENT (Experiments 1B & 2B)										
Produced		**A > C**	YES	YES	**B > D**	YES	YES	**C > D**	YES	YES
		A > B	YES	YES	**B > C**	YES	???	**C > E**	YES	YES
Silent	PAH	**D = E**	NO	YES	**E = F**	YES	YES	**F = B**	NO	NO
	RFM	**D = E**	NO	YES	**E = F**	YES	YES			
	PAH	**D = F**	NO	???	**E = A**	NO	???	**F = A**	NO	NO
	RFM	**D = F**	NO	???						

Note. YES = $BF_{10} > 3$, NO = $BF_{01} > 3$, ??? = $0.33 < BF_{10} < 3$. Highlighted cells represent critical contrasts testing opposing predictions from the two accounts. Empty cells represent contrasts for which the RFM does not make any predictions.

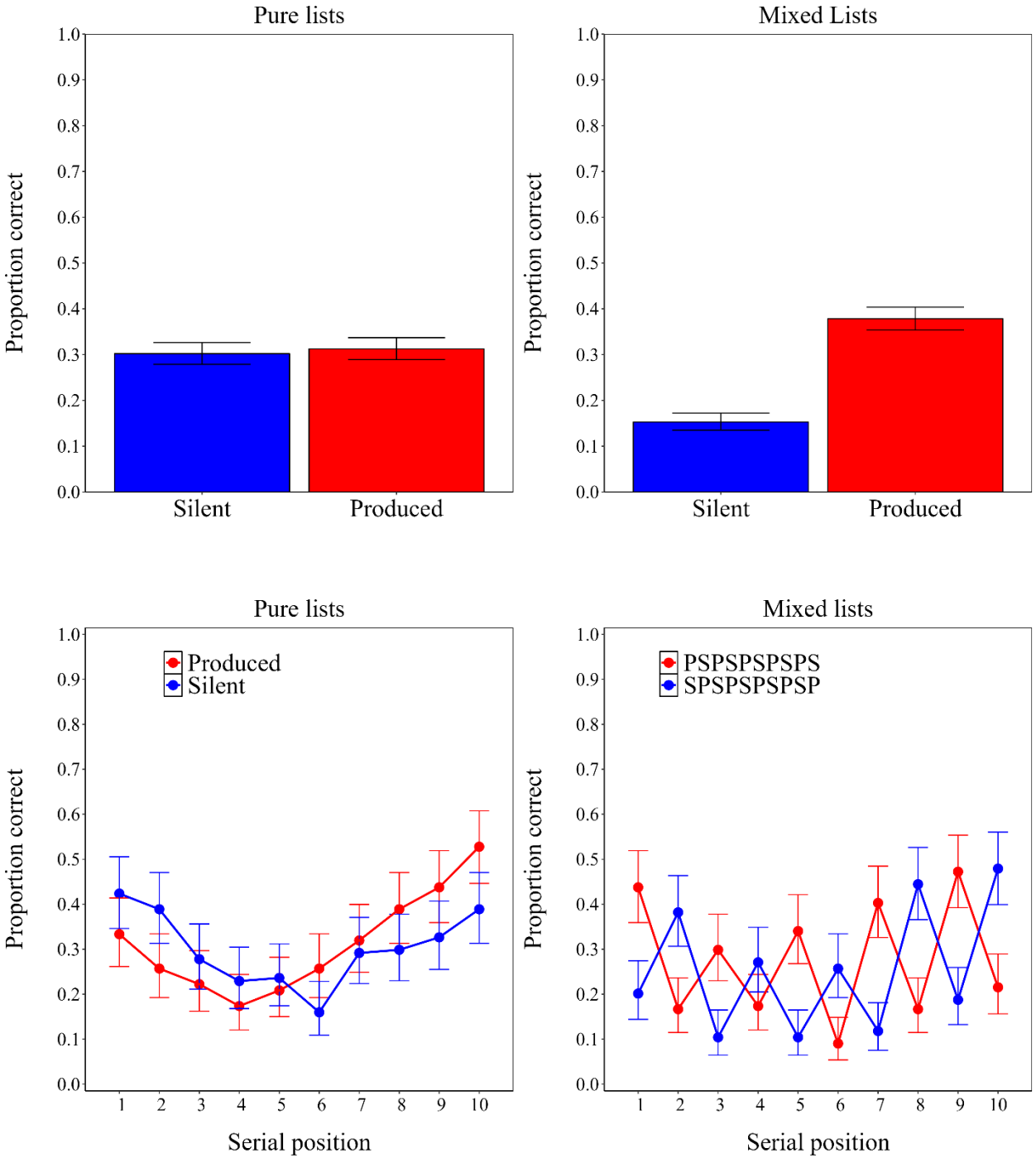
Table 3

Values of fixed parameters, means and 95% HDIs for estimated parameters in each experiment.

Fixed Parameters					
Number of particles		1000			
Number of Modality Independent Features		20			
Number of Modality Dependent Features		2 Visual + 18 Auditory			
Estes parameter $\boldsymbol{\theta}$		0.05			
Estimated Parameters					
Parameter	**Prior**	**EX1A [HDI]**	**EX1B [HDI]**	**EX2A [HDI]**	**EX2B [HDI]**
$\boldsymbol{a}$	Normal (3,1)	5.477 [4.763, 6.308]	5.282 [4.527, 6.360]	5.176 [4.232, 6.142]	4.790 [3.757, 5.811]
$\boldsymbol{\lambda}$ **Overwriting Parameter**	Normal (0.4,0.1)	0.153 [0.00, 0.252]	0.181 [0.058, 0.292]	0.106 [0.000, 0.209]	0.113 [0.000, 0.214]
$\boldsymbol{r_A}$ **Rehearsal parameter for Produced items**	Beta (2,8)	0.295 [0.171, 0.446]	0.332 [0.185, 0.514]	0.158 [0.006, 0.461]	0.178 [0.005, 0.495]
$\boldsymbol{r_S}$ **Rehearsal parameter for Silently read items**	Beta (2,8)	0.317 [0.157, 0.484]	0.275 [0.102, 0.465]	0.240 [0.012, 0.540]	0.219 [0.006, 0.475]
$\boldsymbol{\tau}$ **Temperature parameter**	HalfNormal (0,.1)	0.061 [0.037, 0.085]	0.070 [0.041, 0.104]	0.053 [0.027, 0.083]	0.084 [0.041, 0.140]
$\boldsymbol{floor}$ **Minimum activation of item needed to generate recall**	HalfNormal (0,.1)	0.016 [0.009, 0.024]	0.018 [0.010, 0.028]	0.014 [0.007, 0.025]	0.019 [0.009, 0.038]

Figure 1

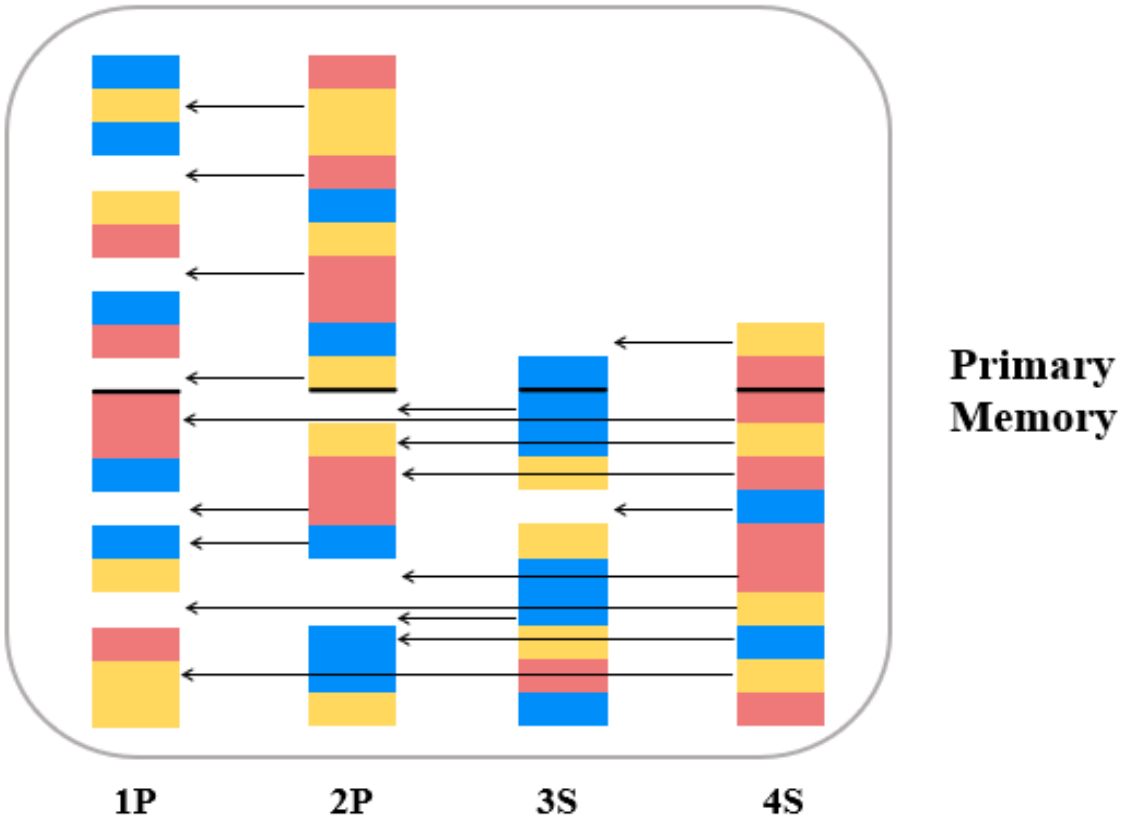
Free recall as a function of production and serial position (Experiment 3 of Cyr et al., 2022)



Note. Redrawn from Experiment 3 of Cyr et al. (2022). Error bars represent 95% Bayesian highest density intervals (HDI) based on the proportion of correct responses.

Figure 2

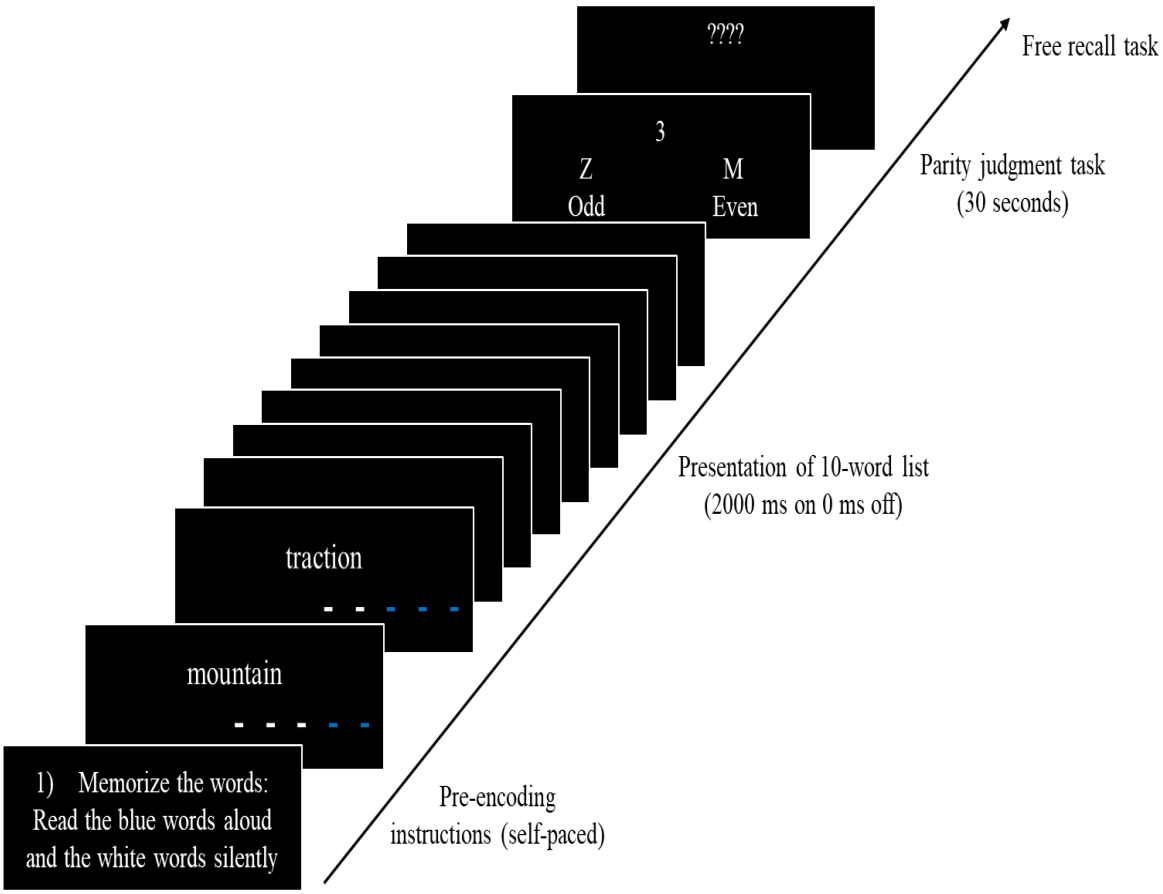
Schematic illustration of the Revised Feature Model's retroactive interference mechanism



Note. This example illustrates the transition between a block of 2 produced items and a block of 2 silent items within a mixed list. Each column represents an item vector, while colored rectangles represent distinct features (yellow, red, and blue rectangles represent values of 1, 2, and 3). For illustrative purposes, the bottom ten rectangles represent modality-independent features, and the top ten rectangles represent modality-dependent features. The left-pointing arrows illustrate the retroactive interference mechanism described within the RFM. When the same feature occupies the same position for two items, the previous item's feature can be overwritten by the identical feature of the subsequent item (shown by the white rectangles). The probability of overwriting is also inversely proportional to the distance between the items. Finally, items with the most intact features will have the greatest probability of being recalled at test.

Figure 3

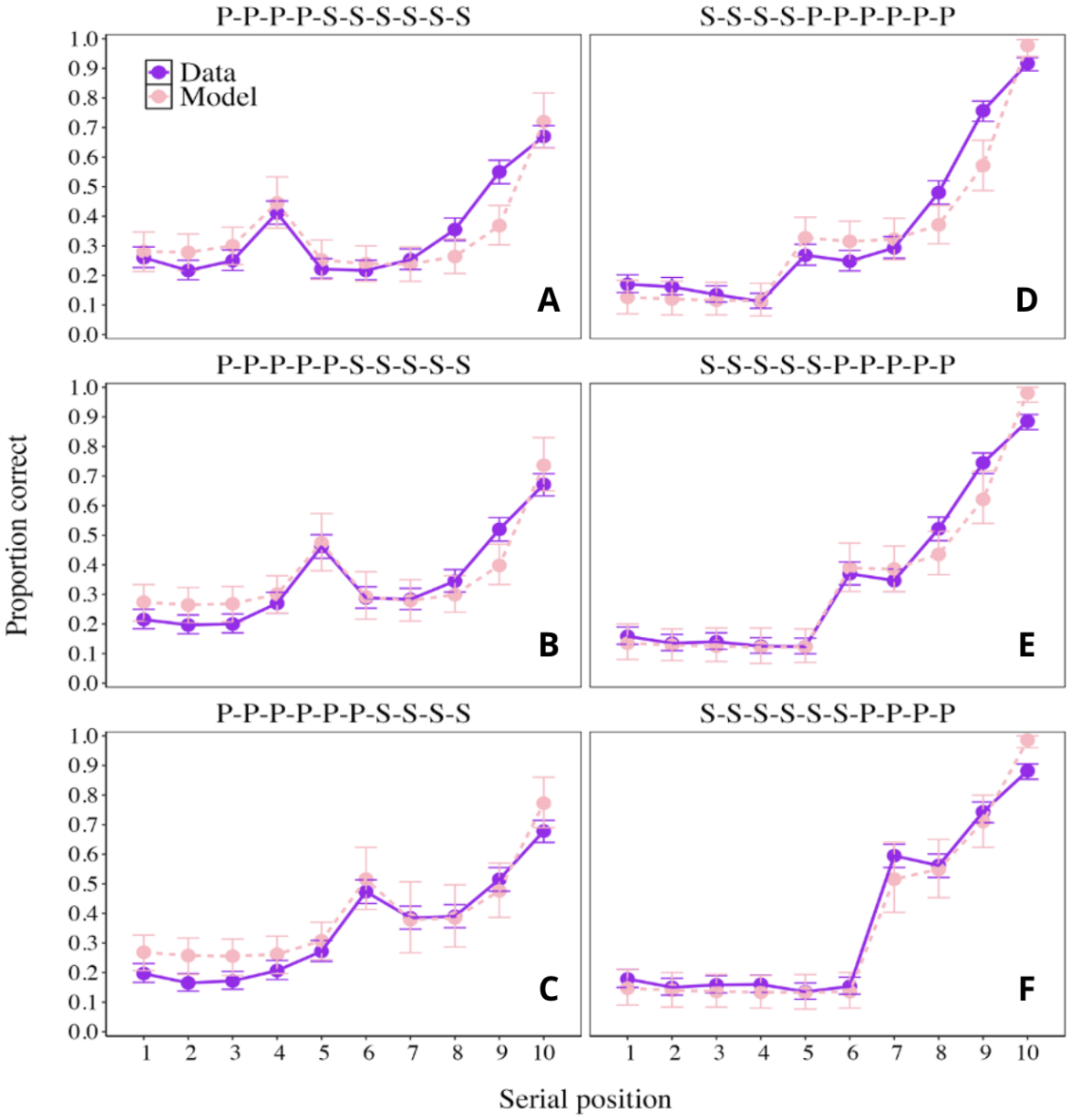
Illustration of the General Procedure Followed in All Experiments



Note. The parity judgment task was presented only in Experiments 2A and 2B.

Figure 4

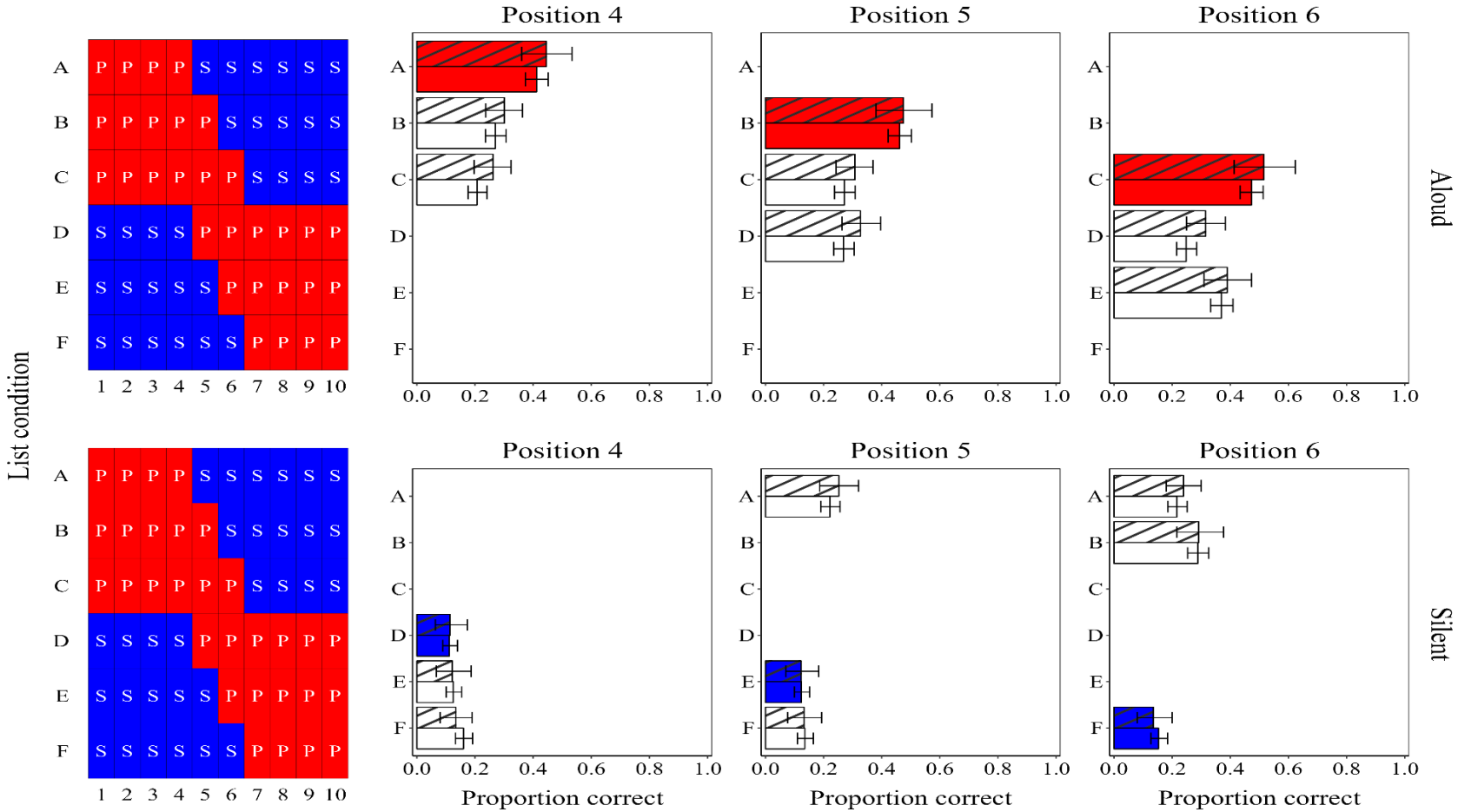
Experiment 1A – Experimenter Present: Mean Free Recall Performance as a Function of List Condition and Serial Position



Note. Error bars represent 95% Bayesian highest density intervals (HDI) based on the proportion of correct responses. Error bars were computed separately for the model and the data. Labels inside each panel represent the list conditions presented in Table 1.

Figure 5

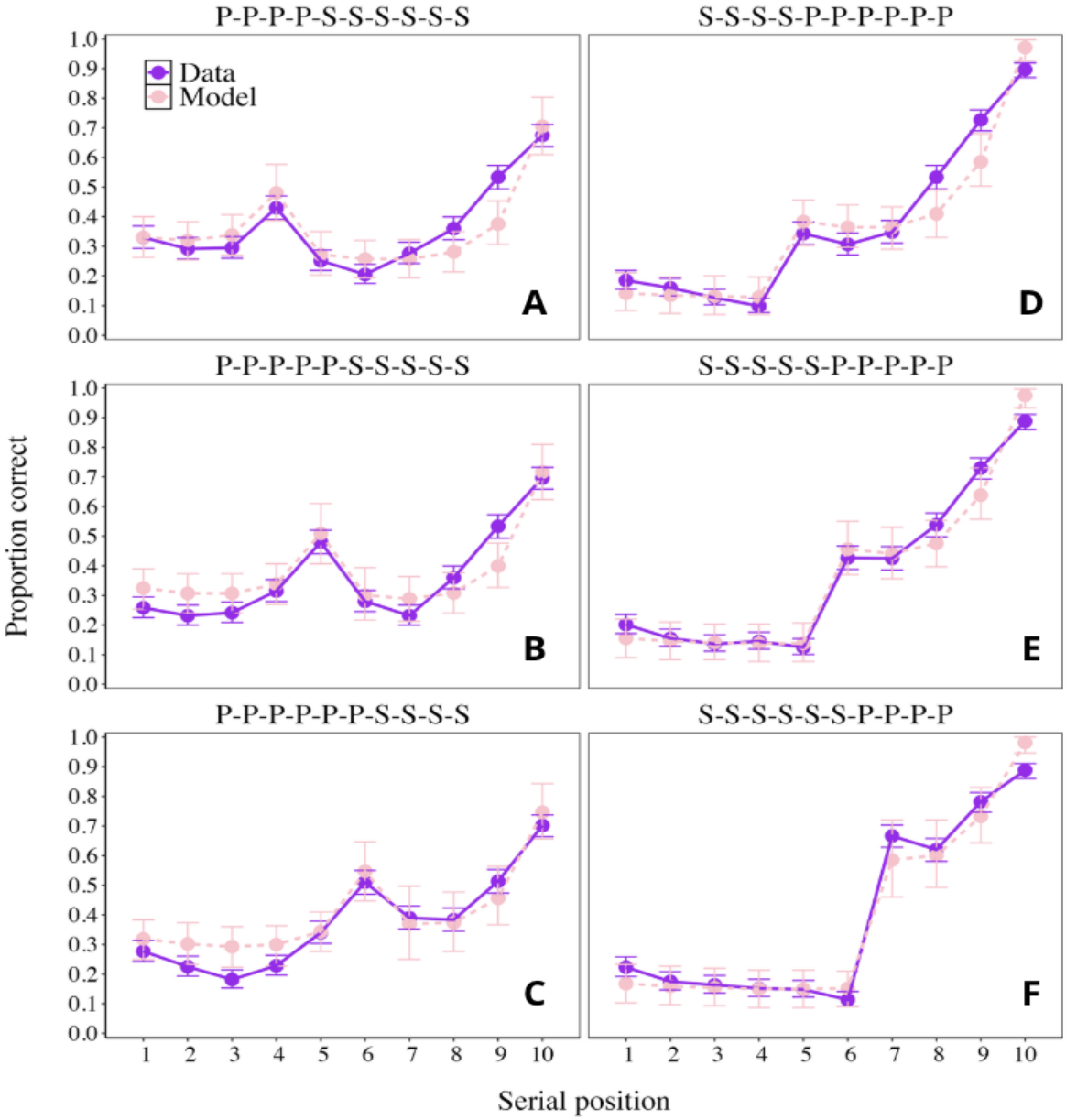
Experiment 1A – Experimenter Present: Mean Free Recall Performance at Positions 4, 5 and 6 as a Function of List Condition.



Note. Empty bars = data, Striped bars = Model predictions, Colored bars = Last item of its block. Error bars represent 95% Bayesian highest density intervals based on the proportion of correct responses. Error bars were computed separately for the model and the data.

Figure 6

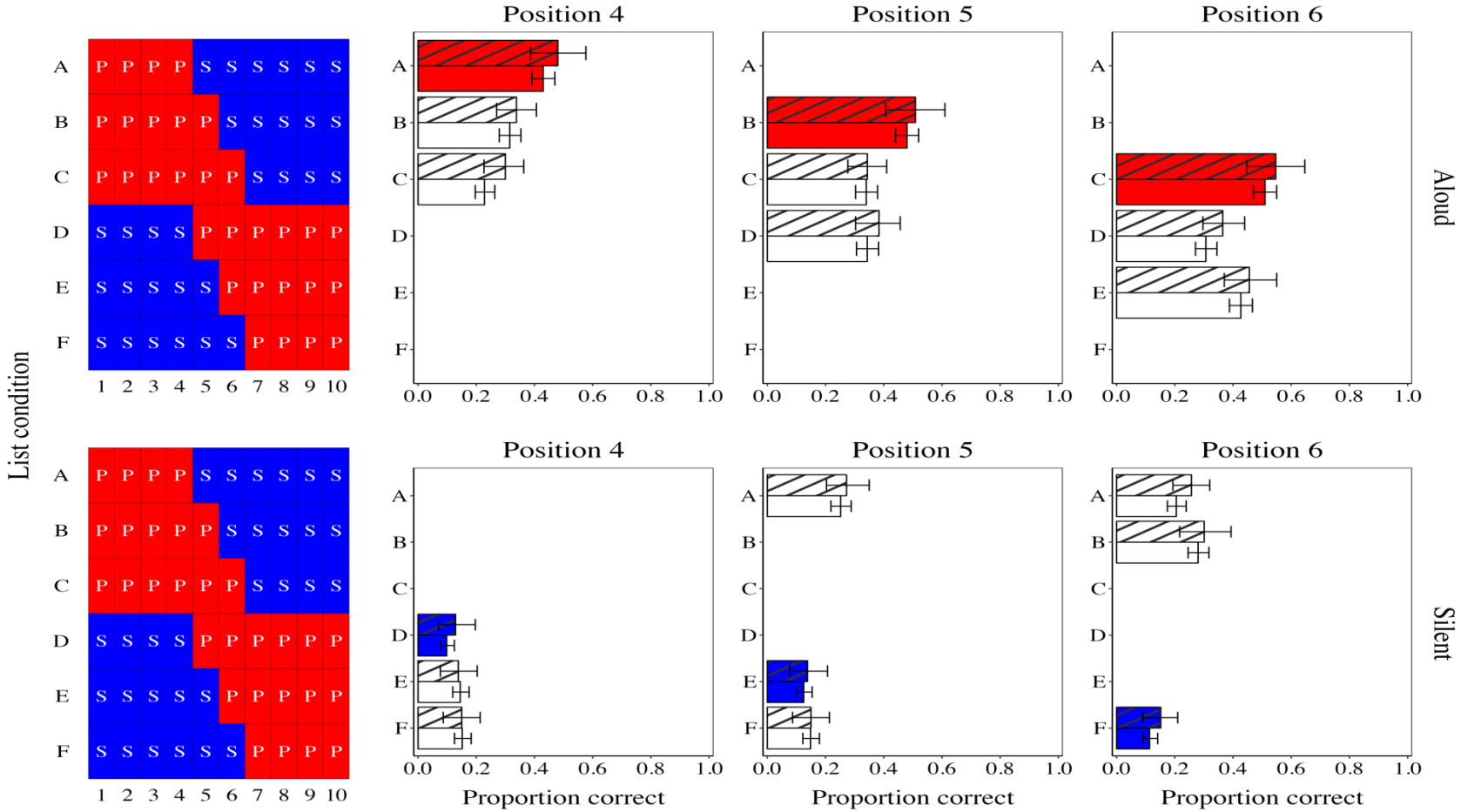
Experiment 1B – Experimenter Absent: Mean Free Recall Performance as a Function of List Condition and Serial Position



Note. Error bars represent 95% Bayesian highest density intervals (HDI) based on the proportion of correct responses. Error bars were computed separately for the model and the data. Labels inside each panel represent the list conditions presented in Table 1.

Figure 7

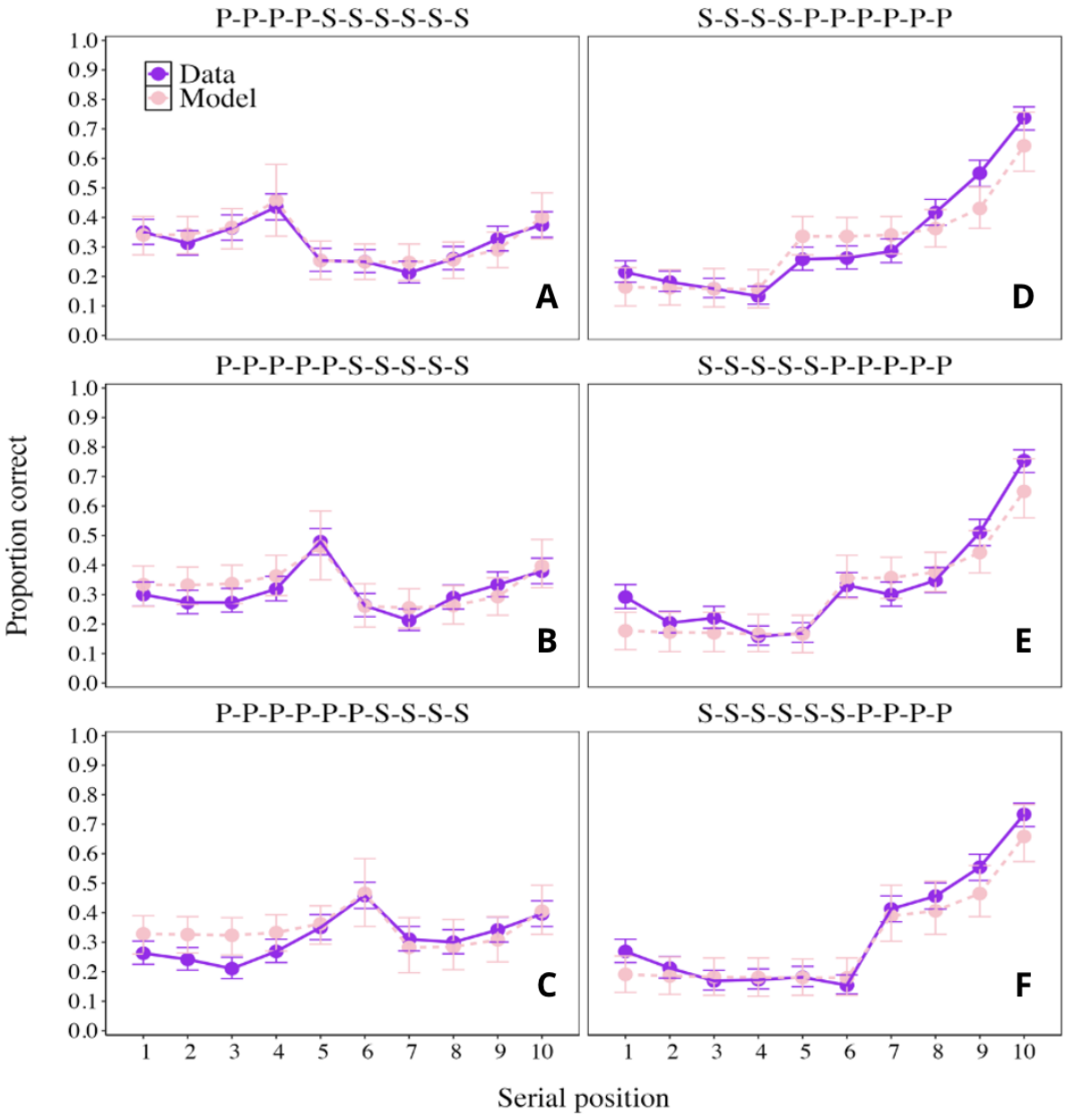
Experiment 1B – Experimenter Absent: Mean Free Recall Performance at Positions 4, 5 and 6 as a Function of List Condition



Note. Empty bars = data, Striped bars = Model predictions, Colored bars = Last item of its block. Error bars represent 95% Bayesian highest density intervals based on the proportion of correct responses. Error bars were computed separately for the model and the data.

Figure 8

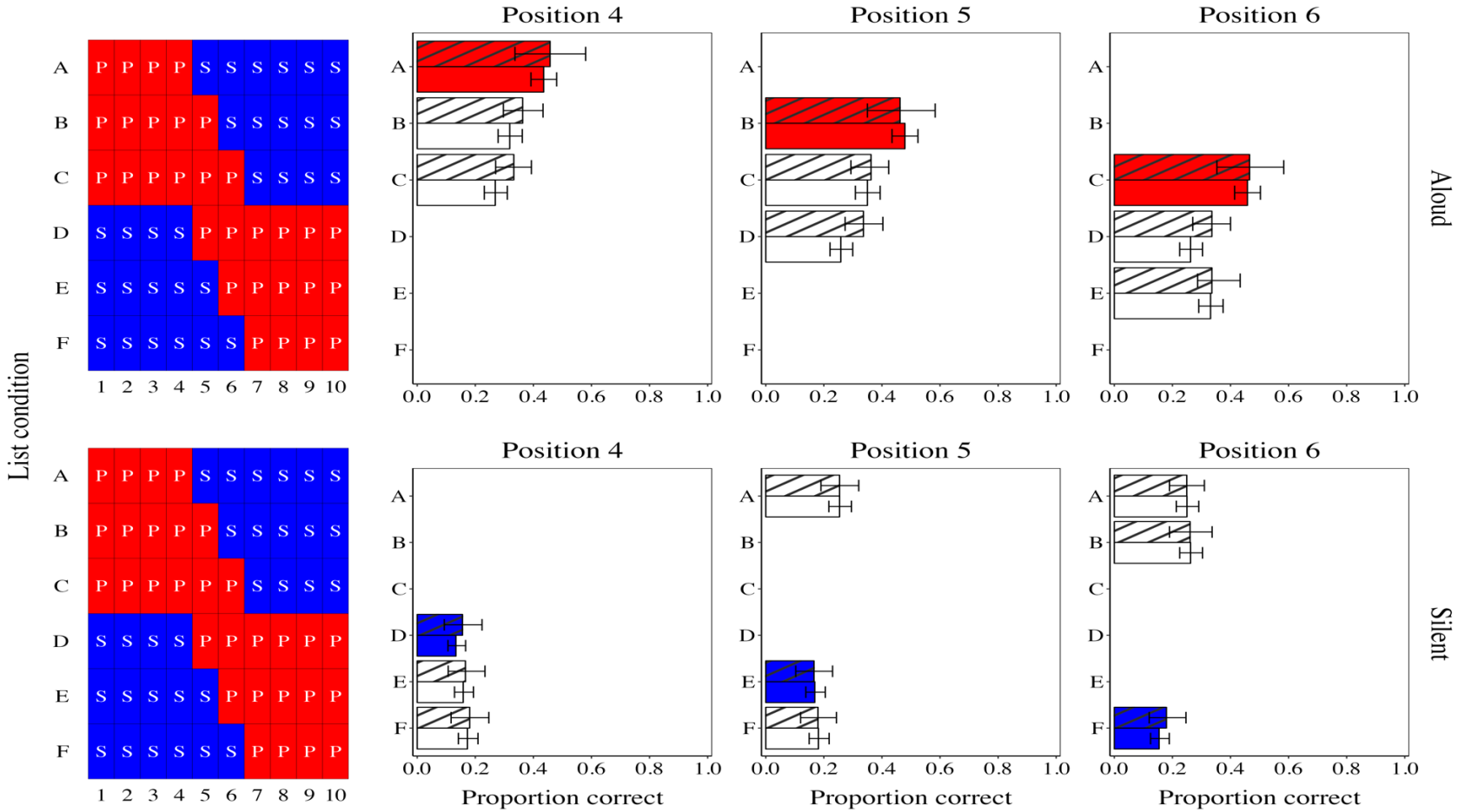
Experiment 2A – Experimenter Present: Mean Free Recall Performance as a Function of List Condition and Serial Position



Note. Error bars represent 95% Bayesian highest density intervals (HDI) based on the proportion of correct responses. Error bars were computed separately for the model and the data. Labels inside each panel represent the list conditions presented in Table 1.

Figure 9

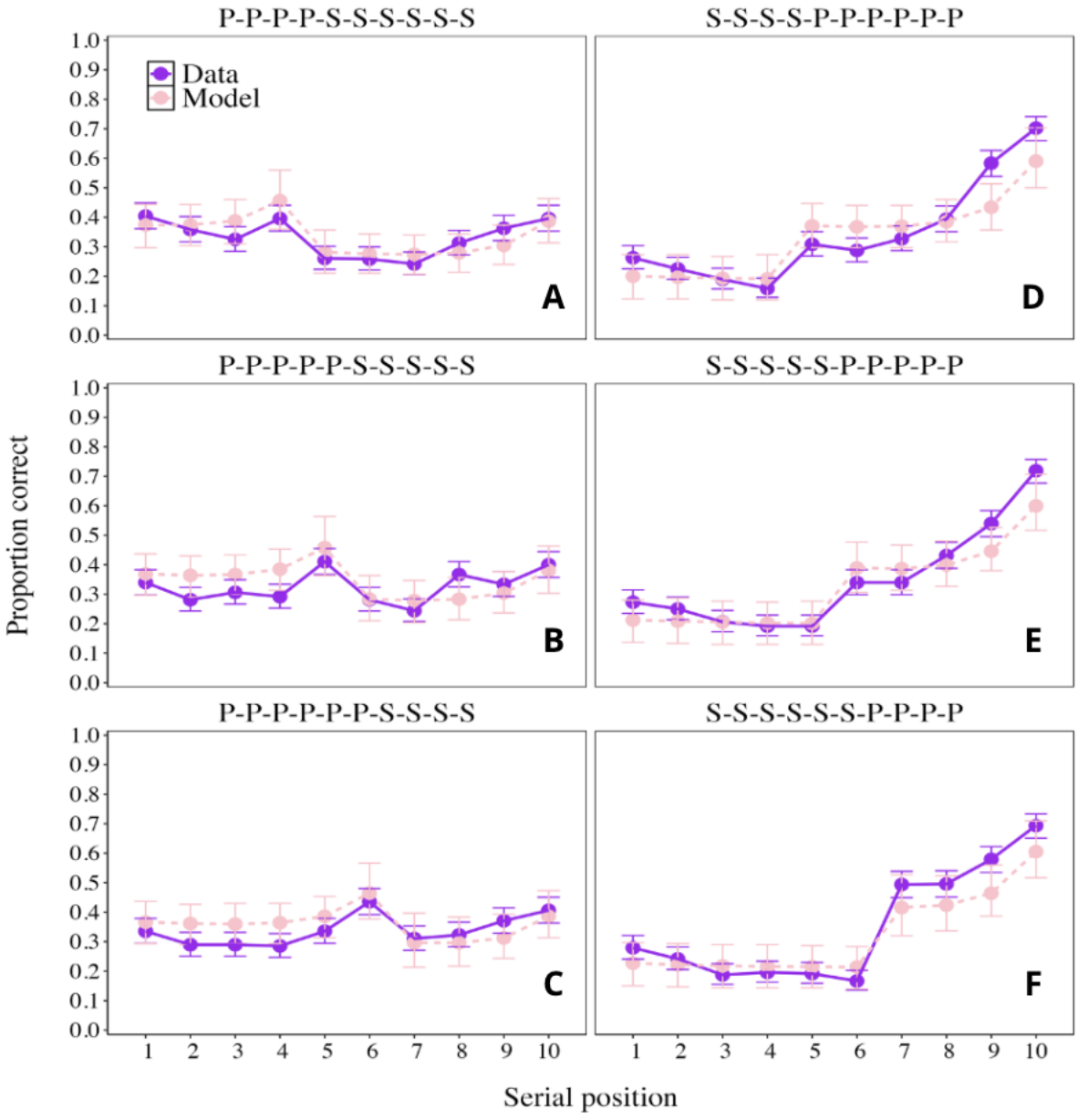
Experiment 2A – Experimenter Present: Mean Free Recall Performance at Positions 4, 5 and 6 as a Function of List Condition



Note. Empty bars = data, Striped bars = Model predictions, Colored bars = Last item of its block. Error bars represent 95% Bayesian highest density intervals based on the proportion of correct responses. Error bars were computed separately for the model and the data.

Figure 10

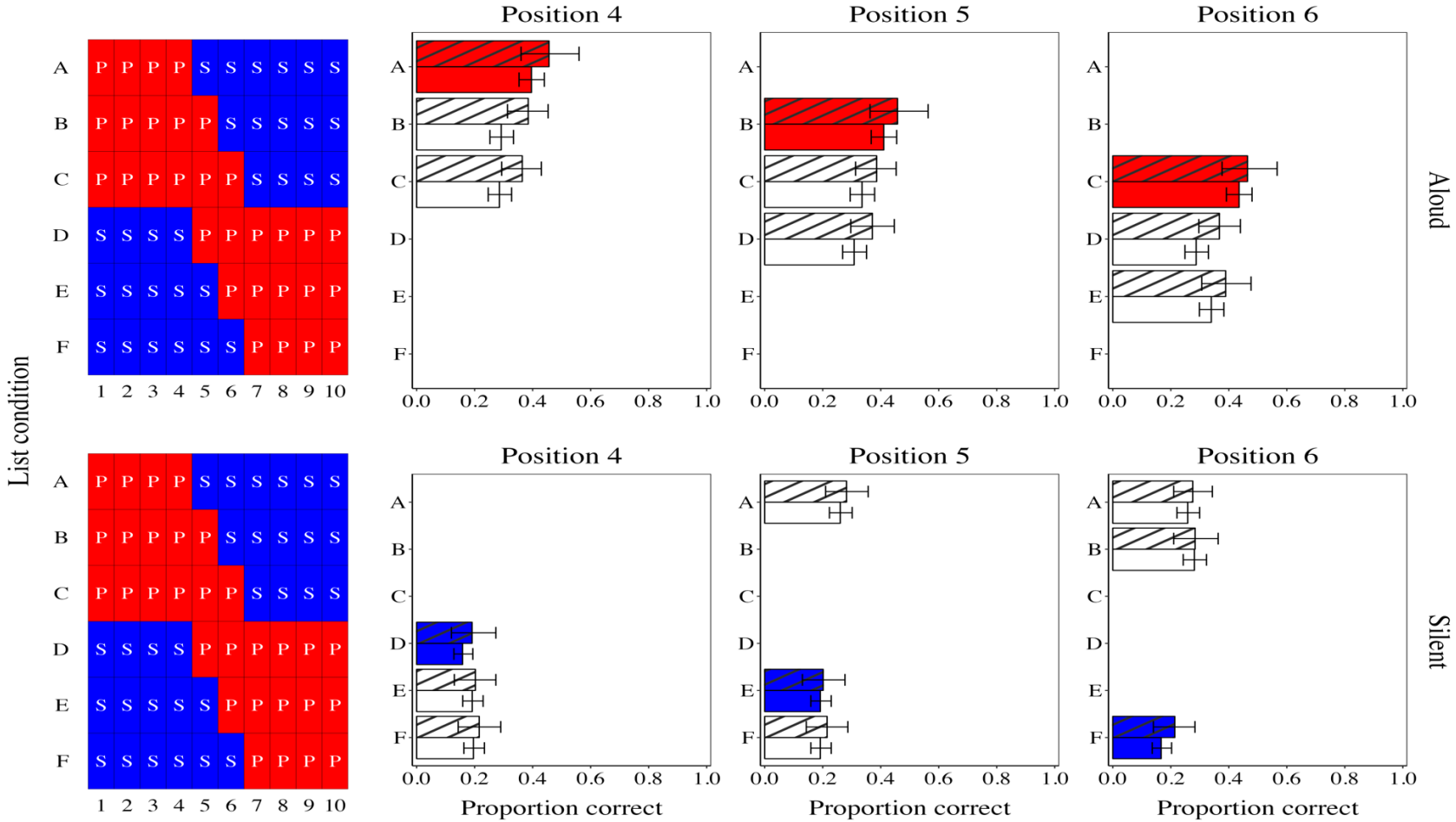
Experiment 2B – Experimenter Absent: Mean Free Recall Performance as a Function of List Condition and Serial Position



Note. Error bars represent 95% Bayesian highest density intervals (HDI) based on the proportion of correct responses. Error bars were computed separately for the model and the data. Labels inside each panel represent the list conditions presented in Table 1.

Figure 11

Experiment 2B – Experimenter Absent: Mean Free Recall Performance at Positions 4, 5 and 6 as a Function of List Condition



Note. Empty bars = data, Striped bars = Model predictions, Colored bars = Last item of its block. Error bars represent 95% Bayesian highest density intervals based on the proportion of correct responses. Error bars were computed separately for the model and the data.

Appendix A – 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 (Turner & Van Zandt, 2012; Marin et al., 2012). ABC methods allow for Bayesian model fitting even in cases where the likelihood cannot be computed, by using simulated data to obtain an approximate likelihood. Specifically, we used a procedure known as ABC Partial Rejection Control (ABC-PRC; Sisson et al., 2007, 2009) which we have previously used to fit the original Feature Model (Poirier et al., 2019) and the Revised Feature Model (Saint-Aubin et al., 2021; Cyr et al., 2022).

ABC-PRC works by repeatedly sampling from a prior over the parameter space until it finds a set of parameters that generate summary statistics 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 are associated with parameter sets, the algorithm gives each particle a weight depending on the prior and 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 produces summary statistics even closer to the data. For full details, see Sisson et al. (2007) (Also note the errata, Sisson et al., 2009).

Under ABC-PRC, posterior estimates for each parameter are 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. Fits were run such that tolerances for each experimental fit were roughly equal.

Appendix B – Comparing two rehearsal approaches for segmented lists

As discussed in the main text, there is some ambiguity about how participants might respond to the transition between silent and produced items in these sorts of segmented lists. In particular, the RFM assumes that rehearsal becomes less likely as the number of presented items increases. However, we assume in the models here that the rehearsal process is "reset" in some sense when the production modality (produced or silent) changes, so that rehearsal may restart part way through the list. That this is necessary seems clear from our preliminary attempt to fit the data with a model without this resetting.

However, there are many different ways to implement such a "reset", and we have no good empirical or theoretical justification for choosing one over the others. In particular, there are two extreme cases we might consider,

(1) At presentation of the first S (P) item, participants begin the rehearsal process again, rehearsing all subsequently presented items as if they were the first ones in a list. However, they ignore (or do not attempt to rehearse) any previously presented P (S) items that came before the transition.

(2) At presentation of the first S (P) item, participants begin the rehearsal process again, rehearsing all subsequently presented items as if they were the first ones in a list. In addition, they also rehearse all previously presented P (S) items that came before the transition with the same likelihood as the first S (P) item.

Neither of these possibilities is especially realistic. Option (1) is arguably more plausible (consider the case of a 100-item list, it is extremely unlikely that participants would successfully rehearse all 50 preceding items on presentation of item 51). However, in all likelihood, the truth lies somewhere between them. Importantly, these two options have the virtue of being possible to implement without adding any extra parameters or flexibility to the RFM. In the main text, we

showed fits from option (1), but we also wanted to compare the fits from the two options. We did this partly for completeness, but also to show that the model's ability to reproduce the qualitative features of the data does not depend on this choice.

We fit a new version of the RFM where rehearsal includes all items presented before the transition to compare it with the version used in the main text which does not attempt rehearsal of the items presented before the transition. We refer to these two versions as 'Full Rehearsal' and 'Limited Rehearsal' respectively. Fits were performed in an identical way for both versions, and there were no additional parameters required. Results are shown in the figures below.

In sum, we can see the following,

(1) Overall fits are better for the Limited Rehearsal version. This is perhaps unsurprising in view of what we mentioned above.

(2) There are nevertheless some conditions where the Full Rehearsal version does well, suggesting that the truth is probably somewhere in between the two extremes.

(3) The qualitative patterns in the data are reproduced by both model versions. The basic predictions of the RFM are independent of how we deal with the impact of a mid-list modality change.

As we outlined above, both investigated versions have somewhat unrealistic elements to them. It would be interesting to gather data on exactly how participants approach this, with the obvious caveat that we expect some individual variations in strategy. Nonetheless, the important point is that the key predictions of the RFM do not depend on this detail, so we can safely leave this for a future study.

Figure B1

Comparison between data and the two model versions for Experiment 1A – Experimenter Present

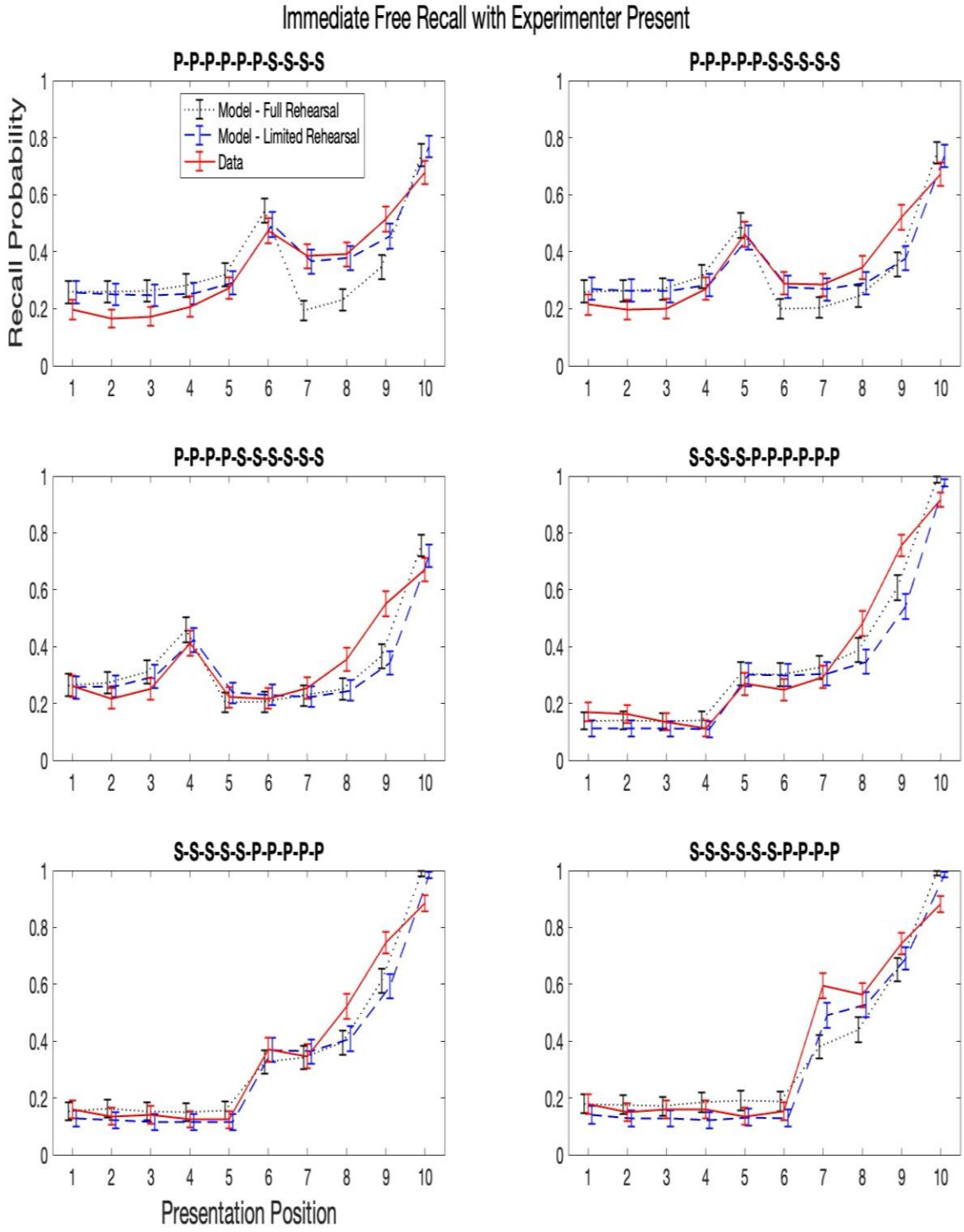


Figure B2

Comparison between data and the two model versions for Experiment 1B – Experimenter Absent

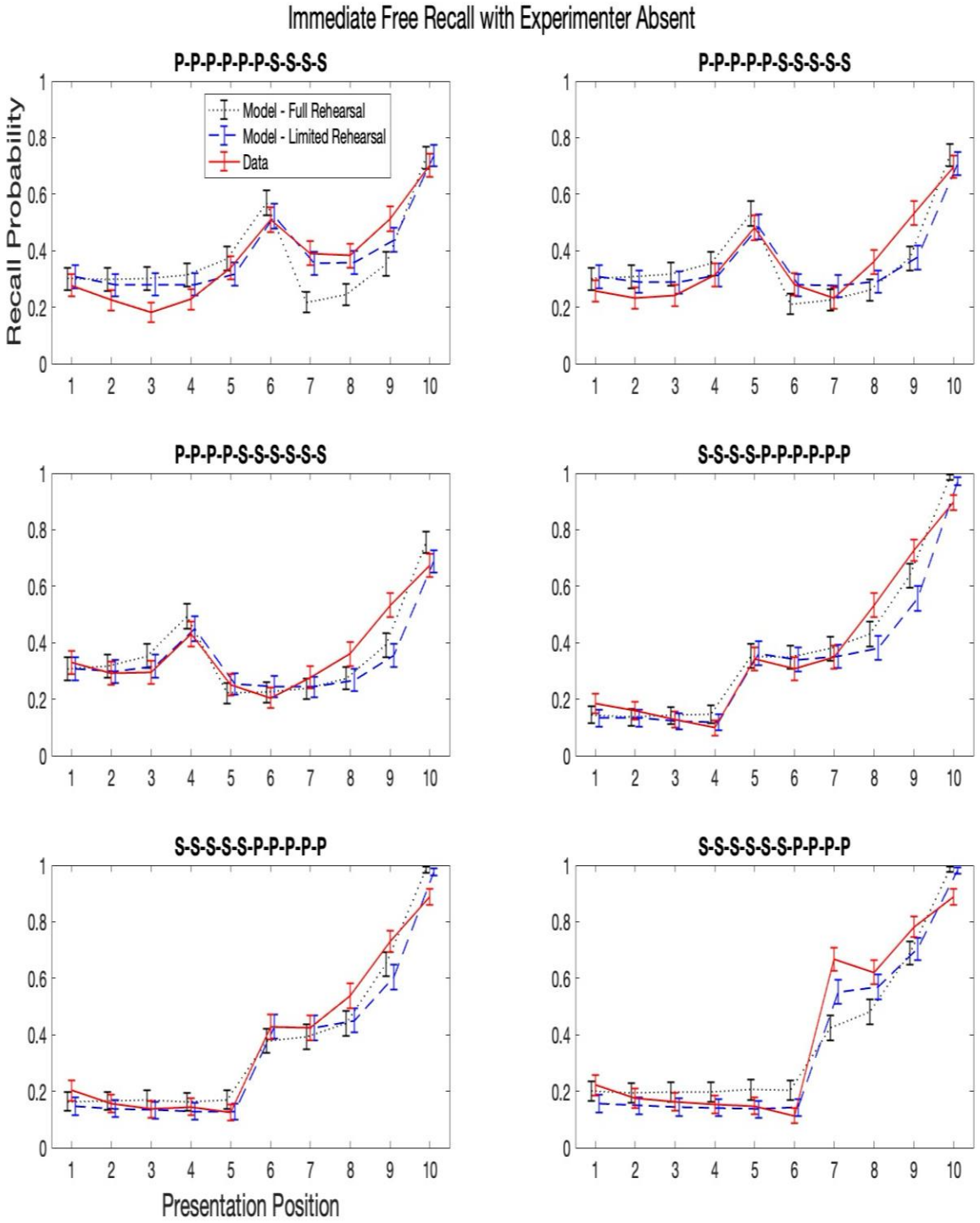


Figure B3

Comparison between data and the two model versions for Experiment 2A – Experimenter Present

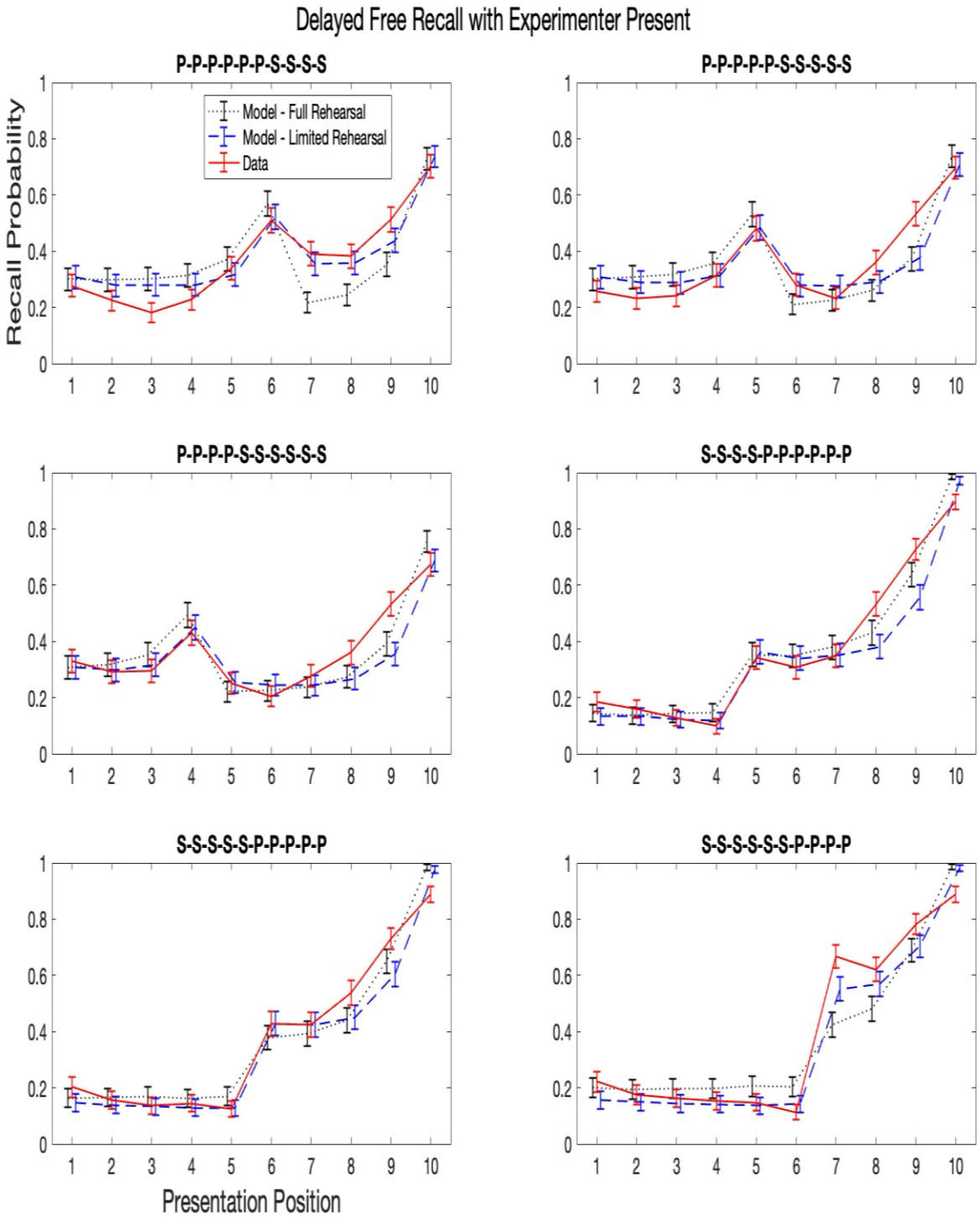


Figure B4

Comparison between data and the two model versions for Experiment 2B – Experimenter Absent