Do semantic and phonological categories lead to proactive interference in immediate and delayed serial recall tasks?

Sho Ishiguro[1,2], René-Pierre Sonier[1], Dominic Guitard[3], Jean Saint-Aubin[1]

[1]Université de Moncton, Canada

[2]Japan Society for the Promotion of Science, Japan

[3]Cardiff University, United Kingdom

Correspondence: École de psychologie, Université de Moncton, Moncton, New Brunswick, E1A 3E9, Canada. (Email: sho.ishiguro@umoncton.ca)

Abstract

Proactive interference (PI) has received considerable attention in research on working memory (WM), but empirical findings have been mixed. The present study examined whether this inconsistency may partly reflect the distinction between item and order memory. We hypothesized that PI would be more evident in item memory than in order memory and tested this hypothesis by applying item and order memory scoring methods to serial recall tasks. We also systematically manipulated two types of categories (semantic and phonological) and two recall tasks (immediate and delayed serial recall). For semantic categories, PI effects were consistently observed in item memory: performance increased following category shifts (e.g., animal words vs. flower words) in both immediate recall (Experiment 1A) and delayed recall (Experiments 1B and 3). In contrast, for phonological categories, reliable PI effects in item memory were observed under delayed recall (Experiment 2B), but not under immediate recall (Experiment 2A). Across these experiments, no clear PI effect was observed in order memory. These findings support the hypothesis that PI is more evident in item memory than in order memory and further suggest that the type of information may also influence PI. We also discuss how our delayed task relates to and differs from the Brown-Peterson task, a well-established task used to assess PI, in light of the item-order hypothesis. The distinction between item and order memory appears to be key to understanding PI in WM and may provide a useful framework for synthesizing past findings and guiding future research.

[246/250 words]

Keywords: working memory, proactive interference, semantic maintenance, phonological maintenance

Do semantic and phonological categories lead to proactive interference in immediate and delayed serial recall tasks?

After changing your password, you keep typing the old one and cannot remember the new one. This is an everyday example of *proactive interference* (PI), a phenomenon in which old memories interfere with new ones (M. Hamilton et al., 2022; Kliegl & Bäuml, 2021; Underwood, 1957).

As forgetting and remembering are two sides of the same coin, identifying the causes of forgetting is essential for our understanding of the memory system (Levy & Anderson, 2002; Nishiyama et al., 2025; Nørby, 2015). As a cause of forgetting, PI has received considerable attention in memory research (Bäuml & Kliegl, 2013; Craik & Birtwistle, 1971; Kliegl & Bäuml, 2021; Underwood, 1957), including studies of working memory (WM) (Andersson et al., 2025; A. C. Hamilton & Martin, 2007; M. Hamilton et al., 2022; Kane & Engle, 2000; Loosli et al., 2014; Lustig et al., 2001; May et al., 1999; Zhang et al., 2010).

Some theorists have emphasized the importance of PI in WM. Jonides and Nee (2006 , p. 181) pointed out “proactive interference is an important determinant of the success of working memory.” Engle (2002 , p. 20) even noted “dealing with the effects of proactive interference is one of the primary functions of WM.” These perspectives support the theoretical position that WM is susceptible to PI and managing its effects is critical to WM functioning.

However, an alternative view has also been proposed: WM is free from PI, and observed PI effects arise from long-term memory (LTM) (Oberauer & Awh, 2022; Oberauer et al., 2017). Distinguishing WM from LTM would not, by itself, resolve the debate between these two accounts, because the same PI effect could be attributed to WM or LTM depending on the theoretical assumptions adopted. Nevertheless, the WM-LTM distinction provides a useful framework for considering the source of PI effects.

This distinction has been widely recognized (Healy & McNamara, 1996; Thorn & Page, 2008) and is supported by neuropsychological and computational evidence. Neuropsychological studies have documented a double dissociation between WM and LTM, with selective impairment of either system alongside preservation of the other (Baddeley & Warrington, 1970; Warrington & Shallice, 1969; for recent studies, see Martin et al., 2020; Rapp et al., 2016). Computationally, conceptualizing WM solely as the activated portion of LTM cannot solve the “problem of two” (Norris, 2017). Activation of an item’s LTM representation cannot distinguish between multiple occurrences of that item (e.g., “1” in “1, 3, 1”). These findings are often used to support the existence of WM mechanisms that are functionally distinct from LTM.

Given the distinction between WM and LTM, proponents of the PI-free WM view have argued that PI does not occur when the number of items remains within WM capacity–approximately four items, assumed to be held in the focus of attention (e.g., Cowan et al., 2020). Once the number of items exceeds this capacity, maintenance may draw on LTM, which is vulnerable to PI. These proponents have supported this account by showing that PI is observed with longer lists (or larger set sizes) but not with shorter lists (or smaller set sizes) (Bartsch & Oberauer, 2023; Cowan et al., 2005; Halford et al., 1988; Oberauer & Awh, 2022). For instance, Bartsch and Oberauer (2023) provided evidence for PI with larger set sizes (six and eight items) and evidence against PI with smaller set sizes (two, three, and four items). One concern, however, is that maintaining a small number of items may produce ceiling effects, thereby obscuring PI (Cowan et al., 2005). More critically, other researchers have demonstrated that PI can occur even when three items are maintained (Carroll et al., 2010; Ralph et al., 2011; Shipstead & Engle, 2013).

Conversely, Oberauer and Bartsch (2023) failed to observe PI with a longer list length (i.e., seven items), which should have produced PI according to the capacity-based account.

Furthermore, the finding that WM capacity is limited to approximately four items can, in principle, be accommodated within the PI-prone view of WM. Endress and Potter (2014) demonstrated that capacity estimates fell within the conventional range of approximately four items under high-PI conditions, whereas under low-PI conditions, they substantially exceeded this limit. The authors argued that PI reduces capacity estimates to approximately four items.

Together, these findings call the capacity-based account into question, suggesting that list length may not determine whether PI is observed. They also shift the focus toward identifying the conditions under which PI emerges and the factors that modulate its magnitude (e.g., Ralph et al., 2011). One potentially important but largely overlooked factor is the type of memory being assessed. Specifically, PI may differentially affect memory for item identities (item memory) and serial order (order memory). This possibility is plausible because research on WM phenomena other than PI has demonstrated dissociable effects on item and order memory. For example, rhyme-based phonological similarity within a list facilitates item memory but impairs order memory (Fallon et al., 1999; Gupta et al., 2005; Nimmo & Roodenrys, 2004). Likewise, within-list semantic similarity benefits item memory while disrupting order memory (Dauphinee et al., 2026; Saint-Aubin et al., 2005). Accordingly, the distinction between item and order memory has served as a useful analytical framework in WM research (Saint-Aubin & Poirier, 1999).

To evaluate the empirical basis of this possibility, we briefly review previous PI studies involving sequences of items, such as serial recall and serial order reconstruction. These tasks are informative because they can provide rich information about item and order memory (Ishiguro & Saito, 2024; Sonier et al., 2025).

Previous findings through the lens of item and order memory

Studies placing relatively high demands on order memory have generally provided limited evidence for PI. Oberauer and Bartsch (2023) conducted experiments in which mini-blocks of four serial recall trials alternated with mini-blocks involving another task. At the beginning of each mini-block, a new stimulus pool was introduced, and the same pool was then used in the mini-block. In serial recall trials, seven words were sampled from a pool of 10, resulting in substantial item overlap across trials within a mini-block. PI was therefore expected to build up over successive trials and to be released at the transition to the next mini-block, when both the task and the stimulus pool changed. Overall, however, their results provided little evidence for PI.

Importantly, Oberauer and Bartsch's serial recall procedure reduced item memory demands and placed greater emphasis on order memory for two reasons. First, because a fixed pool of words was used within each mini-block, participants repeatedly encountered familiar item identities rather than an entirely new set of words on every trial. The PI manipulation, in fact, relied on the repeated presentation of the same items. Second, all 10 words in the pool were presented as response options at test, further reducing the need to retrieve item identities. With this response format, their serial recall task resembled an order reconstruction task, which is designed to measure order memory (Whiteman et al., 1994).

Using an order reconstruction task, Baddeley et al. (2018) obtained suggestive evidence of a PI effect for phonological information. In a repeat-similar condition, the same set of phonologically similar items was repeated across trials but in a different order (e.g., Trial 1: pan, cat, bat, ban, pad, man; Trial 2: ban, man, bat, pan, cat, pad). In a change-similar condition, items were drawn from different sets on each trial (e.g., Trial 1: pan, cat, bat, ban, pad, man; Trial 2: hen, bed, jet, pen, fed, pet). The six items used on each trial were shown again as response options at test, and participants were asked to reconstruct their original order. The results showed that overall performance averaged across trials was better for the change-similar condition than for the repeat-similar condition, a pattern consistent with PI.

However, their detailed inspection of trial-by-trial performance in the repeat-similar condition did not indicate the decline expected from PI buildup (i.e., a decline in performance across trials). Furthermore, they estimated the within-list similarity effect using dissimilar lists (e.g., man, pen, rim, cod, bud, peel) and showed the within-list similarity effect was larger than the PI effect. Baddeley and colleagues therefore emphasized the importance of the within-list similarity over PI for phonological information.

A different pattern has emerged in tasks that place greater demands on item memory (Emery et al., 2008; for item recognition, see Carroll et al., 2010). Emery et al. (2008) demonstrated clear PI effects with an operation span task (Bunting, 2006), using a procedure that emphasized item memory. Across three successive trials, participants were presented with different items drawn from the same semantic category, with no item overlap across trials (e.g., animals; Trial 1: lion, goat, zebra, bull; Trial 2: lamb, squirrel, cow, donkey; Trial 3: panther, moose, camel, pig). The semantic category was then changed after every third trial (e.g., body parts; Trial 4: foot, ankle, hip, shoulder). Unlike in an order reconstruction task, no response options were presented at test; participants had to retrieve the words themselves. Furthermore, recall was scored without regard to output order, a scoring method generally considered to reflect item memory (Saint-Aubin & Poirier, 1999). Their results suggest that PI built up across trials within a semantic category and was released when the category changed.

The Current Study

Taken together, the available evidence tentatively suggests that PI is more readily observed in item memory than in order memory (hereafter *the item-order hypothesis*). However, previous studies differed not only in their memory demands but also in their procedures and the types of information (e.g., phonological versus semantic categories), leaving their findings open to alternative explanations. For example, the contrasting findings of Baddeley et al. (2018) and Emery et al. (2008) may reflect the use of phonological versus semantic categories rather than a difference between item and order memory. In addition, the distractor task embedded in the operation-span procedure may have contributed to the PI observed by Emery et al. (2008). Specifically, each memory item was preceded by an arithmetic equation, such as "(2 + 7) – 7 = 2," which participants were required to verify. By contrast, neither Baddeley et al. (2018) nor Oberauer and Bartsch (2023) included a distractor task.

The present study therefore tests the item-order hypothesis of PI by systematically manipulating two factors: category type (semantic versus phonological) and recall type (immediate serial recall versus delayed serial recall). Each of the first four experiments addresses one of the four combinations of category type and recall type. For example, Experiment 1B tests whether semantic categories produce PI in delayed serial recall with a distractor-filled retention interval. A follow-up experiment addresses methodological issues left unresolved in the first four experiments.

In all experiments, six words are presented sequentially on each trial, and participants are asked to recall them in the same order, either immediately or after a distractor-filled retention interval. Trials are organized into four-trial blocks. Within each block, all words are drawn from a single semantic category or phonological category. A new category is introduced at each block transition. This manipulation of category repetition and change builds on previous WM studies on

PI (Emery et al., 2008) as well as classic studies using the Brown-Peterson task (Wickens, 1970, 1973).

We adopt the well-established category-based PI paradigm but extend it by assessing item and order memory separately. In addition to the standard *correct-in-position* scoring, we use *item-correct* as a measure of item memory and *conditionalized order recall* as a measure of order memory (Ishiguro et al., 2026). The latter two methods allow us to test whether PI differs between item and order memory.[1]

The choice of tasks and list length is based on two considerations. First, both immediate and delayed serial recall with six-word lists have previously been used to assess WM (Ishiguro et al., 2026; Kowialiewski & Majerus, 2020). Memory models also often focus on serial recall of lists containing six or more items because such tasks provide detailed information about order memory (Farrell, 2012; Guitard et al., 2025b; Saint-Aubin et al., 2023). Second, both tasks fall within a broad definition of WM. Although several definitions exist in the literature, Cowan provides the following *generic definition of WM*: “the ensemble of components of the mind that

[1] Oberauer (2022) repeated and changed stimulus classes (e.g., digits vs. words) to test PI in serial recall. The reported results for correct-in-position scores were mixed: Experiment 1 provided evidence against PI, whereas Experiment 3 provided extreme evidence for PI (see Table 2 in Oberauer, 2022). From the perspective of the present study, this inconsistency underscores the importance of considering measures other than correct-in-position when assessing PI.

hold a limited amount of information temporarily in a heightened state of availability for use in ongoing information processing" (Cowan, 2017 p. 1163). By design, this definition does not specify the underlying processes or mechanisms. It, for example, leaves open the possibility that LTM contributes to WM functioning (Cowan, 2017). Rather than committing to a strong theoretical position, we adopt this theoretically neutral definition as our starting point. Consistent with this definition, both immediate and delayed serial recall require participants to temporarily maintain a limited amount of information for subsequent use.

In sum, the item–order hypothesis predicts that PI should be more consistently evident in item-correct performance than in conditionalized order recall. Such a dissociation may, however, vary as a function of category type and recall type. We test our prediction across these conditions in a series of experiments.

Experiment 1

Experiments 1A and 1B presented six blocks of four trials and an additional block of two trials. Critically, semantic categories were repeated in successive trials, but they varied across blocks. We predicted that repeating semantic categories would result in PI while switching them would lead to PI release. This prediction was tested by (1) a main effect of PI level (i.e., performance for the first, second, third, and last trials of each block) and (2) a critical contrast (i.e., performance for the first vs. fourth trials). Experiments 1A and 1B differed only in the WM task; the former used the immediate serial recall task, whereas the latter used the delayed serial recall task.

Method

Transparency and Openness

The data, stimuli, and analysis scripts in this and following experiments are available on the Open Science Framework (OSF) repository at: https://osf.io/wtnxr. None of the experiments, including their designs and analyses, was preregistered. All experiments were approved by the research ethics committee of Cardiff University. Informed consent was obtained from all participants.

Sample Size

We used Bayesian inference in the data analysis, and thus sample size planning was not necessary (Wagenmakers et al., 2018). Nevertheless, for compatibility with frequentist inference, we ran a power analysis. We set a sample size of 60 participants and estimated the sensitivity using G*Power (Faul et al., 2009, 2007) with setting ($1 - \beta$) at .80 and α at .05. The power analysis revealed that a minimum effect size of Cohen's $f = 0.15$ (a small to medium effect) was detectable for the main effect of PI level in a repeated-measure ANOVA and a minimum effect size of Cohen's $dz = 0.37$ (a small to medium effect) was detectable for the critical contrast in a paired *t*-test. Therefore, a sample size of 60 participants was used for all experiments.

Participants

Each experiment recruited 60 volunteer students from Cardiff University. All participants self-identified as fluent English speakers and had no visual, hearing or cognitive impairment. They received course credit as compensation. One participant in Experiment 1A responded "SKIP" on most trials, resulting in zero accuracy for the targeted trials (Trials 3–26); their data were therefore excluded from the analysis. In Experiment 1A ($N = 59$), nine participants self-identified as men, 49 self-identified as women, and a participant self-identified as other. Their ages ranged from 18 to 32 years ($M = 19.39$, $SD = 1.88$). In Experiment 1B ($N = 60$), five participants self-identified as men, 51 self-identified as women, and four participants self-identified as other. Their ages ranged from 18 to 23 years (*M* = 19.25, *SD* = 1.26). Experiments 1A and 1B took approximately 15 and 25 minutes, respectively.

Materials

We selected eight semantic categories from the category norms (Van Overschelde et al., 2004). For each category, 24 words were chosen as reported in Appendix A. For three of the eight categories (*four-footed animal, fruit,* and *part of the human body*), all 24 words were directly taken from the norms. For two categories (*fish* and *musical instrument*), we replaced potentially problematic words (e.g., phonologically similar words) with more suitable words fitting the category. The remaining three categories (*flower*, *color, occupation or profession*), which contained fewer than 24 unique words in the norms, were augmented with additional ones. The procedure for selecting categories and words is detailed below.

First, 25 categories meeting the minimum requirement of 24 words were retrieved from a total of 70 categories in the category norms (Van Overschelde et al., 2004). Then, from these 25 categories, three categories were retained after excluding (a) categories including mostly proper nouns (e.g., male name and football team name categories), (b) categories having many phonologically similar words (e.g., most sports words ended with "-ing") or compound words (e.g., "cruise ship"), (c) an ad-hoc category (i.e., a thing that makes noise), (d) categories similar to another category (e.g., fruit category was retained but its similar category, vegetable, was discarded), (e) a category containing uncommon words (i.e., chemical elements). For the retained categories, we retrieved 24 words from the norms. Then, we adopted two of the previously rejected categories (i.e., categories having phonologically similar or compound words) by modifying their words. Specifically, we replaced problematic words with other words fitting the semantic category.

Lastly, we returned to three categories with fewer than 24 words and added suitable words to each until they had 24 words in total.

Our category and word selection procedure ensured that the eight semantic categories were distinct from one another and each category included phonologically dissimilar words. These categories had equivalent word length (M = 5.85, SD = 1.84, min = 4.13, max = 6.75).

Procedure

Experiments 1A and 1B were programmed with PsyToolkit (Stoet, 2010, 2017). Participants were informed that a list of six words would appear in the center of the screen, and their task was to recall the words in the same order as presented. In each trial, six words were presented sequentially at a rate of one word per second (1000ms on, 0ms off) in white, 20-point, uppercase Times New Roman font on a black background. In Experiment 1A, the offset of the last word was immediately followed by the recall phase. In contrast, in Experiment 1B, recall was delayed by an 18-second distractor task. During this filled delay period, participants judged whether a single digit (0–9) presented on the screen was even or odd by pressing the M or Z key, respectively. Each digit was randomly selected and remained on the screen until a response key was pressed, after which the next digit appeared.

In the recall phase, participants recalled the words by typing them on a keyboard. Each response was displayed at the center of the screen in white text inside a blue box until the enter key was pressed, after which the response was cleared, and the participant could recall the next word. Above the response field, the position of recall was indicated in red text (i.e., “Type the first word”, “Type the second word”, and so forth) so that participants were aware of which word was asked about. Participants were instructed to type “SKIP” if they could not recall a word at a given position. They were informed that they could not backtrack to change a previous response. After the last recall, they pressed the space bar to move on to the next trial.

Both experiments consisted of six blocks of four trials each, followed by a final block of two trials, for a total of 26 trials. For each participant, seven of the eight categories were randomly selected and assigned to blocks. The 24 words in each category were divided into four six-word lists, which were presented across four successive trials (one block). The final block, however, included only two lists for two trials. This is because a key contrast was performance before and after a category shift. The addition of the final block increased the number of observations of a category shift (i.e., the shift from the sixth block to the final block). No practice trials were included in the experiments. The category selection, category order, word allocation to four lists, and word order within each list were randomized freshly for each participant. Our PI manipulation was based on category repetition rather than word repetition; each participant saw each word only once in the experiment.

Analysis Plan

Recall performance was scored using three methods: correct-in-position, item correct, and conditionalized order recall (Ishiguro et al., 2026). Correct-in-position scoring counted words that were recalled in their correct positions; this scoring method captured both item and order memory. Item correct scoring counted correctly recalled words regardless of their recalled position. As this procedure focused on item availability, item correct was used as an item memory measure. Conditionalized order recall was calculated as correct-in-position scores divided by item correct scores. This corresponded to the ratio of words in the correct position among correctly recalled words, and served as an order memory measure.

The current study focused on PI across trials and thus quantified performance as proportion correct at each trial. Proportion correct data were analyzed with the R package, *BayesFactor* (version 0.9.12–4.7) (Morey et al., 2024). Specifically, we used the *anovaBF* function for Bayesian ANOVAs and the *ttestBF* function for Bayesian *t*-tests. The default prior, Jeffreys-Zellner-Siow (JZS), was applied. Bayes Factors (BFs) were interpreted according to a classification scheme (Schönbrodt & Wagenmakers, 2018): 1–3 (anecdotal evidence); 3–10 (moderate evidence); 10–30 (strong evidence); 30–100 (very strong evidence); and > 100 (extreme evidence). We used two notations: BF_{10} indicates the BF for the alternative hypothesis while BF_{01} ($1/BF_{10}$) indicates the BF for the null hypothesis.

Typographical errors in participants' responses were corrected prior to scoring. These errors included, for example, *silveer* for *silver*, *cler* for *clerk*, and *lavendar* for *lavender*. Responses were compared against all words presented within a trial and corrected to the closest word with a Levenshtein distance of one. Specifically, a Levenshtein distance of one corresponds to a single edit operation of either insertion (e.g., *silveer*, adding "e" to *silver*), deletion (e.g., *cler*, removing "k" from *clerk*), or substitution (e.g., *lavendar*, replacing "e" in *lavender* with "a"). In rare cases, responses were ambiguous regarding whether they represented errors or intrusions. For example, the response *golf*, when appearing along with other color-word responses, could reflect a typo of *gold* or an intrusion. To ensure reproducibility and minimize researcher degrees of freedom, the above-mentioned procedure was strictly applied to all responses. Note that we also analyzed raw uncorrected data and confirmed the key findings reported below (see https://osf.io/wtnxr).

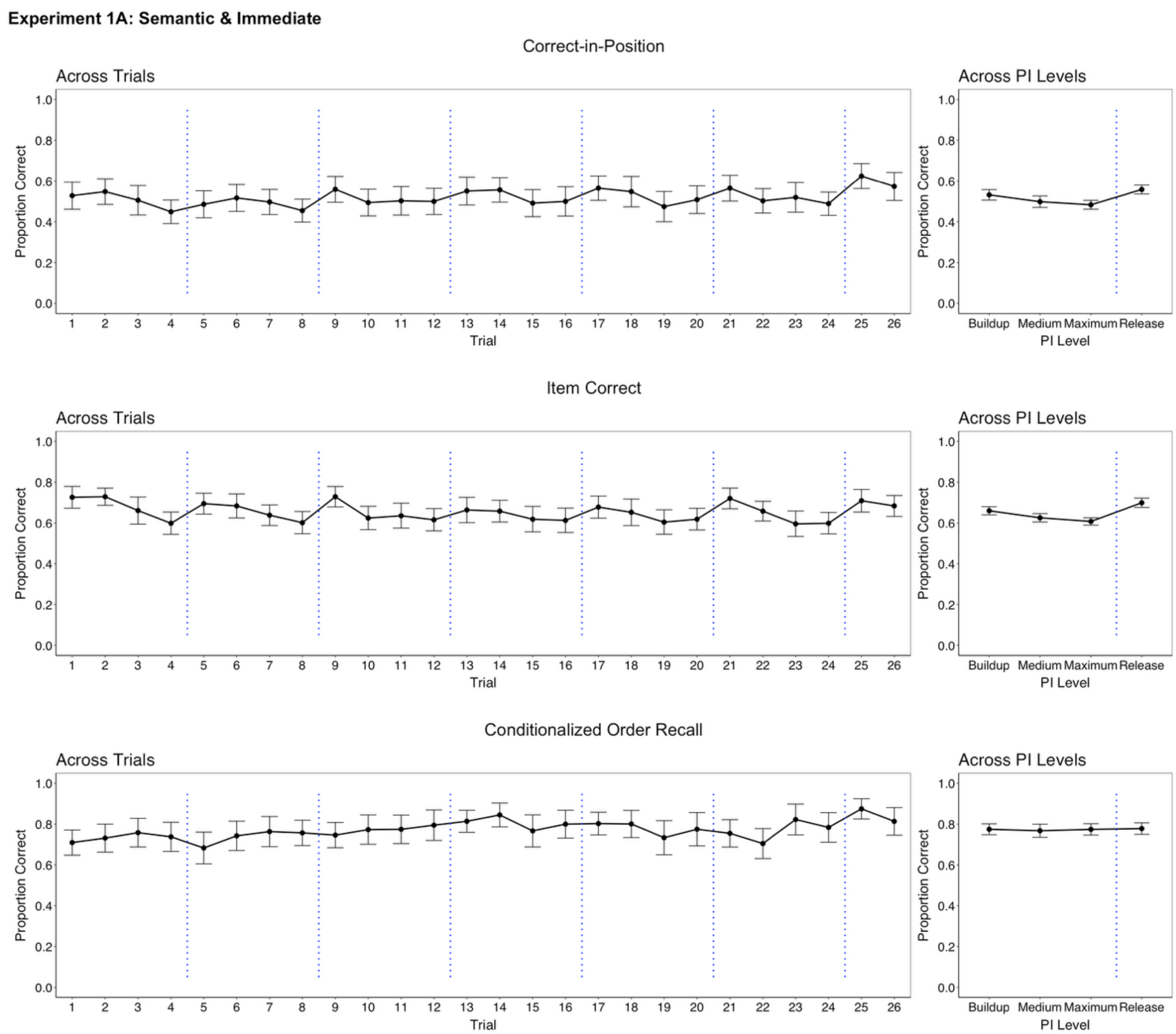
Results and Discussion

Figure 1 shows two sets of results from Experiment 1A: (1) performance across trials and (2) performance across PI levels. The trial-by-trial data indicate that semantic category shifts led to PI release for both correct-in-position and item correct scores. Performance improved following a category shift (top-left and middle-left subplots; compare performance before and after the blue dotted lines). By contrast, this pattern was absent for conditionalized order recall (bottom-left subplot). Similarly, in Experiment 1B (Figure 2), scalloped patterns were present for both correct-in-position and item-correct scores, but absent for conditionalized order recall.

To examine performance across PI levels, trial performance was averaged according to trial position within each block (see the right subplots in Figures 1 and 2). Trials were categorized based on expected PI levels: first trials of each block as *PI release*, second trials as *PI buildup*, third trials as *PI medium*, and fourth trials as *PI maximum*. The first two trials of the first block (Trials 1 and 2) were excluded while the two trials of the final block (Trials 25 and 26) were included, ensuring an equal number of trials per PI level. Beyond balancing the number of trials, there were two additional reasons for excluding the very first trials. First, because the first block had no preceding block, there was no category shift or corresponding PI release. Second, as our experiments did not include practice trials, performance on the first few trials was potentially unstable. The following analyses targeted performance across PI levels.

Figure 1

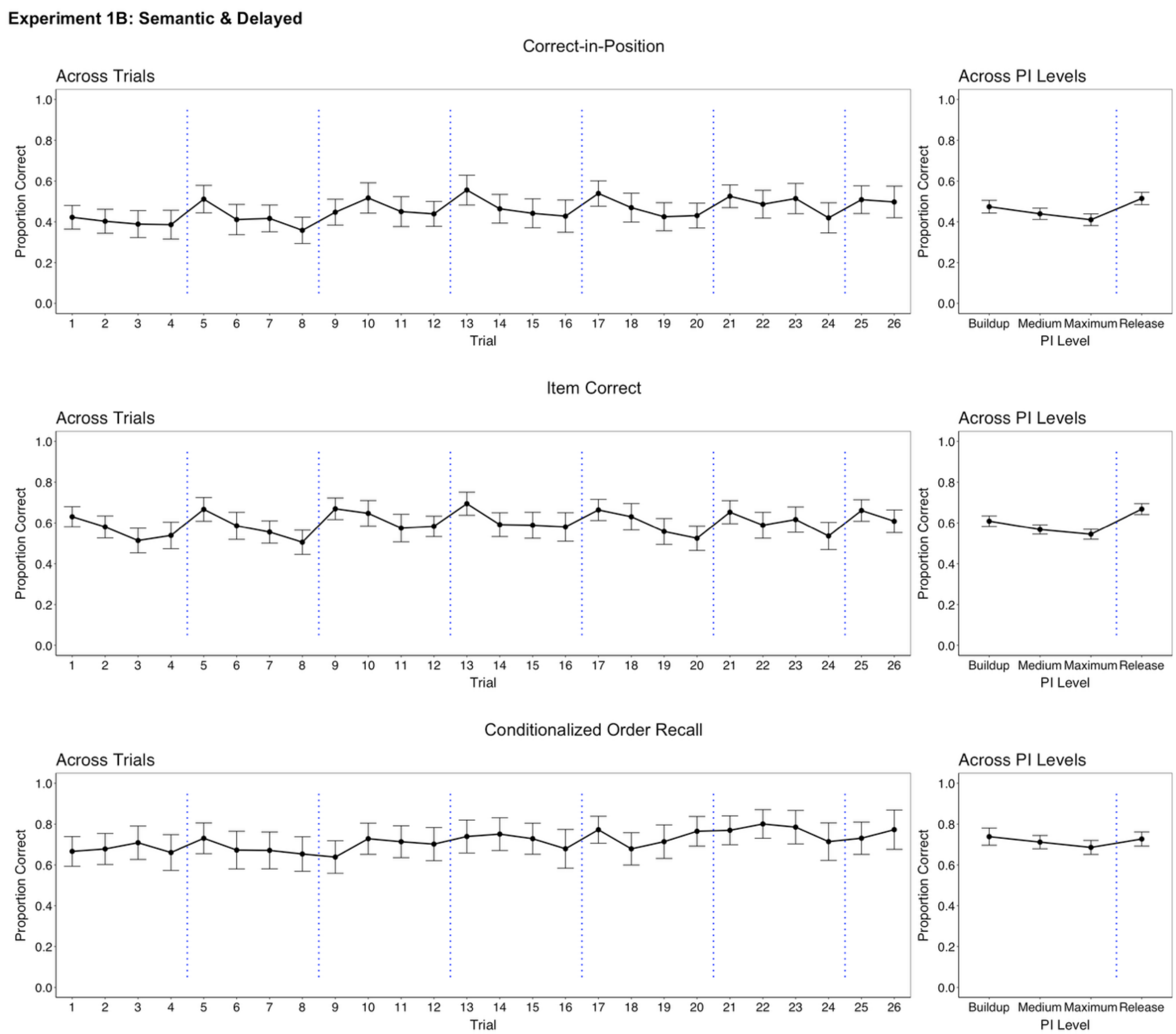
Results of Experiment 1A



Note. Error bars represent 95% confidence intervals for a within-participants factor (Morey, 2008).

Figure 2

Results of Experiment 1B



Note. Error bars represent 95% confidence intervals for a within-participants factor (Morey, 2008).

Experiment 1A

Correct-in-Position. Immediate recall in Experiment 1A was first assessed. For correct-in-position (top-right subplot in Figure 1), Bayesian one-way ANOVA with four levels indicated a main effect of PI level, $BF_{10} = 222.48$, $\eta_p^2 = 0.116$, supporting performance differences among PI buildup ($M = 0.53$, $SD = 0.10$), PI medium ($M = 0.50$, $SD = 0.11$), PI maximum ($M = 0.48$, $SD = 0.08$), and PI release ($M = 0.56$, $SD = 0.08$). Critically, Bayesian paired *t*-test revealed that performance was higher for PI release than for PI maximum, $BF_{10} = 1.13 \times 10^4$, $dz = 0.70$. Other contrasts were not supported, with some evidence for the null effect: $BF_{01} = 1.79$ for PI buildup vs. PI medium; $BF_{01} = 5.00$ for PI medium vs. PI maximum; and $BF_{01} = 2.10$ for PI release vs. PI buildup.[2]

Item Correct. Similar patterns were observed with the item memory index (middle-right subplot). Bayesian ANOVA provided evidence for performance differences across PI levels, $BF_{10} = 2.79 \times 10^6$, $\eta_p^2 = 0.214$. A critical comparison suggested that PI release ($M = 0.70$, $SD = 0.09$) outperformed PI maximum ($M = 0.61$, $SD = 0.07$), $BF_{10} = 3.29 \times 10^5$, $dz = 0.82$. Other contrasts

---

[2] Although PI is assumed to build up across trials, the corresponding decline in performance is not necessarily linear (e.g., Hendrikx, 1986). We therefore used ANOVA and *t*-tests for the primary analyses. Results from supplementary linear analyses, in which the four PI levels were coded as -1.5 (release), -0.5 (buildup), 0.5 (medium), and 1.5 (maximum), are available on the OSF page.

were less clear: $BF_{10} = 2.15$ for PI buildup ($M = 0.66$, $SD = 0.08$) vs. PI medium ($M = 0.63$, $SD = 0.08$); $BF_{01} = 2.88$ for PI medium vs. PI maximum; and $BF_{10} = 2.78$ for PI release vs. PI buildup.

Conditionalized Order Recall. A flat pattern of conditionalized order recall was confirmed by statistical tests (bottom-right subplot). Bayesian ANOVA indicated equivalent performance across four levels, $BF_{01} = 43.88$, $\eta_p^2 = 0.002$ (PI buildup: $M = 0.77$, $SD = 0.10$; PI medium: $M = 0.77$, $SD = 0.12$; PI maximum: $M = 0.77$, $SD = 0.11$; PI release: $M = 0.78$, $SD = 0.11$). Follow-up Bayesian *t*-tests targeting the aforementioned contrasts showed evidence for the null effect, $6.66 < BF_{01} < 6.91$.

Experiment 1B

Correct-in-Position. With a delayed recall task (Figure 2), Experiment 1B yielded results comparable to Experiment 1A. For correct-in-position, we confirmed a main effect of PI level, $BF_{10} = 1.81 \times 10^3$, $\eta_p^2 = 0.137$ (PI buildup: $M = 0.47$, $SD = 0.12$; PI medium: $M = 0.44$, $SD = 0.11$; PI maximum: $M = 0.41$, $SD = 0.11$; PI release: $M = 0.51$, $SD = 0.12$) and a critical difference between PI release versus PI maximum, $BF_{10} = 1.37 \times 10^4$, $dz = 0.70$. Other contrasts remained anecdotal evidence for the null effect ($1.53 < BF_{01} < 2.61$).

Item Correct. With item correct scores, the PI level's main effect, $BF_{10} = 8.28 \times 10^7$, $\eta_p^2 = 0.243$ (PI buildup: $M = 0.61$, $SD = 0.10$; PI medium: $M = 0.57$, $SD = 0.09$; PI maximum: $M = 0.54$, $SD = 0.09$; PI release: $M = 0.67$, $SD = 0.10$), and the critical release-maximum difference, $BF_{10} = 3.17 \times 10^6$, $dz = 0.89$, were supported. The differences in buildup vs. medium ($BF_{10} = 2.64$) and medium vs. maximum ($BF_{01} = 2.63$) were ambiguous. However, PI release scores were higher than PI buildup scores, supported by strong evidence, $BF_{10} = 11.12$.

Conditionalized Order Recall. Conditionalized order recall was largely insensitive to PI levels (PI buildup: $M = 0.74$, $SD = 0.16$; PI medium: $M = 0.71$, $SD = 0.13$; PI maximum: $M = 0.69$, *SD* = 0.13; PI release: *M* = 0.73, *SD* = 0.13): Neither the main effect in ANOVA (BF_{01} = 6.83, η_p^2 = 0.026), nor the critical contrast in *t*-test (BF_{01} = 1.21, *dz* = 0.25) was found. Other contrasts also indicated equivalent performance ($3.98 < BF_{01} < 6.55$).

In sum, the results of correct-in-position scores suggest that PI buildup/release occurred via semantic category shifts in both immediate and delayed recall. More fine-grained analyses suggested that this PI effect was driven by item memory. The magnitude of PI in item memory was similar for immediate and delayed recall. For instance, the difference between PI release and PI maximum in item memory was supported by comparable effect sizes (*dz* = 0.82 and *dz* = 0.89 for immediate and delayed recall, respectively). By contrast, neither immediate nor delayed recall provided evidence of PI in order memory; if anything, the results suggested the absence of PI. This selective effect of PI on item memory supported the item-order hypothesis.

Appendix B also reports performance within trials (i.e., serial position curves). Importantly, PI level did not interact with serial position in any experiment or scoring method. The observed PI effects were not restricted to a specific position within a trial.

Experiment 2

Following Experiment 1 that focused on semantic categories, Experiment 2 addressed rhyme-based phonological categories. The general procedure was identical to that of the previous experiment, except that Experiment 2 used phonological categories instead of semantic categories and switched from laboratory-based to online experiments. The immediate and delayed serial recall tasks were again used in Experiments 2A and 2B, respectively.

Method

Participants

Separate samples of 60 participants were recruited for Experiments 2A and 2B through the online platform Prolific. They were compensated £2.20 for their participation. Recruitment criteria were as follows: (a) age between 18 and 30 years old; (b) native English speakers; (c) British nationality; (d) normal or corrected-to-normal vision; (e) no cognitive impairment or dementia; (f) no language-related disorders; and (g) an approval rating of at least 90% on prior submissions at Prolific. In Experiment 2A, 31 participants self-identified as men, 28 self-identified as women and one participant preferred not to report; with a mean age of 25.65 years old (SD = 3.33). In Experiment 2B, 46 participants self-identified as men and 14 self-identified as women; with a mean age of 24.73 years old (SD = 3.63). Experiments 2A and 2B took approximately 15 and 20 minutes, respectively.

Materials

Eight phonological categories were constructed, each consisting of 24 words with the same rhyme (Appendix A). Rhymes were defined as shared vowel–consonant endings. For example, the category EYT (eɪt) included *orate*, *eight*, *bait*, and *date,* while the category IHP (ɪp) included *dip*, *hip*, *kip*, and *lip*. Details of the category and word selection are described below.

We started by retrieving all words commonly listed in both the English Lexicon Project (Balota et al., 2007) and the Carnegie Mellon Pronouncing Dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict). By filtering words according to their last syllable, we identified 98 different rhyme categories, of which 93 categories contained at least 24 words. To keep the equivalent word length across categories, we then excluded words with more than five syllables, resulting in 39 out of 93 remaining categories. From these 39 categories, 13 distinct categories were selected by retaining a single category from each set of categories sharing the same vowel. For example, although our initial vowel–consonant criterion classified the categories IYK (iːk), IYD (iːd), and IYZ (iːz) as separate ones, only IYK was retained whereas IYD and IYZ were discarded. This procedure increased phonological differences across categories, thereby increasing the probability of PI release. Word pronunciations were further verified by referring to British English transcriptions of the International Phonetic Alphabet (IPA) (Mu-Sonic Ltd; https://tophonetics.com). Specifically, for each category, we included 24 words that had only single pronunciation and indeed contained the same final syllable. Finally, from these 13 categories of 24 words, eight were selected based on the lowest mean word length and syllable count. Shorter words were prioritized over longer ones as the latter were more likely to include additional unique syllables, thereby reducing within-category similarity. Our category and word selection was intended to minimize across-category similarity and maximize within-category similarity. All categories included similar mean word length (M = 4.27, SD = 0.70, min = 3.63, max = 4.75) and syllable count (M = 3.36, SD = 0.54, min = 2.92, max = 3.54).

Procedure and Scoring

The overall structure, consisting of six blocks and an additional block, followed that of the previous experiments. A key distinction in Experiment 2 was the use of web-based tasks combined with the manipulation of phonological categories. The category selection (i.e., sampling seven from the eight categories), category order, word allocation to four lists, and word order within each list were randomized for each participant same as in Experiment 1. Experiment 2A asked participants to perform immediate recall, whereas Experiment 2B involved delayed recall.

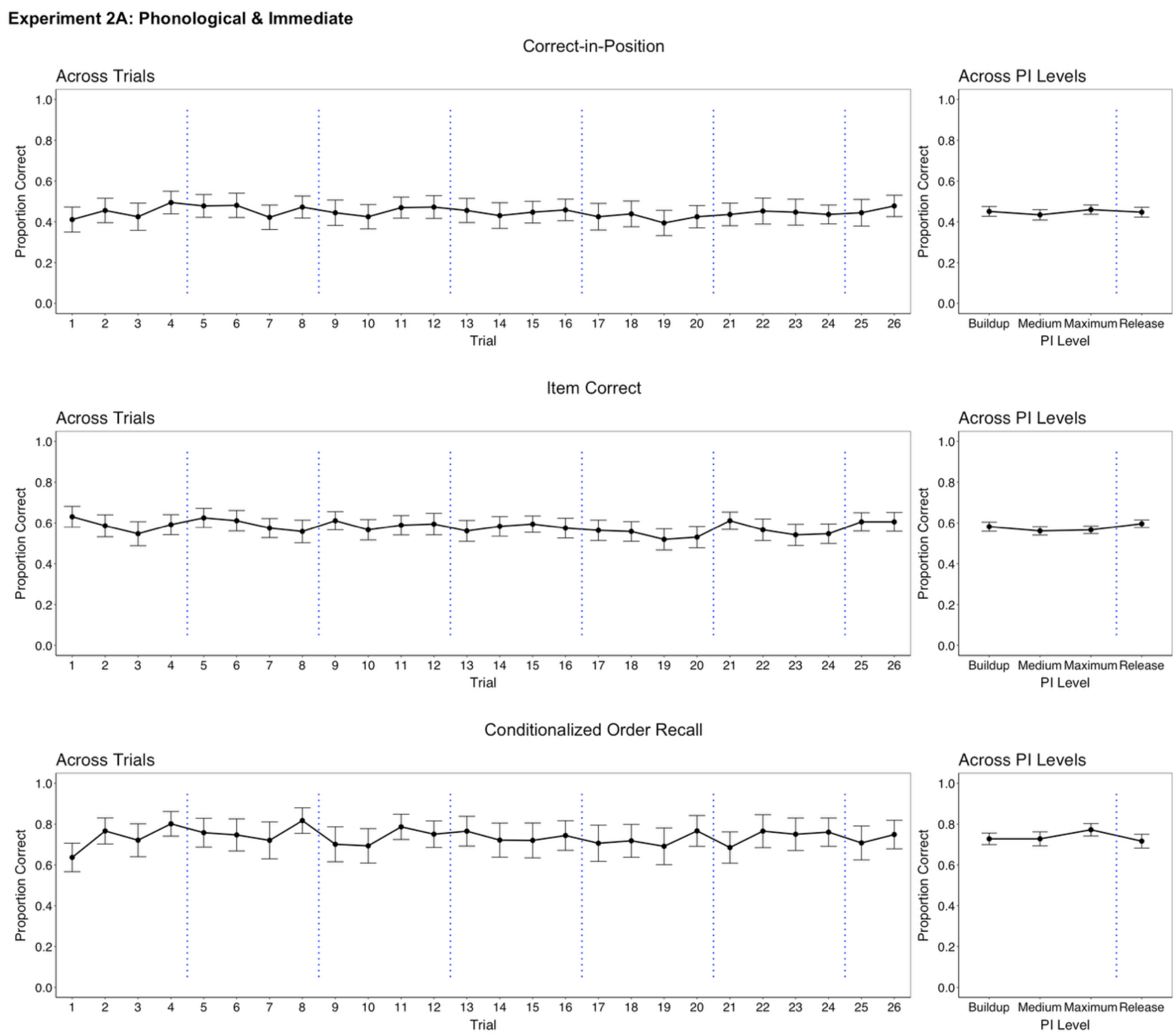
Response data from Experiments 2A and 2B were left uncorrected. This is because phonologically similar words tended to be orthographically similar and there were multiple correction candidates for a single response. For instance, the response *ip* could correspond to either *dip*, *hip*, *kip*, or *lip*, which made correction impossible.

Results and Discussion

Figures 3 and 4 represent the results of Experiments 2A and 2B, respectively. Visual inspection of trial-by-trial data indicates that shifts in phonological categories did not consistently lead to release from PI compared to the data of semantic categories. Performance across PI levels was examined using formal analyses.

Figure 3

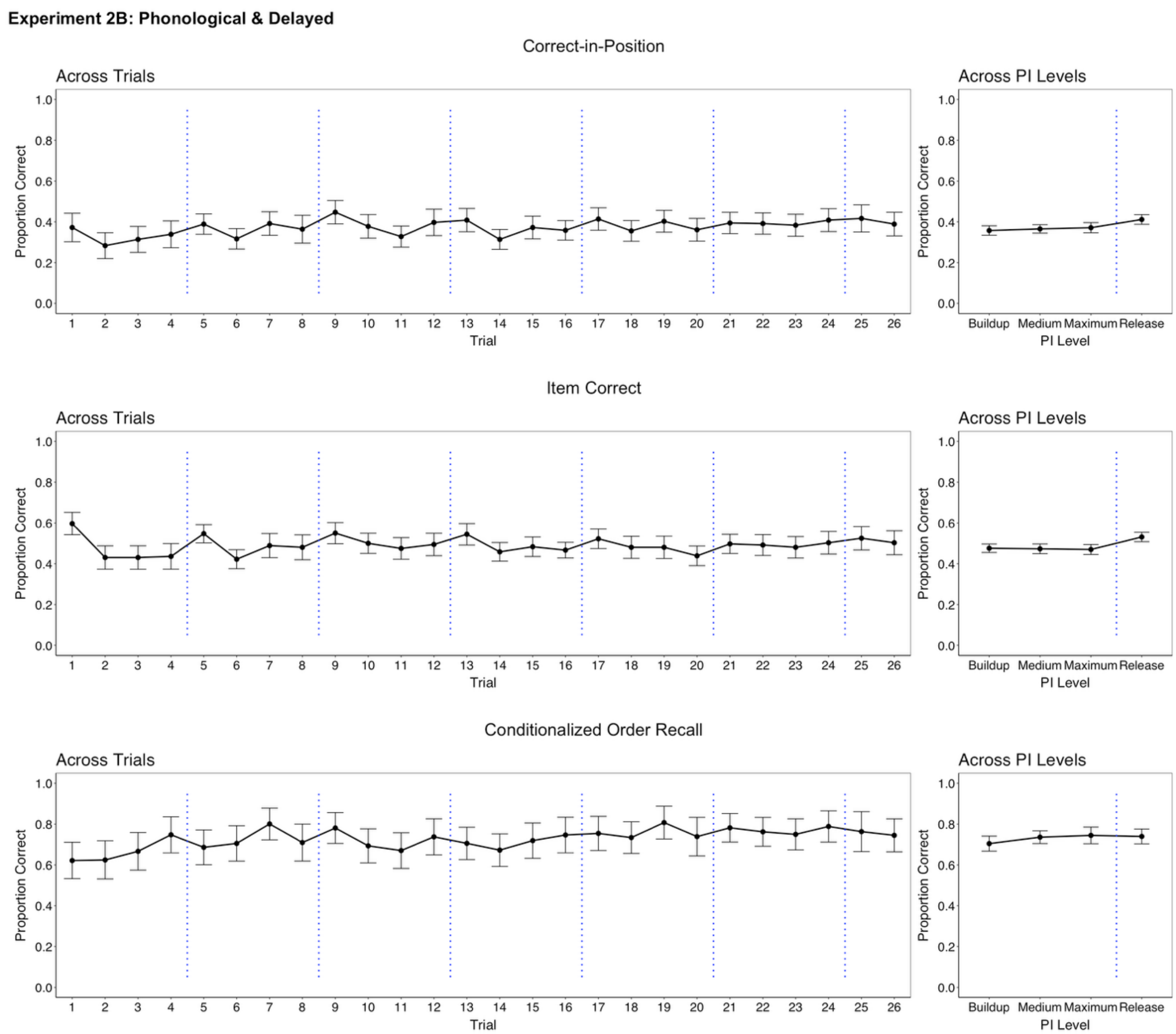
Results of Experiment 2A



Note. Error bars represent 95% confidence intervals for a within-participants factor (Morey, 2008).

Figure 4

Results of Experiment 2B



Note. Error bars represent 95% confidence intervals for a within-participants factor (Morey, 2008).

Experiment 2A

Correct-in-Position. For immediate recall (Figure 3), correct-in-position scores showed neither main effect of PI level, $BF_{01} = 19.12$, $\eta_p^2 = 0.013$, nor the difference between PI maximum versus PI release, $BF_{01} = 5.24$, $dz = 0.10$. Performance was equivalent among PI buildup ($M = 0.45$, $SD = 0.09$), PI medium ($M = 0.43$, $SD = 0.10$), PI maximum ($M = 0.46$, $SD = 0.09$), and PI release ($M = 0.45$, $SD = 0.09$). Other contrasts–buildup versus medium, medium versus maximum, and release versus buildup–indicated, if anything, evidence against a difference ($2.36 < BF_{01} < 6.91$).

Item Correct. Item correct scores exhibited anecdotal evidence against a difference among the four PI levels, $BF_{01} = 2.06$, $\eta_p^2 = 0.042$ (PI buildup: $M = 0.58$, $SD = 0.09$; PI medium: $M = 0.56$, $SD = 0.08$; PI maximum: $M = 0.57$; $SD = 0.07$; PI release: $M = 0.60$, $SD = 0.07$). Evidence that PI release was better than PI maximum ($BF_{10} = 2.94$, $dz = 0.33$) remained anecdotal. Other contrasts suggested, if anything, evidence against a difference ($2.86 < BF_{01} < 6.63$).

Conditionalized Order Recall. No clear pattern was observed in the conditionalized order recall data: PI buildup ($M = 0.73$, $SD = 0.11$), PI medium ($M = 0.73$, $SD = 0.13$), PI maximum ($M = 0.77$, $SD = 0.12$), and PI release ($M = 0.72$, $SD = 0.13$). There was no credible evidence for the main effect of PI level ($BF_{01} = 2.47$, $\eta_p^2 = 0.039$), the critical release-versus-maximum comparison ($BF_{10} = 1.67$, $dz = 0.30$), and the other comparisons ($1.05 < BF_{01} < 7.08$).

Experiment 2B

Correct-in-Position. In delayed recall, correct-in-position scores showed moderate evidence for a difference among the four PI levels, $BF_{10} = 4.45$, $\eta_p^2 = 0.069$ (PI buildup: $M = 0.36$, $SD = 0.09$; PI medium: $M = 0.37$, $SD = 0.08$; PI maximum: $M = 0.37$, $SD = 0.10$; PI release: $M = 0.41$, $SD = 0.09$). Evidence for the maximum versus release difference, however, remained anecdotal, $BF_{10} = 2.09$, $dz = 0.31$. Performance of PI buildup was equivalent to that of PI medium

(BF_{01} = 6.08), which was equivalent to that of PI maximum (BF_{01} = 6.64). Unexpectedly, performance was lower on PI buildup than on PI release (BF_{10} = 12.27, *dz* = 0.41).

Item Correct. For item correct in delayed recall, we found very strong evidence for the PI effect, both in the main effect of the four PI levels (BF_{10} = 57.02, η_p^2 = 0.099) and the critical contrast of PI release versus PI maximum (BF_{10} = 45.48, *dz* = 0.47). Performance across PI levels was as follows: PI buildup (*M* = 0.48, *SD* = 0.08), PI medium (*M* = 0.47, *SD* = 0.09), PI maximum (*M* = 0.47, *SD* = 0.09), and PI release (*M* = 0.53, *SD* = 0.09). PI buildup was equivalent to PI medium (BF_{01} = 6.97), which was equivalent to PI maximum (BF_{01} = 6.96). PI buildup was, however, lower than PI release (BF_{10} = 48.30, *dz* = 0.47). Notably, performance was lower for PI buildup than for PI release in both correct-in-position and item correct. Although the primary focus was the maximum–release contrast and no a priori hypothesis was formulated for the buildup–release contrast, these results suggest that PI may emerge rapidly from the first to the second trials of a block.

Conditionalized Order Recall. Consistent with the results of other experiments, no clear pattern was observed in order memory (PI buildup: *M* = 0.70, *SD* = 0.14; PI medium: *M* = 0.74, *SD* = 0.12; PI maximum: *M* = 0.74, *SD* = 0.16; PI release: *M* = 0.74, *SD* = 0.14). Evidence argues against a difference among the four PI levels (BF_{01} = 14.43, η_p^2 = 0.017) and against the critical difference between PI release and PI maximum (BF_{01} = 6.96, *dz* = 0.02). Anecdotal to moderate evidence against the other contrasts was also observed ($2.62 < BF_{01} < 6.71$).

Experiment 2 provided evidence for phonology-based PI in item memory, although its strength differed between recall tasks. In delayed recall (Experiment 2B), the item-correct scores provided very strong evidence for PI, whereas the corresponding evidence in immediate recall (Experiment 2A) was only anecdotal. This pattern suggests that phonology-based PI may not be as pronounced as semantics-based PI. Semantics-based PI was clearly and consistently observed in both immediate and delayed recall.

However, the four experiments left two methodological issues unresolved. First, Experiments 1 and 2 differed in both their mode of administration and participant population. Experiment 1 was conducted in the laboratory with university students, whereas Experiment 2 was conducted online with workers recruited through Prolific. Although Experiment 1 provided clear evidence for semantic PI, it remained possible that this effect depended on laboratory-based testing, the university student sample, or both. This procedural difference also complicated the comparison of the four experiments.

Second, the delayed-recall tasks in Experiments 1 and 2 used only one type of distractor task–parity judgement. It therefore remained unclear whether PI would also be observed with other distractor tasks, potentially limiting the generalizability of the findings. Moreover, the comparison between immediate and delayed recall performance suggested that the parity judgement task was relatively easy. Focusing on correct-in-position performance, the difference between immediate and delayed recall was approximately 6 percentage points in Experiment 1 ($M = 0.52$ vs. $M = 0.46$) and 7 percentage points in Experiment 2 ($M = 0.45$ vs. $M = 0.38$). Despite procedural differences, such as list length, Brown-Peterson studies have reported larger detrimental effects of distractor-filled retention intervals (Murdock, 1961; Peterson & Peterson, 1959). For instance, Murdock (1961) showed a 70-percentage-point decline in recall for word triads, from .93 following a 0-sec delay (no distractor) to .23 following an 18-sec distractor-filled delay.[3] The next follow-up experiment addressed both issues.

[3] List length may be a critical procedural difference: the present study used six-word lists, as is typical in immediate serial recall tasks, whereas Murdock (1961) used word triads, consistent with the standard Brown–Peterson procedure. We return to this issue in the General Discussion.

Experiment 3

Experiment 3 followed up on Experiment 1B to test semantic PI in delayed recall using web-based administration and a new distractor task.

Method

Participants

A new sample of 60 participants was recruited through Prolific with the previous recruitment criteria (see Experiment 2). Forty-one participants self-identified as women, and 17 as men. One participant selected “other,” and another selected “prefer not to report”. Their ages ranged from 18 to 30 years (M = 25.12, SD = 3.43). They were compensated £2.20 for their participation. Experiment 3 took approximately 25 minutes.

Materials, Procedure, and Scoring

The materials, procedure, and scoring were identical to those of Experiment 1B, except for the following two changes. First, Experiment 3 was administered online, as was Experiment 2. Second, the parity judgement task was replaced with a math verification task. During the 18-sec delay, participants verified as many mathematical equations as possible. Each equation took the form $a \times b + c = d$, where a, b, c were integers randomly selected from 1 to 9. For correct equations, d was equal to $a \times b + c$; for incorrect equations, d differed from the correct answer by ± 1. This task involved mental arithmetic and was therefore expected to impose greater cognitive demands than the previous parity judgement task.

Results and Discussion

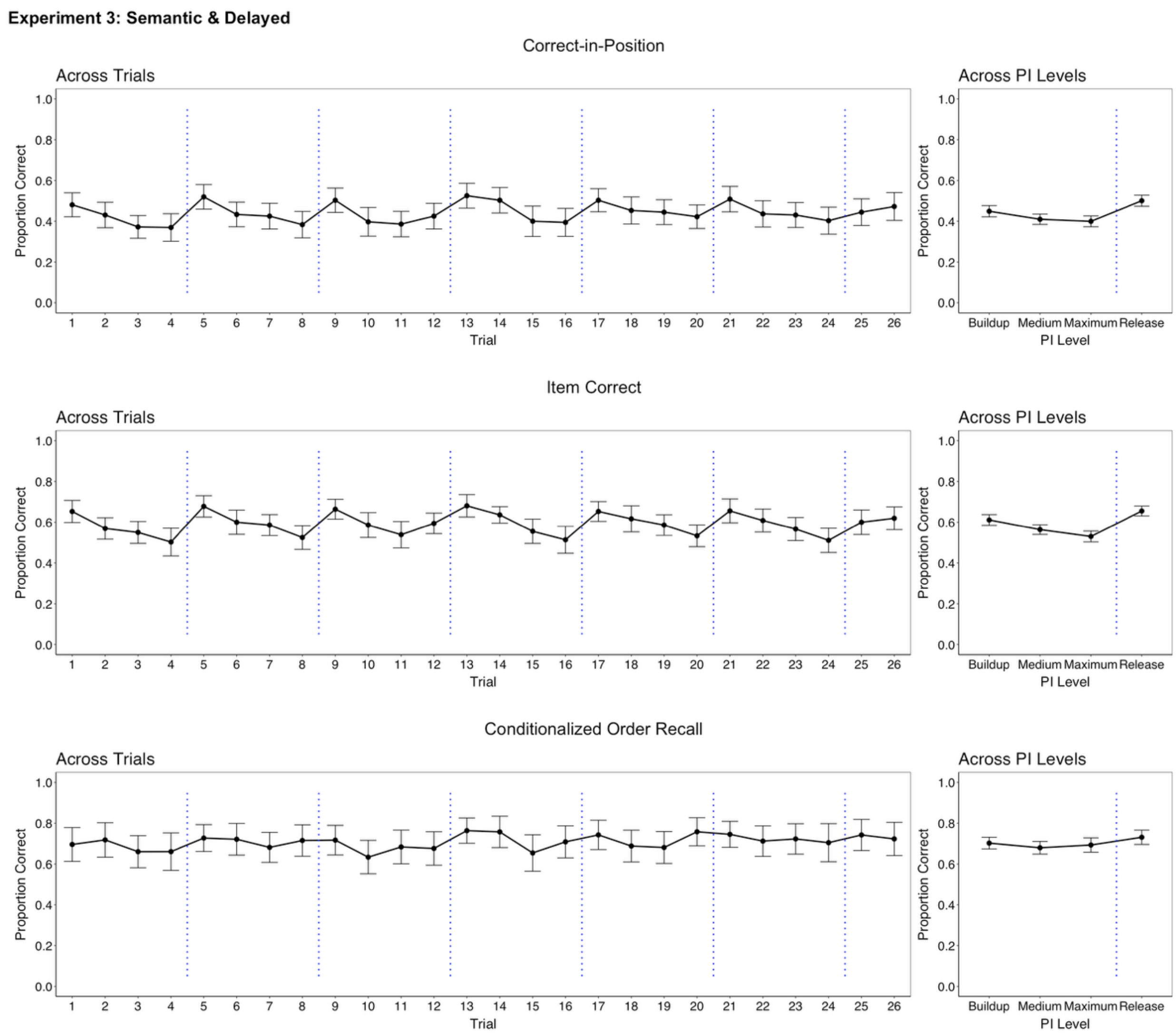
Before conducting the main analyses, we compared correct-in-position performance in Experiment 3 with that in Experiment 1B. Recall performance was numerically lower in

Experiment 3 with the math verification task (M = 0.44, SD = 0.22) than in Experiment 1B with the parity judgement task (M = 0.46, SD = 0.18), but the Bayes factor provided moderate evidence in favor of no difference (BF_{01} = 4.76, d = 0.08). Thus, contrary to our expectation, the presumably more demanding math verification task did not produce an additional impairment in recall performance relative to the parity judgement task. Nonetheless, Experiment 3 extended Experiment 1B by using a different distractor task, thereby providing valuable evidence regarding the generalizability of the findings.

Figure 5 shows the main results of Experiment 3. The overall pattern, including the scalloped pattern in the trial-by-trial data in item memory, was similar to that observed in Experiment 1B despite the methodological differences between the experiments. Formal analyses of PI levels are reported below.

Figure 5

Results of Experiment 3



Note. Error bars represent 95% confidence intervals for a within-participants factor (Morey, 2008).

Correct-in-Position

For correct-in-position performance (top-right panel of Figure 5), Experiment 3 replicated two key findings from Experiment 1B: an overall effect of PI level and a difference between PI release and PI maximum. A Bayesian ANOVA revealed a main effect of PI level, $BF_{10} = 3.38 \times 10^4$, $\eta_p^2 = 0.168$ (PI buildup: *M* = 0.45, *SD* = 0.11; PI medium: *M* = 0.41, *SD* = 0.10; PI maximum: *M* = 0.40, *SD* = 0.10; PI release: *M* = 0.50, *SD* = 0.10). A Bayesian paired *t*-test showed that performance was higher for PI release than for PI maximum, $BF_{10} = 6.82 \times 10^3$, *dz* = 0.67. The other contrasts provided anecdotal evidence for a difference between PI buildup and PI medium, $BF_{10} = 1.04$, *dz* = 0.27, moderate evidence against a difference between PI medium and PI maximum, $BF_{01} = 6.04$, *dz* = 0.07, and moderate evidence for a difference between PI release and PI buildup, $BF_{10} = 3.69$, *dz* = 0.35.

Item Correct

Turning to the item memory measure (middle-right panel of Figure 5), Experiment 3 also replicated the key pattern observed in Experiment 1B. A Bayesian ANOVA revealed a main effect of PI level, $BF_{10} = 6.87 \times 10^7$, $\eta_p^2 = 0.241$ (PI buildup: *M* = 0.61, *SD* = 0.10; PI medium: *M* = 0.56, *SD* = 0.09; PI maximum: *M* = 0.53, *SD* = 0.10; PI release: *M* = 0.66, *SD* = 0.09). A Bayesian *t*-test provided extreme evidence that performance was higher at PI release than at PI maximum, $BF_{10} = 4.41 \times 10^6$, *dz* = 0.90. Among the remaining contrasts, PI buildup outperformed PI medium, $BF_{10} = 4.24$, *dz* = 0.35. Evidence for the difference between PI medium and PI maximum, $BF_{01} = 1.22$, *dz* = 0.25, and that between PI release and PI buildup, $BF_{10} = 2.63$, *dz* = 0.33, was not clear.

Conditionalized Order Recall

In line with Experiment 1B, there was no clear evidence for PI in order memory (bottom-right panel of Figure 5). A Bayesian ANOVA did not support a main effect of PI level, BF_{01} = 5.36, η_p^2 = 0.03 (PI buildup: *M* = 0.70, *SD* = 0.11; PI medium: *M* = 0.68, *SD* = 0.12; PI maximum: *M* = 0.69, *SD* = 0.14; PI release: *M* = 0.73, *SD* = 0.14). A Bayesian *t*-test provided anecdotal evidence against the difference between PI release and PI maximum, BF_{01} = 2.62, *dz* = 0.19. The other contrasts, if anything, suggested equivalent performance ($2.93 < BF_{01} < 5.92$).

Experiment 3 replicated the findings of Experiment 1B across all three scoring measures. For item-correct performance, the measure of primary interest, the effect sizes for both the overall effect of PI and the difference between PI release and PI maximum were nearly identical across the two experiments: η_p^2 = 0.243 and *dz* = 0.89 in Experiment 1B, compared with η_p^2 = 0.241 and *dz* = 0.90 in Experiment 3. These findings would indicate that neither the presence nor the magnitude of the PI effect was specific to the mode of administration (i.e., lab-based vs. web-based), participant sample (i.e., university students vs. participants on Prolific), or the distractor task (i.e., parity judgement vs. math verification). Replication across these procedural variations supports the generalizability of the semantic PI effect.

Across the five separate experiments, we investigated the PI effect. It is also desirable to facilitate comparisons across experiments and synthesize the findings. To this end, we report effect sizes and their credible intervals in the final analysis.

Cross-Experiment Comparison

We indexed the PI effect as *performance on PI release trials minus performance on PI maximum trials*, following the critical contrast used in the previous *t*-tests. A positive value indicates a PI effect, reflecting better performance on PI release trials than on PI maximum trials. Using the *ttestBF* function with its default JZS prior, we obtained 100,000 posterior samples of the standardized effect size, δ. Based on these samples, we report the median δ and its 95% credible interval.

In previous analyses, we reported BFs, which quantify the relative evidence for the null and alternative hypotheses. Thus, even when the estimated effect size was numerically different from zero, evidence for the alternative hypothesis could remain weak if the effect was small. By contrast, the present cross-experiment comparison focuses on estimation rather than hypothesis comparison, providing the magnitude of the effect and its uncertainty conditional on the alternative hypothesis (van Doorn et al., 2021).

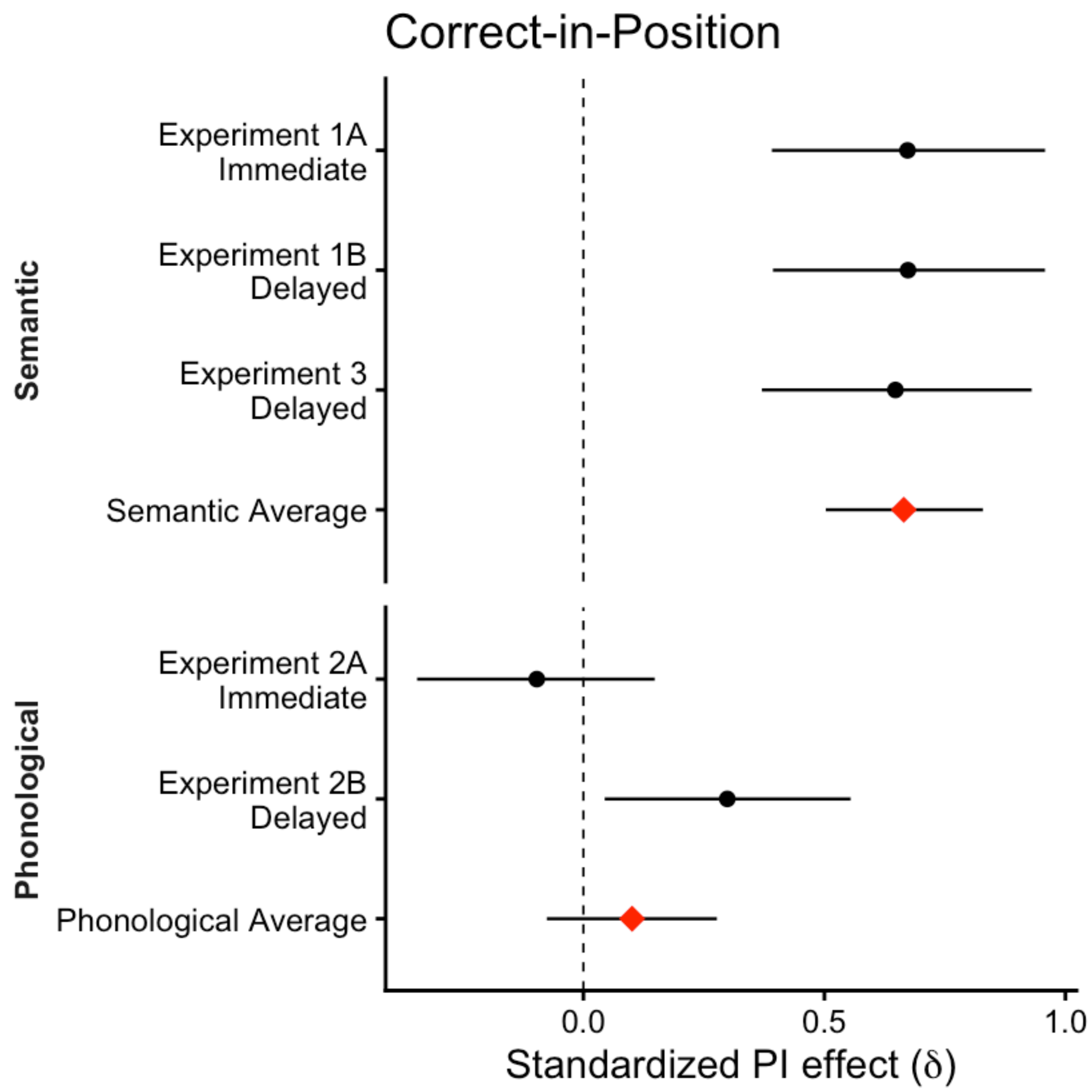
Because a separate sample of participants was recruited for each experiment, comparisons across experiments should be interpreted with caution. That said, each experiment included approximately 60 participants, which should help improve the precision of the effect size estimates.

Correct-in-Position

Figure 6 presents the effect sizes and credible intervals across the five experiments. Semantics-based PI was apparent in Experiment 1A, median δ = 0.67, 95% CrI [0.39, 0.96]; Experiment 1B, median δ = 0.67, 95% CrI [0.39, 0.96]; and Experiment 3, median δ = 0.65, 95% CrI [0.37, 0.93]. By contrast, the effects of phonology-based PI were smaller or in the opposite direction: Experiment 2A, median δ = -0.10, 95% CrI [-0.35, 0.15]; Experiment 2B, median δ = 0.30, 95% CrI [0.04, 0.55]. To further compare semantics-based and phonology-based PI, we also calculated the average effects across Experiments 1A, 1B, and 3 and across Experiments 2A and 2B, respectively. The average effect was numerically larger for semantics-based PI (δ = 0.66, 95% CrI [0.50, 0.83]) than for phonology-based PI (δ = 0.10, 95% CrI [-0.08, 0.28]).

Figure 6

Effect Sizes and 95% Credible Intervals for Correct-in-Position



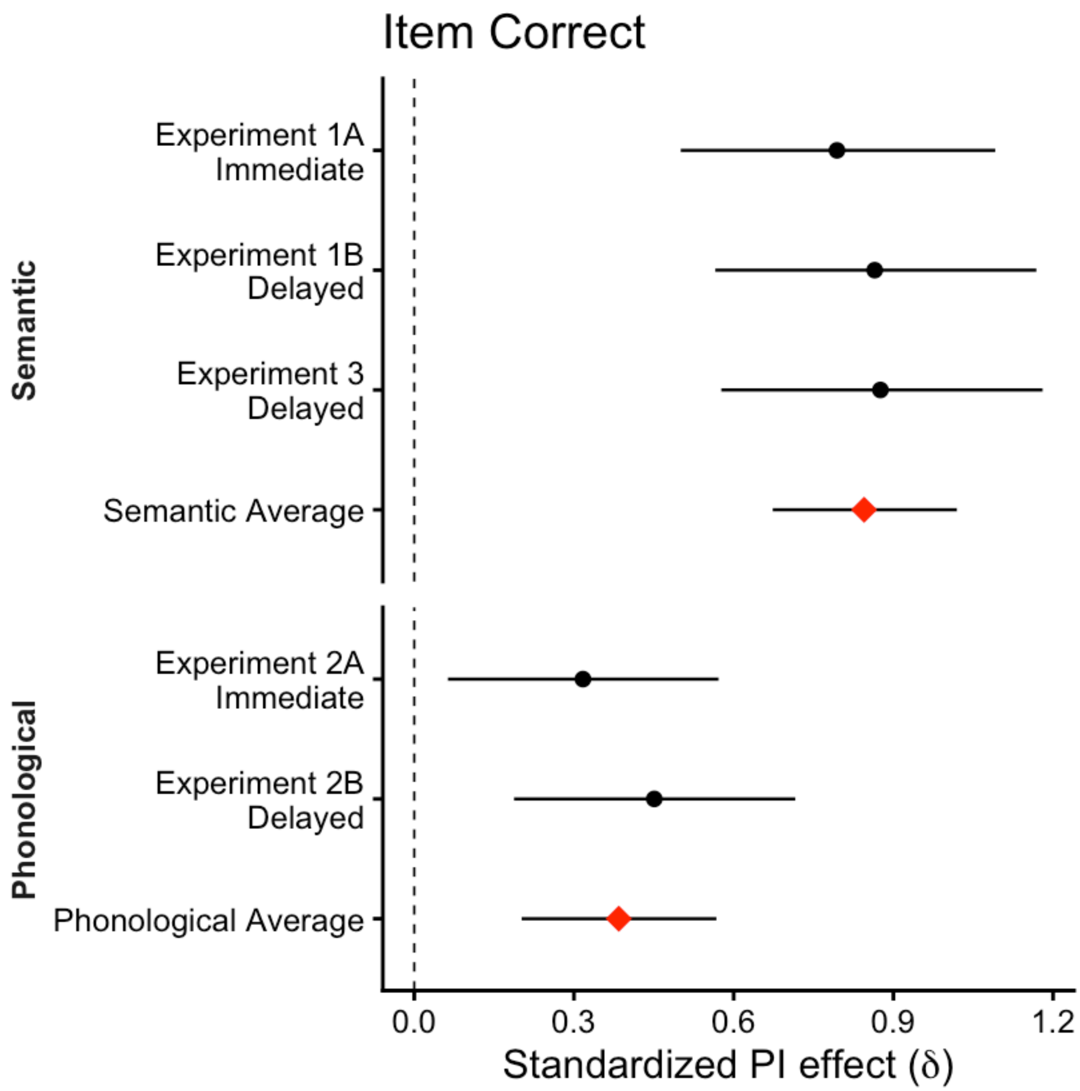
Note. The effect refers to the difference in performance between PI release and PI maximum trials.

Item Correct

Compared with the correct-in-position results, the effects measured by item correct were more pronounced (Figure 7): Experiment 1A, median δ = 0.79, 95% CrI [0.50, 1.09]; Experiment 1B, median δ = 0.86, 95% CrI [0.57, 1.17]; Experiment 3, median δ = 0.88, 95% CrI [0.58, 1.18]; Experiment 2A, median δ = 0.32, 95% CrI [0.06, 0.57]; and Experiment 2B, median δ = 0.45, 95% CrI [0.19, 0.72]. The average effect was again numerically larger for the semantics-based PI (δ = 0.85, 95% CrI [0.67, 1.02]) than for the phonology-based PI (δ = 0.38, 95% CrI [0.20, 0.57]).

Figure 7

Effect Sizes and 95% Credible Intervals for Item Correct



Note. The effect refers to the difference in performance between PI release and PI maximum trials.

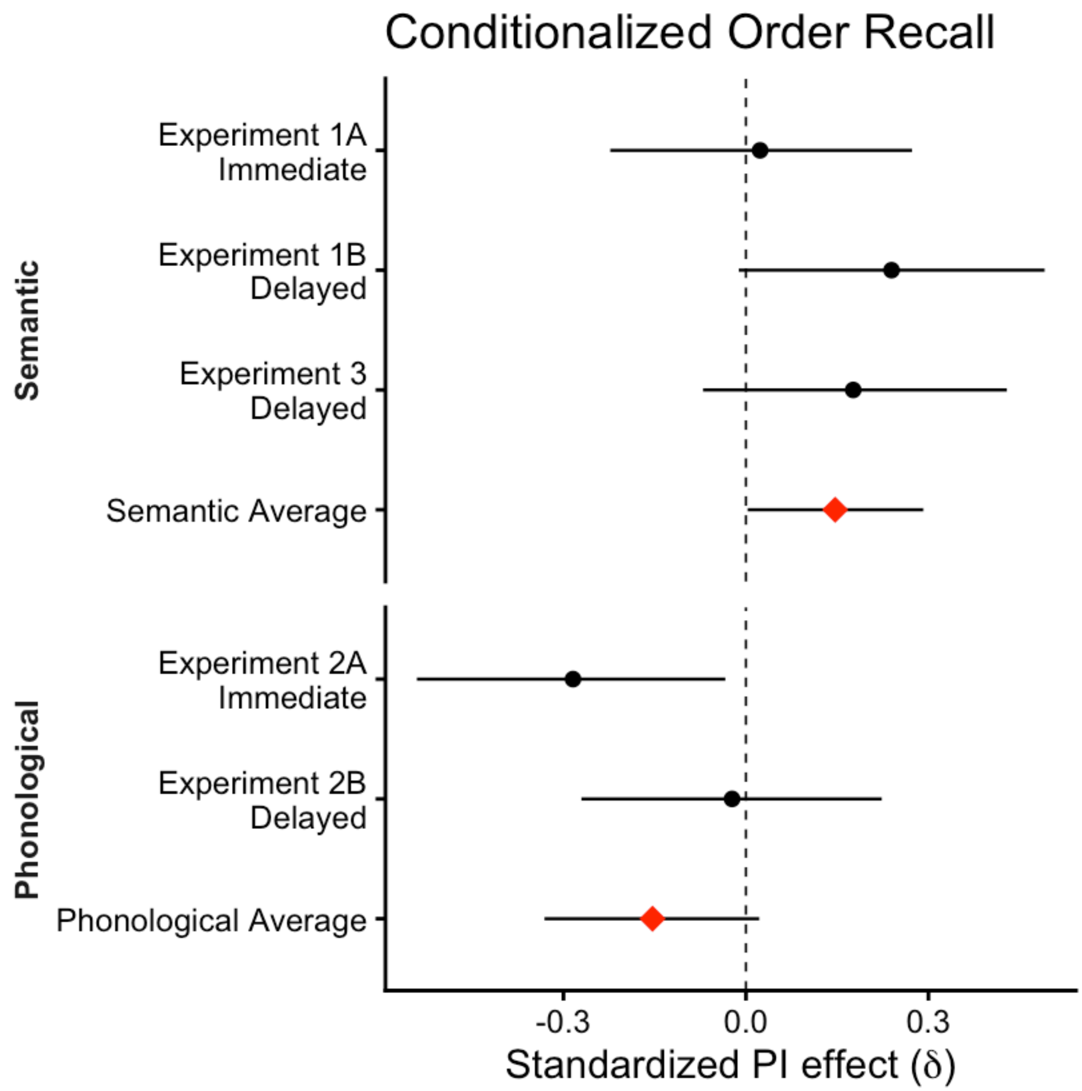
Conditionalized Order Recall

Consistent with separate analyses, the PI effect was less clear for order memory (Figure 8): Experiment 1A, median δ = 0.02, 95% CrI [-0.22, 0.27]; Experiment 1B, median δ = 0.24, 95% CrI [-0.01, 0.49]; Experiment 3, median δ = 0.18, 95% CrI [-0.07, 0.43]; Experiment 2A, median δ = -0.28, 95% CrI [-0.54, -0.03]; and Experiment 2B, median δ = -0.02, 95% CrI [-0.27, 0.22]. The average effect was positive for semantics-based PI (δ = 0.15, 95% CrI [0.003, 0.29]) but negative for phonology-based PI (δ = -0.15, 95% CrI [-0.33, 0.02]).

The cross-experiment comparison was based on the data from each experiment and complemented the individual-experiment results by providing estimates of the average effects of semantics- and phonology-based PI across experiments. The magnitude of PI was numerically larger for semantics-based PI than for phonology-based PI in both correct-in-position and item correct. With regard to the potential difference between correct-in-position and item correct, the magnitude of semantics-based PI was numerically larger in item correct (δ = 0.85) than in correct-in-position (δ = 0.66). Although a direct comparison between the two measures may be affected by differences in measurement properties, such as variability and measurement error, this pattern is consistent with our item-order hypothesis.

Figure 8

Effect Sizes and 95% Credible Intervals for Conditionalized Order Recall



Note. The effect refers to the difference in performance between PI release and PI maximum trials.

General Discussion

PI has garnered considerable attention in WM research. On the one hand, some theories assume that WM is susceptible to PI and that resolving PI is one of its core functions (Engle, 2002; Jonides & Nee, 2006). On the other hand, WM has also been conceptualized as a PI-free system (Oberauer & Awh, 2022; Oberauer et al., 2017). Despite their sharp contrast, these theoretical positions collectively highlight PI as a critical issue in theories of WM.

In addition to this theoretical divergence, empirical evidence for PI in WM has been rather mixed (Baddeley et al., 2018; Emery et al., 2008; Oberauer & Bartsch, 2023). Against this background, the current study focused on memory type as a potential factor determining when PI emerges and proposed the item–order hypothesis. This hypothesis predicted that PI would be more readily detected in item memory than in order memory. To provide a clear test of this hypothesis, we also systematically varied two other factors: category type (semantic vs. phonological) and recall type (immediate vs. delayed recall).

Experiment 1 demonstrated clear semantic PI in item memory under both immediate and delayed recall. The semantic PI effect in delayed recall was also replicated in a follow-up experiment (Experiment 3). In contrast, for phonological categories in Experiment 2, clear evidence was obtained for item memory in delayed recall, but not in immediate recall. Across these experiments, order memory appeared insensitive to PI. Thus, the item-order hypothesis was generally supported, although phonological PI in item memory was not reliably observed under immediate recall.

To summarize, our results suggest that (1) PI is more strongly associated with item memory than with order memory; (2) semantics-based PI is robust; (3) phonology-based PI can be observed under certain experimental conditions; and (4) semantics-based PI is more consistently observed across experiments than phonology-based PI. A major contribution of the present study was to propose and provide support for the item–order hypothesis, thereby refining our understanding of PI in WM. This study also identified category type as an additional factor that may influence the magnitude of PI. We suggest that the item-order distinction and category type provide useful perspectives on when and how PI emerges in WM.

Brown-Peterson and Serial Recall Tasks

The present study bears notable similarities to classic PI studies using the Brown–Peterson task, as both manipulated stimulus categories to examine PI. These Brown-Peterson studies documented the buildup of PI within categories and its release following category shifts across semantic categories (e.g., Wickens et al., 1976) and phonological categories (Baldwin & Wickens, 1974; Turvey & Egan, 1970).

In a typical trial of the Brown-Peterson task, a single item (e.g., a digit trigram or word triad) is presented, followed by an irrelevant distractor for several seconds (e.g., 18 sec, Wickens et al., 1976). After the distractor activity, participants are asked to recall the item. Thus, the PI effects observed in the Brown-Peterson task are consistent with the present study's findings under delayed recall.

However, the Brown-Peterson studies have documented larger PI effects than those observed in the present study. For example, Goggin and Wickens (1971) reported an approximately 23-percentage-points increase in recall performance following a semantic category shift, indicating substantial release from PI (see their Figure 1). By contrast, in the present study, item-correct performance differed by only 13 percentage points between the PI-maximum and PI-release conditions in both Experiments 1B and 3. We suggest that the item-order distinction may help account for these differences in the magnitudes of the effects.

Compared with the present study's task, the classic Brown-Peterson task may place fewer demands on order memory, thereby increasing the relative contribution of item memory to performance (for trade-offs between item and order processing, see Guitard et al., 2022). Because the present findings indicate that PI is more strongly associated with item memory than with order memory, a greater contribution of item memory may result in a larger observed PI effect. This difference in memory demands may arise partly because classic Brown-Peterson studies typically used shorter lists. For example, with word-triad lists, if participants remember the first and last words and their positions (e.g., the primacy and recency effects), the remaining word can be inferred to occupy the middle position. Such a list structure may therefore reduce the demands placed on order memory. Conversely, longer lists may place greater demands on order memory.

Although item memory was assessed using a separate measure in the present study, the overall task demands may still have influenced how participants encoded the lists. Specifically, the use of longer lists may have encouraged participants to devote greater attention to order information, and the requirement to recall items in serial order may have further strengthened this tendency (i.e., serial recall).

We acknowledge that this interpretation remains speculative and extends the item-order hypothesis beyond the present findings. This interpretation can, however, be tested in future research by manipulating list length and recall format. Specifically, shorter lists and free recall, relative to longer lists and serial recall, are predicted to produce larger PI effects. This empirically testable prediction underscores the potential value of the item–order hypothesis.

Relationship Between WM and LTM

We suggest that the immediate and delayed serial recall tasks used in the present study are consistent with the generic definition of WM (Cowan, 2017). Nevertheless, this definition is deliberately operational, as it “cautiously avoids a statement about the mechanism or about functions other than retention of information (i.e., memory)” (Cowan, 2017, p. 1163). Accordingly, the PI effects observed in the present study may reflect mechanisms associated with LTM rather than mechanisms intrinsic to WM. Given the distinction between WM and LTM, this possibility complicates the interpretation of these effects as evidence that WM itself is susceptible to PI.

Importantly, however, distinguishing WM from LTM does not preclude interactions between the two systems. For example, Cowan’s elaboration of the generic definition allows the activated portion of LTM to be considered part of WM without treating WM and LTM as identical constructs (Cowan, 2017; for related accounts, see also Ericsson & Kintsch, 1995; Unsworth & Engle, 2007). Similarly, Baddeley et al. (2019) regarded WM and LTM as separate systems but emphasized the importance of understanding how they interact. Consistent with this view, the *episodic buffer* in Baddeley and colleagues’ multicomponent model has been conceptualized as an interface between WM and LTM (Baddeley, 2000, 2012; Baddeley et al., 2021). This component provides access to long-term knowledge and makes that information available within WM.

It is therefore reasonable to assume that LTM contributed to WM performance in the present study. During the tasks, participants likely drew on pre-existing semantic and phonological knowledge stored in LTM. Had they lacked the relevant semantic knowledge (e.g., the color category: blue, red, green) or phonological knowledge (e.g., the AYD category: bide, dyed, sighed), the corresponding PI effects would not have emerged. We therefore interpret our findings as indicating that performance on WM tasks is susceptible to PI, presumably through interactions between WM and LTM.

Semantic and Phonological Representations

Although participants likely relied on semantic and phonological knowledge, the contributions of the two forms of knowledge appear to differ. Across the individual experiments, semantic categories produced more reliable PI effects than phonological categories. This difference is unlikely to reflect a weak phonological manipulation. Short words were selected to maximize within-category phonological similarity (e.g., dip, hip, kip), whereas categories were chosen to minimize between-category phonological similarity (e.g., dip, hip, kip vs. cot, dot, got).

The difference between the semantic and phonological effects may reflect representational differences. Several memory models adopt vector representations for semantic and phonological information under a single architecture (Guitard et al., 2025a; Oberauer et al., 2012; Reid et al., 2026). For instance, in a computational implementation of the eCFM, encoding involves concatenating vectors representing serial position (e.g., 300 features) and item information (e.g., 300 features) (Guitard et al., 2025a). Critically, the item features can be semantic and phonological while the resulting vectors operate under the same mechanisms.

The number of features allocated to different representation types is a key consideration in such models. According to the attentional subsetting theory (Caplan & Guitard, 2024a, 2024b), phonological information is represented by relatively few features, forming a dense representational space, whereas semantic information is represented by a larger number of features, forming a sparse space. Moreover, due to its density, phonological information is thought to be more prone to introduce similarity-based confusions than semantic information.

Intuitively, the notion of few phonological features and many semantic features seems plausible: Phonology relies on a limited number of phonemes within a language, while semantics may include a variety of features (e.g., ad hoc categories, Ishiguro et al., 2025). Furthermore, practical conventions in dimensionality reduction techniques support this distinction. Semantic vectors are often reduced to a larger number of features (e.g., 300 features, Landauer & Dumais, 1997) while phonological ones can be reduced to a smaller number (e.g., 50 features, Parrish, 2017).

Such representation differences would account for the present findings. Specifically, we assume that participants may down-weight their reliance on a particular type of information (e.g., phonological or semantic information) when confusion along that dimension reaches saturation, and that representation differences may determine how rapidly saturation occurs for different types of information. Due to a small number of phonological features, phonology-based confusions may arise quickly, immediately after seeing six items in the first trial of a block (e.g., dip, hip, kip, lip, rip, sip).[4] Participants may therefore rapidly down-weight their reliance on phonological information. As a result, phonological PI may be overshadowed by the within-list phonological similarity effect. In contrast, semantics-based confusions would develop more

[4] If phonological similarity were purely harmful to performance, optimal encoding would disregard phonological information. However, phonological similarity can be beneficial to item memory (Fallon et al., 1999; Gupta et al., 2005; Nimmo & Roodenrys, 2004). Accordingly, we suggest that phonological information is encoded even in phonologically similar lists.

gradually and are therefore more detectable as PI. For instance, even when presented with six semantically similar items (e.g., dog, cat, horse, lion, bear, tiger), participants may distinguish them from each other by drawing on rich semantic features (e.g., a dog is loyal, but a cat is fickle).

Findings on within-list similarity effects provide indirect support for this account. For example, Salamé and Baddeley (1986) found a phonological similarity effect for lists of five to seven items, but the effect disappeared for eight-item lists. They interpreted this disappearance as indicating that phonological coding became ineffective for longer lists and was therefore abandoned in favor of alternative forms of coding (e.g., visual or semantic coding). Although this null effect cannot be taken as direct evidence that phonological confusion develops rapidly, the overall pattern is consistent with the possibility that phonologically similar items generate substantial confusion even within relatively short lists, beyond which phonological information ceases to provide an effective basis for recall. Accordingly, with the six-item lists used in the present study, there may have been little opportunity for further phonological confusion to accumulate across trials. By contrast, Dauphinee et al. (2026) reported a detrimental effect of semantic similarity on order memory even for ten-item lists. This finding suggests that semantic information continues to influence performance over longer lists and is therefore consistent with the possibility that semantic confusion remains sensitive to further accumulation across trials.

As semantic and phonological features are pre-experimentally learned, they are likely retrieved from LTM. Therefore, the reasoning based on representational differences is relevant to LTM and is consistent with the episodic buffer notion. It further suggests that differences in the number of features underlying semantic and phonological representations may account for the divergent effects observed for the two types of information.

Conclusions

A series of experiments revealed that PI was detected more reliably in item memory than in order memory and for semantic categories than for phonological categories. We suggest that the item-order distinction is an important factor in when and how PI emerges in WM. This distinction may aid in synthesizing previous findings, facilitating comparisons across different memory tasks, and generating testable predictions. We further argue that the difference between the semantic and phonological effects may be attributable to differences in the number of representational features, thereby shedding light on how long-term knowledge influences WM.

Declarations

Funding

This project was supported by Natural Sciences and Engineering Research Council of Canada via a grant to Jean Saint-Aubin (grant number RGPIN-2023-05943). It was conducted as a part of Sho Ishiguro's Overseas Research Fellowship funded by Japan Society for the Promotion of Science (grant number 202460272).

Conflicts of interest/Competing interests

The authors have no competing interests to declare.

Ethics approval

The current study was approved by the institutional ethics committee at Cardiff University (reference EC.22.09.20.6627G).

Consent to participate

Informed consent for participation was obtained from all individual participants in the study.

Consent for publication

Informed consent for publication of findings and anonymized data was obtained from all participants.

Open Practices Statement

Materials, anonymized data, and analysis scripts are available on the OSF. None of the experiments were preregistered.

Availability of data and materials

Anonymized data and materials are available on the OSF (https://osf.io/wtnxr).

Code availability

Analysis scripts are available on the OSF (https://osf.io/wtnxr).

Authors' contributions

Sho Ishiguro: Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **René-Pierre Sonier:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing. **Dominic Guitard:** Conceptualization, Methodology, Software, Investigation, Data curation, and Writing – review & editing. **Jean Saint-Aubin:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, and Funding acquisition.

References

Andersson, P., Schrooten, M. G. S., & Persson, J. (2025). Age differences in brain functional connectivity underlying proactive interference in working memory. *Human Brain Mapping*, *46*(5), e70189. https://doi.org/10.1002/hbm.70189

Baddeley, A. D. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/s1364-6613(00)01538-2

Baddeley, A. D. (2012). Working memory: theories, models, and controversies. *Annual Review of Psychology*, *63*(1), 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Baddeley, A. D., Hitch, G., & Allen, R. (2021). A multicomponent model of working memory. In R. Logie, V. Camos, & N. Cowan (Eds.), *Working Memory: The State of the Science* (pp. 10–43). Oxford University Press. https://doi.org/10.1093/oso/9780198842286.003.0002

Baddeley, A. D., Hitch, G. J., & Allen, R. J. (2019). From short-term store to multicomponent working memory: The role of the modal model. *Memory & Cognition*, *47*(4), 575–588. https://doi.org/10.3758/s13421-018-0878-5

Baddeley, A. D., Hitch, G. J., & Quinlan, P. T. (2018). Is the phonological similarity effect in working memory due to proactive interference? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *44*(8), 1312–1316. https://doi.org/10.1037/xlm0000509

Baddeley, A. D., & Warrington, E. K. (1970). Amnesia and the distinction between long- and short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *9*(2), 176–189. https://doi.org/10.1016/s0022-5371(70)80048-2

Baldwin, R. B., & Wickens, D. D. (1974). Release from PI and the physical aspects of words. *Bulletin of the Psychonomic Society*, *3*(4), 305–306. https://doi.org/10.3758/bf03333478

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/bf03193014

Bartsch, L. M., & Oberauer, K. (2023). The contribution of episodic long-term memory to working memory for bindings. *Cognition*, 231, 105330. https://doi.org/10.1016/j.cognition.2022.105330

Bäuml, K.-H. T., & Kliegl, O. (2013). The critical role of retrieval processes in release from proactive interference. *Journal of Memory and Language*, *68*(1), 39–53. https://doi.org/10.1016/j.jml.2012.07.006

Bunting, M. (2006). Proactive interference and item similarity in working memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *32*(2), 183–196. https://doi.org/10.1037/0278-7393.32.2.183

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), 104556. https://doi.org/10.1016/j.jml.2024.104556

Carroll, L. M., Jalbert, A., Penney, A. M., Neath, I., Surprenant, A. M., & Tehan, G. (2010). Evidence for proactive interference in the focus of attention of working memory. *Revue Canadienne de Psychologie Experimentale [Canadian Journal of Experimental Psychology]*, *64*(3), 208–214. https://doi.org/10.1037/a0021011

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review*, *24*(4), 1158–1170. https://doi.org/10.3758/s13423-016-1191-6

Cowan, N., Johnson, T. D., & Saults, J. S. (2005). Capacity limits in list item recognition: evidence from proactive interference. *Memory* , *13*(3–4), 293–299. https://doi.org/10.1080/09658210344000206

Cowan, N., Morey, C. C., & Naveh-Benjamin, M. (2020). An embedded-processes approach to working memory: How is it distinct from other approaches, and to what ends? In R. H. Logie, V. Camos, & N. Cowan (Eds.), *Working Memory* (pp. 44–84). Oxford University Press. https://doi.org/10.1093/oso/9780198842286.003.0003

Craik, F., & Birtwistle, J. (1971). Proactive inhibition in free recall. *Journal of Experimental Psychology*, *91*, 120–123. https://doi.org/10.1037/H0031835

Dauphinee, I., Guitard, D., Ishiguro, S., & Saint-Aubin, J. (2026). Semantic similarity disrupts order recall: Returning the black sheep of similarity effects to the flock. *Journal of Experimental Psychology. Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001622

Emery, L., Hale, S., & Myerson, J. (2008). Age differences in proactive interference, working memory, and abstract reasoning. *Psychology and Aging*, *23*(3), 634–645. https://doi.org/10.1037/a0012577

Endress, A. D., & Potter, M. C. (2014). Large capacity temporary visual memory. *Journal of Experimental Psychology. General*, *143*(2), 548–565. https://doi.org/10.1037/a0033934

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*(1), 19–23. https://doi.org/10.1111/1467-8721.00160

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*(2), 211–245. https://doi.org/10.1037/0033-295x.102.2.211

Fallon, A. B., Groves, K., & Tehan, G. (1999). Phonological similarity and trace degradation in the serial recall task: When CAT helps RAT, but not MAN. *International Journal of Psychology: Journal International de Psychologie*, *34*(5–6), 301–307. https://doi.org/10.1080/002075999399602

Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review*, *119*(2), 223–271. https://doi.org/10.1037/a0027371

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

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

Goggin, J., & Wickens, D. D. (1971). Proactive interference and language change in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *10*(4), 453–458. https://doi.org/10.1016/s0022-5371(71)80046-4

Guitard, D., Saint-Aubin, J., & Cowan, N. (2022). Tradeoffs between Item and Order Information in Short-Term Memory. *Journal of Memory and Language*, *122*. https://doi.org/10.1016/j.jml.2021.104300

Guitard, D., Saint-Aubin, J., Reid, J. N., & Jamieson, R. K. (2025a). An embedded computational framework of memory: The critical role of representations in veridical and false recall predictions. *Psychonomic Bulletin & Review*, 1–55. https://doi.org/10.3758/s13423-025-02669-7

Guitard, D., Saint-Aubin, J., Reid, N. J., & Jamieson, R. K. (2025b). 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), 104573. https://doi.org/10.1016/j.jml.2024.104573

Gupta, P., Lipinski, J., & Aktunc, E. (2005). Reexamining the phonological similarity effect in immediate serial recall: the roles of type of similarity, category cuing, and item recall. *Memory & Cognition*, *33*(6), 1001–1016. https://doi.org/10.3758/bf03193208

Halford, G. S., Maybery, M. T., & Bain, J. D. (1988). Set-size effects in primary memory: an age-related capacity limitation? *Memory & Cognition*, *16*(5), 480–487. https://doi.org/10.3758/bf03214229

Hamilton, A. C., & Martin, R. C. (2007). Proactive interference in a semantic short-term memory deficit: role of semantic and phonological relatedness. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *43*(1), 112–123. https://doi.org/10.1016/s0010-9452(08)70449-0

Hamilton, M., Ross, A., Blaser, E., & Kaldy, Z. (2022). Proactive interference and the development of working memory. *Wiley Interdisciplinary Reviews. Cognitive Science*, *13*(3), e1593. https://doi.org/10.1002/wcs.1593

Healy, A. F., & McNamara, D. S. (1996). Verbal learning and memory: does the modal model still work? *Annual Review of Psychology*, *47*(1), 143–172. https://doi.org/10.1146/annurev.psych.47.1.143

Hendrikx, A. J. P. (1986). Short-term proactive interference revisited. *Bulletin of the Psychonomic Society*, *24*(5), 358–360. https://doi.org/10.3758/bf03330151

Ishiguro, S., Guitard, D., & Saint-Aubin, J. (2025). Examining the semantic relatedness effect on working memory with ad hoc categories. *Memory & Cognition*, *53*(6), 1944–1962. https://doi.org/10.3758/s13421-025-01692-2

Ishiguro, S., Guitard, D., & Saint-Aubin, J. (2026). From stork to baby: Semantic relatedness can improve order memory without grouping. *Psychonomic Bulletin & Review*, *33*(1), 42. https://doi.org/10.3758/s13423-025-02840-0

Ishiguro, S., & Saito, S. (2024). The Semantic Similarity Effect on Short-Term Memory: Null Effects of Affectively Defined Semantic Similarity. *Journal of Cognition*, *7*(1), 24. https://doi.org/10.5334/joc.349

Jonides, J., & Nee, D. E. (2006). Brain mechanisms of proactive interference in working memory. *Neuroscience*, *139*(1), 181–193. https://doi.org/10.1016/j.neuroscience.2005.06.042

Kane, M. J., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *26*(2), 336–358. https://doi.org/10.1037/0278-7393.26.2.336

Kliegl, O., & Bäuml, K.-H. T. (2021). Buildup and release from proactive interference - Cognitive and neural mechanisms. *Neuroscience and Biobehavioral Reviews*, *120*, 264–278. https://doi.org/10.1016/j.neubiorev.2020.10.028

Kowialiewski, B., & Majerus, S. (2020). The varying nature of semantic effects in working memory. *Cognition*, *202*, 104278. https://doi.org/10.1016/j.cognition.2020.104278

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037//0033-295x.104.2.211

Levy, B. J., & Anderson, M. C. (2002). Inhibitory processes and the control of memory retrieval. *Trends in Cognitive Sciences*, *6*(7), 299–305. https://doi.org/10.1016/s1364-6613(02)01923-x

Loosli, S. V., Rahm, B., Unterrainer, J. M., Weiller, C., & Kaller, C. P. (2014). Developmental change in proactive interference across the life span: evidence from two working memory tasks. *Developmental Psychology*, *50*(4), 1060–1072. https://doi.org/10.1037/a0035231

Lustig, C., May, C. P., & Hasher, L. (2001). Working memory span and the role of proactive interference. *Journal of Experimental Psychology. General*, *130*(2), 199–207. https://doi.org/10.1037/0096-3445.130.2.199

Martin, R. C., Rapp, B., & Purcell, J. (2020). Domain-specific working memory: Perspectives from cognitive neuropsychology. In R. H. Logie (Ed.), *Working Memory* (Vol. 437, pp. 235–281). Oxford University Press. https://doi.org/10.1093/oso/9780198842286.003.0009

May, C. P., Hasher, L., & Kane, M. J. (1999). The role of interference in memory span. *Memory & Cognition*, *27*(5), 759–767. https://doi.org/10.3758/bf03198529

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

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2024, January 24). *BayesFactor: Computation of Bayes Factors for Common Designs*. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=BayesFactor

Murdock, B. B., Jr. (1961). The retention of individual items. *Journal of Experimental Psychology*, *62*(6), 618–625. https://doi.org/10.1037/h0043657

Nimmo, L. M., & Roodenrys, S. (2004). Investigating the phonological similarity effect: Syllable structure and the position of common phonemes. *Journal of Memory and Language*, *50*(3), 245–258. https://doi.org/10.1016/j.jml.2003.11.001

Nishiyama, S., O'Reilly, R. C., & Saito, S. (2025). Slowdown vs. breakdown of memory recall by retrieval stopping. *Memory & Cognition*, 1–13. https://doi.org/10.3758/s13421-025-01696-y

Nørby, S. (2015). Why forget? On the adaptive value of memory loss. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *10*(5), 551–578. https://doi.org/10.1177/1745691615596787

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

Oberauer, K. (2022). When does working memory get better with longer time? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 48(12), 1754–1774. https://doi.org/10.1037/xlm0001199

Oberauer, K., & Awh, E. (2022). Is there an activity-silent working memory? *Journal of Cognitive Neuroscience*, *34*(12), 2360–2374. https://doi.org/10.1162/jocn_a_01917

Oberauer, K., Awh, E., & Sutterer, D. W. (2017). The role of long-term memory in a test of visual working memory: Proactive facilitation but no proactive interference. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *43*(1), 1–22. https://doi.org/10.1037/xlm0000302

Oberauer, K., & Bartsch, L. M. (2023). When does episodic memory contribute to performance in tests of working memory? *Journal of Cognition*, *6*(1), 44. https://doi.org/10.5334/joc.311

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: an interference model of complex span. *Psychonomic Bulletin & Review*, *19*(5), 779–819. https://doi.org/10.3758/s13423-012-0272-4

Parrish, A. (2017). Poetic sound similarity vectors using phonetic features. *Proceedings. AAAI Artificial Intelligence and Interactive Digital Entertainment Conference*, *13*(2), 99–106. https://doi.org/10.1609/aiide.v13i2.12971

Peterson, L. R., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*(3), 193–198. https://doi.org/10.1037/h0049234

Ralph, A., Walters, J. N., Stevens, A., Fitzgerald, K. J., Tehan, G., Surprenant, A. M., Neath, I., & Turcotte, J. (2011). Immunity to proactive interference is not a property of the focus of attention in working memory. *Memory & Cognition*, *39*(2), 217–230. https://doi.org/10.3758/s13421-010-0030-7

Rapp, B., Purcell, J., Hillis, A. E., Capasso, R., & Miceli, G. (2016). Neural bases of orthographic long-term memory and working memory in dysgraphia. *Brain: A Journal of Neurology*, *139*(Pt 2), 588–604. https://doi.org/10.1093/brain/awv348

Reid, J. N., Guitard, D., & Jamieson, R. K. (2026). MINERVA OPS: A computational framework for the representation and recognition of orthographic, phonological, and semantic associates. *Psychonomic Bulletin & Review*, *33*(2). https://doi.org/10.3758/s13423-025-02792-5

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. (1999). The influence of long‐term memory factors on immediate serial recall: An item and order analysis. *International Journal of Psychology: Journal International de Psychologie*, *34*(5–6), 347–352. https://doi.org/10.1080/002075999399675

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

Salamé, P., & Baddeley, A. (1986). Phonological factors in STM: Similarity and the unattended speech effect. *Bulletin of the Psychonomic Society*, *24*(4), 263–265. https://doi.org/10.3758/bf03330135

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

Shipstead, Z., & Engle, R. W. (2013). Interference within the focus of attention: working memory tasks reflect more than temporary maintenance. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(1), 277–289. https://doi.org/10.1037/a0028467

Sonier, R.-P., Guitard, D., Melanson, E., Jamieson, R. K., & Saint-Aubin, J. (2025). Semantic similarity is not emotional: No effect of similarity defined by valence, arousal, and dominance on short-term ordered recall. *Memory & Cognition*, 1–17. https://doi.org/10.3758/s13421-024-01678-6

Stoet, G. (2010). PsyToolkit: a software package for programming psychological experiments using Linux. *Behavior Research Methods*, *42*(4), 1096–1104. https://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 (Columbia, Mo.)*, *44*(1), 24–31. https://doi.org/10.1177/0098628316677643

Thorn, A., & Page, M. (2008). *Interactions between short-term and long-term memory in the verbal domain*. Psychology Press.

Turvey, M. T., & Egan, J. F. (1970). Release from proactive interference in short-term memory as a function of change in visual and phonemic structure and retention interval. *Perception & Psychophysics*, *7*(3), 169–172. https://doi.org/10.3758/bf03208651

Underwood, B. J. (1957). Interference and forgetting. *Psychological Review*, *64*(1), 49–60. https://doi.org/10.1037/h0044616

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*(1), 104–132. https://doi.org/10.1037/0033-295X.114.1.104

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, Š., Ly, A., Marsman, M., Matzke, D., Komarlu Narendra Gupta, A. R., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.-J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, 28(3), 813–826. https://doi.org/10.3758/s13423-020-01798-5

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*(3), 289–335. https://doi.org/10.1016/j.jml.2003.10.003

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Warrington, E. K., & Shallice, T. (1969). The selective impairment of auditory verbal short-term memory. *Brain: A Journal of Neurology*, *92*(4), 885–896. https://doi.org/10.1093/brain/92.4.885

Whiteman, H. L., Nairne, J. S., & Serra, M. (1994). Recognition and recall-like processes in the long-term reconstruction of order. *Memory* , *2*(3), 275–294. https://doi.org/10.1080/09658219408258949

Wickens, D. D. (1970). Encoding categories of words: An empirical approach to meaning. *Psychological Review*, *77*(1), 1–15. https://doi.org/10.1037/h0028569

Wickens, D. D. (1973). Some characteristics of word encoding. *Memory & Cognition*, *1*(4), 485–490. https://doi.org/10.3758/BF03208913

Wickens, D. D., Dalezman, R. E., & Eggemeier, F. T. (1976). Multiple encoding of word attributes in memory. *Memory & Cognition*, *4*(3), 307–310. https://doi.org/10.3758/BF03213181

Zhang, J. X., Wu, R., Kong, L., Weng, X., & Du, Y. (2010). Electrophysiological correlates of proactive interference in the “Recent Probes” verbal working memory task. *Neuropsychologia*, *48*(7), 2167–2173. https://doi.org/10.1016/j.neuropsychologia.2010.04.008