Memory & Cognition (2019) 47:455–472
https://doi.org/10.3758/s13421-018-0883-8

Set size and long-term memory/lexical effects in immediate serial recall: Testing the impurity principle

Ian Neath[1] · Aimée M. Surprenant[1]

Published online: 7 December 2018

Abstract

The impurity principle (Surprenant & Neath, 2009b) states that because memory is fundamentally reconstructive, tasks and processes are not pure. This principle is based on a long line of research showing the effects of one memory system or process on another. Although the principle is widely accepted, many researchers appear hesitant to endorse it in extreme edge cases. One such case involves the effects of long-term memory and lexical factors when a small, closed set of items is used. According to this view, because the subject knows the set of items, there will be no effect of item information. In contrast, the impurity principle predicts that such effects can still be observed, because immediate serial recall with a small closed set of items is not a pure test of order information. Four experiments tested this edge case. In Experiments 1 and 2, we found concreteness effects when item uncertainty was minimized in both within-subjects (Exp. 1) and between-subjects (Exp. 2) designs. In Experiments 3 and 4, we found frequency effects when item uncertainty was minimized in both within-subjects (Exp. 3) and between-subjects (Exp. 4) designs. Analyses of intrusion and omission errors indicated that the sets of items had been learned. Analyses by experiment half also confirmed that the effects of concreteness and frequency were observable in the latter stages of the experiments, when there should have been even less doubt about the items. The results support the impurity principle and suggest that hesitation about accepting it in edge cases is unwarranted.

Keywords Short term memory · Set size effects · Serial recall · Working memory

Surprenant and Neath (2009b) proposed a number of principles that, they argued, summarized important empirical findings about human memory. One of these, the *impurity principle*, was based on the idea that because memory is a fundamentally reconstructive process, people will recruit and use a wide variety of information and processes to help them remember a particular item. Because of this, tasks are not pure: There is always contamination from multiple sources and multiple processes. This idea is not new; many theorists had previously made similar arguments (e.g., Crowder, 1993; Jacoby, 1991; Kolers & Roediger, 1984; Restle, 1974). As widely accepted as this idea is, edge cases still occur in which researchers seem diffident about wholeheartedly accepting the principle. The purpose of this article is to assess one such case. It is well established that long-term memory and lexical factors affect immediate serial recall when a large stimulus pool is used. In this article, we use four experiments to examine whether these factors still affect immediate serial recall when a small, closed pool of stimuli is used.

Immediate serial recall has featured as the principal way of assessing primary or short-term or immediate or working memory since the 19th century. Jacobs (1887, p. 75) introduced the term *span* as a measure of the ability "of temporarily retaining sounds long enough to reproduce them correctly". Subsequent studies quickly confirmed the influence of long-term memory and lexical factors on determining span (for an early review, see Blankenship, 1938; see also Crowder, 1976; Surprenant & Neath, 2009a, 2009b). Space precludes a listing of all such factors, but they include phonological neighborhood size (Roodenrys, Hulme, Lethbridge, Hinton, & Nimmo, 2002), orthographic neighborhood size (Jalbert, Neath, & Surprenant, 2011), semantic similarity (Saint-Aubin, Ouellette, & Poirier, 2005), pleasantness (Monnier & Syssau, 2008), word frequency (Roodenrys & Quinlan, 2000), and concreteness (Walker & Hulme, 1999). In this article, we focus on the latter two.

Despite the wide acknowledgement that these factors affect performance on immediate serial recall, a number of

✉ Ian Neath
ineath@mun.ca

[1] Memorial University of Newfoundland, St John's, Newfoundland, Canada

researchers appear hesitant to accept the impurity principle in all cases. One such case is when a small, closed pool of items is used. It is tempting to think that subjects will quickly learn the identity of the few possible items, and therefore item will no longer play a role. This apparent hesitation can be seen in a number of ways. Some theorists qualify their statements about the role of item information when considering possible differences between small closed sets and large open sets. Rather than saying that item information continues to affect serial recall, they appear to hesitate and allow for diminished or nonexistent effects. For example, Baddeley (2012, p. 8) noted that "studies that specifically attempt to investigate the [phonological] loop tend to minimize the need to retain item information by repeatedly using the same limited set, for example, consonants. Studies using open sets, for instance, different words for each sequence, are more likely to reflect loss of item information and to show semantic and other LTM-based effects." Similarly, Hughes and colleagues (Hughes, Chamberland, Tremblay, & Jones, 2016; Hughes, Marsh, & Jones, 2009) distinguished between what they termed *pure serial recall*, in which the same set of items from a closed pool is used on every trial, and *nonpure serial recall*, in which new items are used on every trial. In pure serial recall, "the burden falls entirely or primarily on reproducing item order rather than individual item identity" (Hughes et al., 2016, p. 127).

These theoretical statements can be interpreted as positing that when the set of to-be-remembered items are known, only order information is involved, and therefore long-term and lexical factors will not affect performance. For example, Osth and Dennis (2015, p. 1448) stated that "One of the motivations behind conducting studies that use closed sets is that memory for individual items quickly reaches ceiling, and only the order among the items has to be remembered." Similarly, Lin, Chen, Lai, and Wu (2015, p. 541) stated that "a closed set of Chinese characters were selected, since the previous research has shown that memory performance with an open set of stimuli in the immediate serial-recall task might be affected by representations in both WM and LTM."

The latter quote is a common interpretation of Baddeley's (2012) working memory framework. According to this view, the phonological loop—made up of the phonological store and the articulatory control process—retains verbal information over the short term. Items in the phonological store are represented by a phonological code and decay unless refreshed via articulatory rehearsal. There is no place within the loop for nonphonological information. Although an episodic buffer is posited, there is no requirement that it interact with the phonological loop all the time. This allows for the interpretation that if a small closed pool is used, there is no need for the episodic buffer to be involved, and therefore long-term factors are either minimized or play no role.

A quite different view of memory can also be seen as equivocal on whether item information plays a role when a small closed set of items is used. Within the Hughes et al. (2009) framework, serial recall involves primarily perceptual and motor processes. Perceptual objects are mapped onto a motor-planning process, and limits of the ability to reproduce a sequence in order arise naturally from the built-in limitation that only one biological action can be performed at a time. When the items are all known, there is little if any role for long-term memory to play (Hughes et al., 2016)—hence, the distinction between "pure" and "nonpure" serial recall. This view can be taken as predicting no or only minimal effects of long-term or lexical factors when pure serial recall is tested.[1]

In contrast, in some theories long-term and lexical factors are always involved, and the impurity principle is endorsed—either implicitly or explicitly—even in edge cases. For example, Cowan's (1999) embedded processes model views working memory as the activated part of long-term memory rather than as a separate memory store. Long-term memory factors, then, are inherently part of a working memory representation, and because of this, semantic, lexical, linguistic, and other long-term memory factors naturally affect working memory and immediate serial recall, regardless of the set size.

Although Nairne's (1990) feature model differs in almost every way from Cowan's (1999) embedded processes account, impurity is again central. Items are represented as vectors of features, and all recall is from secondary memory, even when the task is immediate serial recall; primary memory is simply where cues are held. Correct recall thus depends on finding the best relative match for a cue from items in secondary memory. As in Cowan's model, this means that semantic, lexical, linguistic, and other long-term memory factors affect immediate serial recall.

Only a handful of studies have directly assessed whether long-term factors affect immediate serial recall when a very small closed set is used. Walker and Hulme (1999) examined the immediate serial recall of abstract and concrete words using a closed stimulus set. In a block of trials, a subject heard a seven-word list drawn from 16 abstract words. For that block, the set of possible to-be-remembered items was therefore known. In another block, the same subject heard a seven-word list drawn from 16 concrete words. Walker and Hulme observed a concreteness effect. However, it is possible to argue that this closed set was not sufficiently small: Even though a block of trials would draw from the same set of 16 words, only seven of those words appeared on any given trial, and

[1] It should be noted that the perceptual–gestural view was intended, by its creators, as a replacement for memory. For example, as D. M. Jones and Macken (2018, p. 351) stated, "Our goal is not to present a new theory of verbal short-term memory (vSTM), but to supplant the concepts used to explain performance on vSTM tasks for some 60 years."

thus some uncertainty could remain about which words had been presented.

Roodenrys and Quinlan (2000) examined immediate serial recall of high- and low-frequency words as a function of set size. On each trial, subjects saw six-item lists of either low- or high-frequency words. For each subject, eight words were drawn randomly from a larger pool of 92 high-frequency words, and eight words were drawn randomly from a larger pool of 92 low-frequency words. The subjects then received blocks of trials in which all trials within the block were from either the open pool (the remaining 84 words of each type) or the closed pool. Roodenrys and Quinlan found word frequency effects with both open and closed pools. Again, however, one might claim that some uncertainty remained on each trial, if only because there were eight high-frequency words, but a trial in the closed-pool condition would present only six of those words. Quinlan, Roodenrys, and Miller (2017, Exp. 1) reported a replication.

The purpose of the four experiments reported here was to remove as much of the remaining uncertainty as possible and examine whether concreteness and frequency effects would still be observed. Experiments 1 and 3 used traditional within-subjects manipulations of concreteness (Exp. 1) and frequency (Exp. 3), but minimized uncertainty as much as is possible for this design: In the closed-pool conditions, the stimulus pool was the same size as the list length. Experiments 2 and 4 used less common between-subjects manipulations of concreteness (Exp. 2) and frequency (Exp. 4), such that for a given subject, the same six items appeared on every trial. The impurity principle predicts that concreteness and frequency effects can still be found under these conditions.

In addition to looking at the difference in recall of the two word types (i.e., abstract vs. concrete, low vs. high frequency), the analyses also included performance as a function of experiment half. It is possible that even with a closed set, some time would be required in order to learn the words in the set. If this were the case, then one possible pattern of results would be that a concreteness effect would be seen in the first half of the experiment but absent in the second half. This pattern would provide evidence against the impurity principle, which predicts effects in both list halves. As a final additional analysis, errors were analyzed.

Experiment 1

Experiment 1 was designed to see whether the concreteness effect that is observable when a large open set is used would also be observable when a small closed set was used. It differed from the experiments reported by Walker and Hulme (1999) in two important details. First, Walker and Hulme used 16 abstract and 16 concrete words for their closed pool, but their list length was only seven, and therefore some uncertainty remained about which items would appear on any given trial. In contrast, in Experiment 1 we used six abstract and six concrete words in the closed pool and a six-item list, thereby removing any doubt about which abstract or concrete words could appear. A second difference was that in the closed-pool condition, all of the subjects in the Walker and Hulme study received the same stimuli. In contrast, the particular words used in the closed pool in Experiment 1 were randomly determined for each subject, thereby mitigating any possible effects of an odd or unusual word in the stimulus set. We also included an open-pool condition in which unique items were used on every trial.

Method

Subjects Sixty volunteers from ProlificAC were paid £3 (prorated from £8.00 per hour) for their participation. For all experiments reported here, the following inclusion criteria were used: (1) native speaker of English, (2) approval rating of at least 90% on prior submissions at ProlificAC, and (3) age between 19 and 39. The mean age was 28.53 years (SD = 5.27, range 19–38); 38 of the subjects self-identified as female, and 22 self-identified as male. The subjects were randomly assigned to one of two groups, open pool or closed pool. The sample size was set at 30 subjects in each of the open- and closed-pool conditions, based on previous immediate serial recall studies using subjects from ProlificAC.

Design The experiment had a 2 Set Size (open vs. closed) × 2 Word Type (abstract vs. concrete) × 6 Serial Position mixed factorial design. Set size was a between-subjects factor, whereas word type and serial position were within-subjects factors.

Stimuli The stimuli were 196 concrete and 196 abstract one-syllable nouns. The words were drawn originally from a much larger pool sampled from Coltheart (1981), and the original pool was reduced in size until the abstract and concrete words were equated for familiarity, number of letters, number of phonemes, orthographic and phonological neighborhood size and frequency, and contextual diversity. Details are provided in Appendix 1. Importantly, all of the concrete words were higher on the measures of concreteness than all of the abstract words.

Procedure The subjects used a mouse to click a "Start next trial" button on a computer screen. One second after the fixation point had disappeared, six words were shown one at a time for 1 s each in uppercase letters in the center of the screen. After the final word had been shown, the subjects saw a message that asked them to type in the words they had just seen, in strict serial order. They were informed that they needed to

enter the first word first, the second word second, and so on. If they were unsure of a response, they were encouraged to guess or else to click on a button labeled "Skip."

There were 32 trials, half with concrete and half with abstract words; the order of these trials was randomly determined for each subject. For subjects in the open-pool condition, a new set of six words were randomly drawn without replacement from the larger pool for each list; for each subject in the closed-pool condition, six abstract and six concrete words were randomly drawn from the larger pool at the beginning of the experiment, and then these words were used on every trial of the appropriate type.

Results

Accuracy analysis The top row of Fig. 1 shows the proportion of concrete and abstract words recalled correctly in order as a function of set size and serial position; the left panel shows the data from the first half of the experiment, and the right panel shows the data from the second half of the experiment. Concreteness effects are apparent in both panels for both set sizes.

The proportion of words recalled correctly and in order were analyzed in a 2 Set Size (open vs. closed) × 2 Word Type (abstract vs. concrete) × 2 Experiment Half (first half vs. second half) × 6 Serial Position mixed factorial analysis of variance (ANOVA).[2] All main effects were significant: More words were correctly recalled in order in the closed group (M = .717, SD = .138) than in the open group (M = .563, SD = .173), $F(1, 58) = 14.418$, $MSE = .588$, $\eta_p^2 = .199$, $p < .001$; more concrete words were recalled correctly (M = .671, SD = .174) than abstract words (M = .609, SD = .183), $F(1, 58) = 34.456$, $MSE = .041$, $\eta_p^2 = .373$, $p < .001$; and more words were recalled in the second half (M = .666, SD = .184) than in the first half (M = .614, SD = .174), $F(1, 58) = 20.970$, MSE = .047, η_p^2 = .266, $p < .001$. The main effect of serial position was also significant, $F(5, 290) = 127.850$, MSE = .066, η_p^2 = .688, $p < .001$.

Importantly, the word type by set size interaction was not significant, $F(1, 58) < 1$. Only one of the two-way interactions was significant: experiment half by position, $F(5, 290)$ = 8.074, MSE = 0.018, η_p^2 = .122, $p < .001$. We observed no improvement from the first to the second half for the early list positions, but there was improvement for the later list positions. The remaining two-way interactions were experiment half by set size, $F(1, 58) = 2.618$, $MSE = 0.047$, $\eta_p^2 = .043$, $p > .10$; experiment half by word type, $F(1, 58) < 1$; word type by position, $F(5, 290) < 1$; and position by set size, $F(5, 290)$ = 1.700, MSE = 0.066, η_p^2 = .028, $p > .10$.

None of the three-way interactions were significant, and all had Fs < 1, except for experiment half by word type by set size, $F(1, 58) = 2.875$, MSE = .044, η_p^2 = .047, p = .095. The four-way interaction was not significant, $F(5, 290) < 1$.

Error analysis Each incorrect response was categorized as either an intrusion error (the word reported was not in the list), an omission error (no response was given), a repetition error (a word that had already been reported was reported again), or a position error (a word in the list was reported in an incorrect position). The proportion of responses that fell in each error category are shown in Table 1, along with the proportion correct.

The number of repetition errors did not differ as a function of set size, $t(58) = 0.142$, $p > .85$. Because of this, and also because (1) the rate of occurrence was very low (they accounted for only 2.7% of all responses) and (2) the impurity principle makes no specific predictions for this error type, no further analyses were performed. The position errors were also not analyzed; although these accounted for the majority of errors, no specific predictions were made.

The two types of errors most directly related to the impurity principle and the pure serial recall hypothesis are intrusions and omissions in the closed set, because the presence of either suggests that the items have not yet been learned. In the open-set condition it was not possible to learn the words, because they were never repeated, and therefore intrusion and omission errors in this condition could be used as a baseline. The middle row of Fig. 1 shows the mean number of intrusions for each half of the experiment, and the bottom row shows the mean number of omissions for each half of the experiment.

The mean number of intrusion errors were analyzed with a 2 Set Size (open vs. closed) × 2 List Type (abstract vs. concrete) × 2 Experiment Half (first half vs. second half) mixed factorial ANOVA.[3] All main effects were significant: There were more intrusion errors in the open condition (M = 8.200, SD = 6.570) than in the closed condition (M = 3.425, SD = 3.017), $F(1, 58) = 16.192$, $MSE = 84.488$, $\eta_p^2 = .218$, $p < .001$; more intrusion errors with abstract (M = 6.317, SD = 6.116) than with concrete (M = 5.308, SD = 5.083) lists, $F(1, 58)$ = 11.242, MSE = 5.427, η_p^2 = .162, $p < .01$; and more intrusion errors in the first half (M = 6.275, SD = 5.508) than in the second half ($M = 5.350$, $SD = 5.745$), $F(1, 58) = 5.884$, MSE = 8.725, η_p^2 = .092, $p < .05$.

The only significant interaction was List Type × Set Size, $F(1, 58) = 4.792$, $MSE = 5.427$, $\eta_p^2 = .076$, $p < .05$. There was little difference between the mean number of intrusion errors in abstract and concrete lists in the closed set (3.600 vs. 3.250, respectively), but there was a much larger difference in the open set (9.267 vs. 7.583). For all other interactions, Fs < 1.

[2] See Appendix 2 for details on the scoring.

[3] Position was not included as a factor because there were no a priori predictions concerning position, and also because including it would make the results more difficult to parse.

The same 2 Set Size (open vs. closed) × 2 List Type (abstract vs. concrete) × 2 Experiment Half (first half vs. second half) mixed factorial ANOVA was performed on the mean number of omission errors. As with the intrusion errors, we found more omission errors in the open condition (M = 8.683, SD = 10.047) than in the closed condition (M = 2.667, SD = 3.988), $F(1, 58) = 9.959$, $MSE = 218.106$, $\eta_p^2 = .147$, $p < .01$. There were also more omission errors with abstract (M = 6.433, SD = 8.554) than with concrete (M = 5.417, SD = 7.836) lists, $F(1, 58) = 9.454$, $MSE = 6.560$, $\eta_p^2 = .140$, $p < .01$. Unlike with the intrusion data, the main effect of experiment half was not significant: The same number of omission errors occurred in the first half (M = 5.925, SD = 8.157) as in the second half (M = 5.425, SD = 8.272), $F(1, 58) = 1.781$, $MSE = 8.421$, $\eta_p^2 = .030$, $p > .15$.

Unlike with the intrusion data, the only significant interaction was Experiment Half × Set Size, $F(1, 58) = 6.658$, $MSE = 8.421$, $\eta_p^2 = .103$, $p < .05$. In the closed condition, there was a decrease in the number of omissions from the first to the second half (3.400 vs. 1.933, respectively), whereas in the open condition there was no decrease (8.450 vs. 8.917). The rest of the interactions were List Type × Set Size, $F(1, 58) = 2.137$, $MSE = 6.560$, $\eta_p^2 = .036$, $p > .14$; Experiment Half × List Type, $F < 1$; and the three-way interaction, $F(1, 58) = 2.897$, $MSE = 3.889$, $\eta_p^2 = .048$, $p = .094$.

Discussion

The concreteness effect observed in Experiment 1 supports the impurity principle. In particular, in the closed condition in the second half of the experiment, we found, on average, 3.000 intrusion errors and 1.933 omission errors. These very low rates indicate that the subjects had learned the words in the closed pool, and therefore there should have been no concreteness effect. Despite this, evidence of a concreteness effect emerged in both experiment halves, but no evidence that the concreteness effect in the closed condition in the second half was different, either from that in the open condition or those observed in the first half of the experiment.

These results replicated those of Walker and Hulme (1999) and extended them by showing that even reducing the uncertainty of the list items to a minimum, given the type of design, does not reduce the concreteness effect. It further extends them both by analyzing each half of the experiment and by showing that the pattern of errors is also fully consistent with the earlier results.

Experiment 2

It is possible to argue that the subjects might still have had some uncertainty about the items that would appear on each trial, because half the time abstract words would appear, whereas the other half of the time concrete words would appear. The purpose of Experiment 2 was to address this point by taking advantage of the fact that the concreteness effect is readily observable in between-subjects designs (e.g., Neath, 1997; Ruiz-Vargas, Cuevas, & Marschark, 1996; Yuille & Paivio, 1968). In Experiment 2, the subjects in the abstract condition saw the same six abstract words on every trial, whereas the subjects in the concrete condition saw the same six concrete words on every trial. Despite the fact that there could be no doubt about which words would be shown, the impurity principle still predicts that a concreteness effect would be observed, because immediate serial recall is not a pure test, or even a relatively pure test.

Method

Subjects Sixty different volunteers from ProlificAC were paid £3 (prorated from £8.00 per hour) for their participation. The mean age was 28.68 years (SD = 5.97, range 19–38); 28 of the subjects self-identified as female, and 32 self-identified as male. The subjects were randomly assigned to one of two groups, abstract or concrete.

Design The experiment had a 2 Word Type (abstract vs. concrete) × 6 Serial Position mixed factorial design. Word type was a between-subjects factor, whereas serial position was a within-subjects factor.

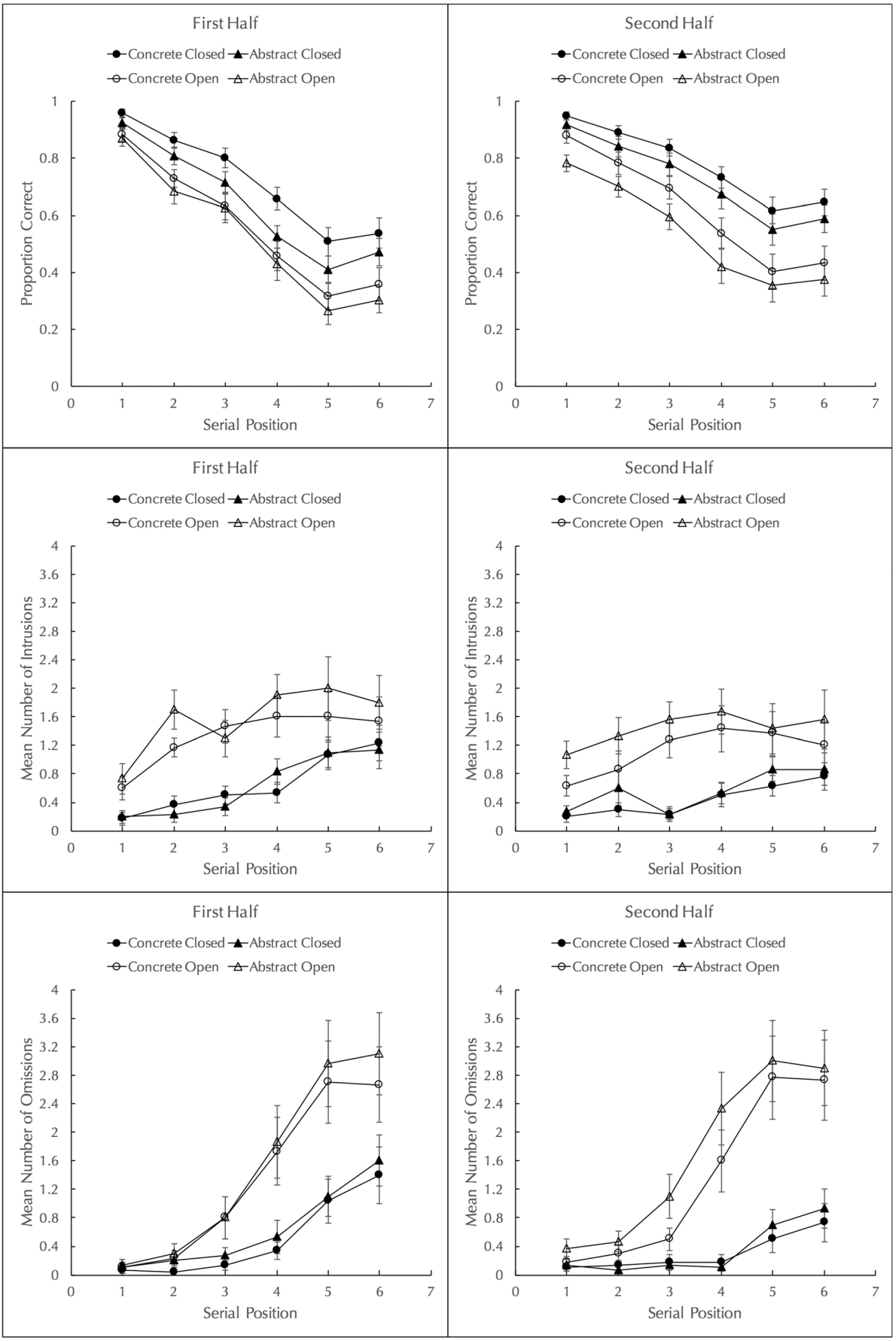
Stimuli The stimuli were the same as in Experiment 1.

Procedure The procedure was identical to that of Experiment 1, except for the following. For each subject, six words were randomly drawn from the larger pool, either six abstract words or six concrete words, depending on the condition. These six words then appeared in random order on every trial. Each subject received 18 trials.

Results

Accuracy analysis The top row of Fig. 2 shows the proportion of concrete and abstract words correctly recalled in order as a function in the first half of the experiment (left panel) and the second half of the experiment (right panel). A concreteness effect is apparent in both figures.

The proportion of words recalled correctly and in order were analyzed in a 2 Word Type (abstract vs. concrete) × 2 Experiment Half (first half vs. second half) × 6 Serial Position mixed factorial ANOVA. All main effects were significant: More concrete words were recalled correctly (M = .743, SD = .109) than abstract words (M = .667, SD = .169), $F(1, 58) = 4.302$, $MSE = .243$, $\eta_p^2 = .069$, $p < .05$; more words were recalled in the second half (M = .737, SD = .158) than in the first (M = .673, SD = .156), $F(1, 58)$ =


First Half
Second Half
Concrete Closed
Abstract Closed
Concrete Open
Abstract Open
Proportion Correct
Mean Number of Intrusions
Mean Number of Omissions
Serial Position

◀ **Fig. 1** Proportion of concrete and abstract words recalled correctly in order (top row), mean number of intrusion errors (middle row), and mean number of omission errors (bottom row) for the first eight lists (left panels) or last eight lists (right panels) in Experiment 1. Error bars show the standard error of the means

17.273, MSE = .043, η_p^2 = .229, p < .001; and finally, the main effect of serial position was significant, F(5, 290) = 81.027, MSE = .044, η_p^2 = .583, p < .001.

The only significant interaction was Experiment Half × Position, F(5, 290) = 4.412, MSE = 0.016, η_p^2 = .071, p < .01. For all other interactions, Fs < 1.

Error analysis The proportion of responses that fell in each error category are shown in Table 2, along with the proportion correct. The repetition errors, which accounted for 3.8% of all responses, did not differ as a function of word type, t(58) = 1.514, p > .10. The middle row of Fig. 2 shows the mean number of intrusion errors in each experiment half, and the bottom row of Fig. 2 shows the mean number of omission errors in each half.

The mean number of intrusion errors were analyzed with a 2 List Type (abstract vs. concrete) × 2 Experiment Half (first half vs. second half) mixed factorial ANOVA. There were more intrusion errors in the first half (M = 1.717, SD = 2.187) than in the second half (M = 1.183, SD = 2.318), F(1, 58) = 4.137, MSE = 2.063, η_p^2 = .067, p < .05. However, the number of intrusion errors did not differ as a function of list type, F(1, 58) = 1.339, MSE = 8.067, η_p^2 = .023, p > .25, with a mean of 1.150 (SD = 2.335) for abstract, as compared to 1.750 (SD = 2.160) for concrete lists. The interaction was not significant, F < 1.

The mean number of omission errors were analyzed with a 2 List Type (abstract vs. concrete) × 2 Experiment Half (first half vs. second half) mixed factorial ANOVA. More omission errors occurred in the first half (M = 2.517, SD = 3.934) than in the second half (M = 0.967, SD = 3.092), F(1, 58) = 19.316, MSE = 3.731, η_p^2 = .250, p < .001. However, the number of omission errors did not differ as a function of list type, F(1, 58) = 2.005, MSE = 20.957, η_p^2 = .033, p > .15, with a mean of 2.333 (SD = 4.550) for abstract, as compared to 1.150 (SD = 2.200) for concrete. The interaction was not significant, F < 1.

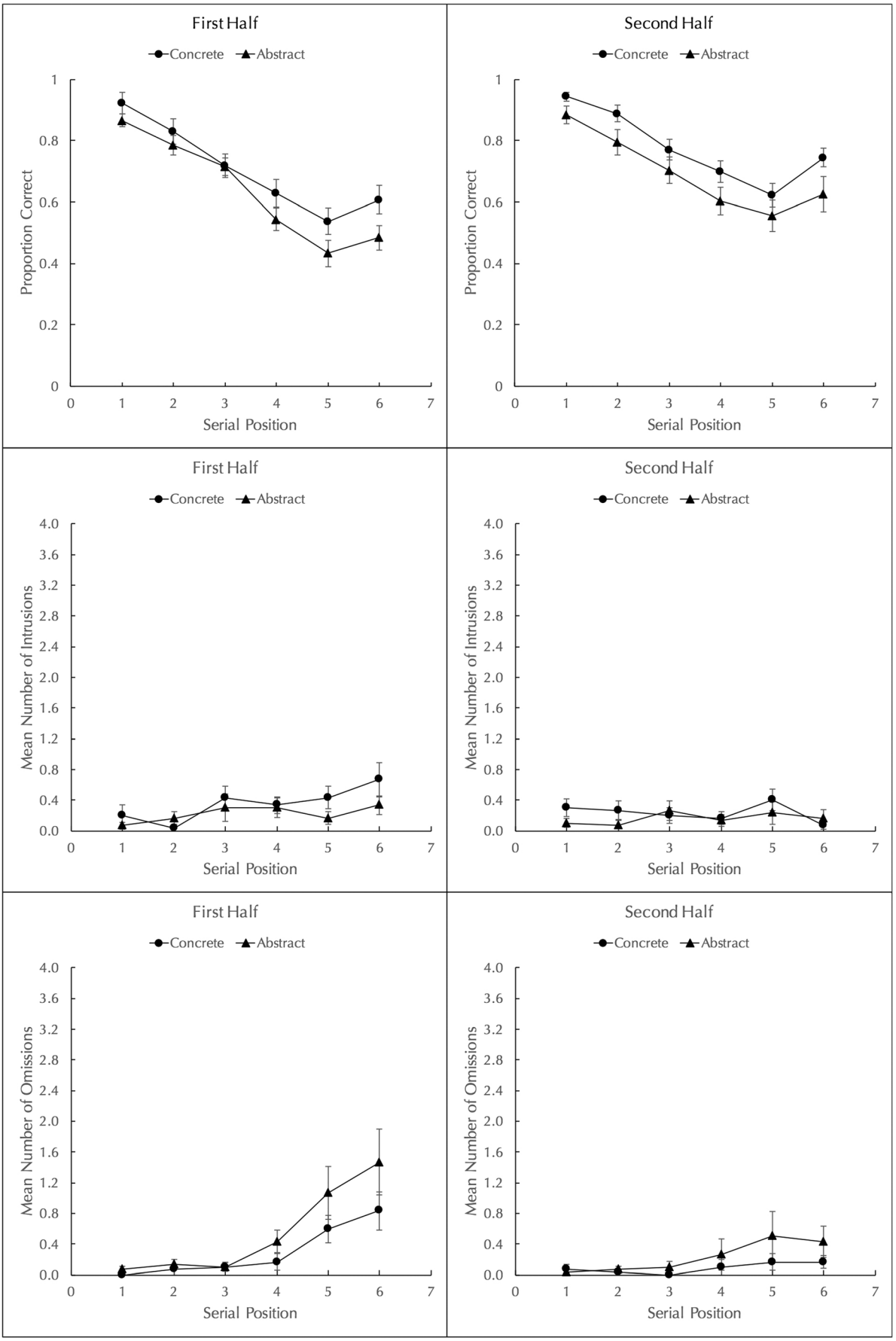
Discussion

Concreteness effects were observed in both experiment halves, even though a given subject saw the same six words on every trial. The error analysis suggests that the subjects had learned the six words by the second half of the experiment: The mean number of intrusion errors in the second half was 1.183, and the mean number of omission errors was 0.967. Put another way, over the course of the final nine lists, on average, a given subject made only one omission error and only one intrusion error, but concrete words were still recalled more accurately than abstract words. This result is contrary to the pure serial recall hypothesis, which predicts that when the role of long-term memory is minimized by using the same items on every trial, long-term memory effects would not be apparent. In contrast, the impurity principle states that tasks are not pure. Immediate serial recall is not a pure measure—or even a relatively pure measure—of order information, even when a small closed set is used.

One possible objection to Experiments 1 and 2 is that the abstract–concrete dimension is potentially not an appropriate test of long-term or lexical influence. Although some accounts of the concreteness effect are based on semantic properties (e.g., G. V. Jones, 1988; Schwanenflugel, 1991), alternate explanations invoke differential processing (e.g., Marschark & Hunt, 1989; Paivio, 1991)—because concrete words afford the construction of an image, whereas abstract words do not—and it may be that this difference remains even when uncertainty about the identity of the to-be-remembered items is minimized. Therefore, in the next two experiments the dimension of interest was changed to word frequency. The predictions of the impurity principle remain the same: A word frequency effect would obtain despite the use of a small closed

Table 1 Proportion of each type of response in Experiment 1

			Errors				
			Intrusion	Omission	Repetition	Position	Correct
Closed	First half	Abstract	.080	.079	.040	.158	.642
		Concrete	.081	.063	.019	.117	.721
	Second half	Abstract	.070	.043	.030	.132	.725
		Concrete	.055	.038	.026	.103	.779
Open	First half	Abstract	.197	.191	.027	.056	.530
		Concrete	.166	.172	.026	.073	.563
	Second half	Abstract	.180	.212	.027	.043	.538
		Concrete	.141	.168	.022	.047	.622


First Half
Concrete
Abstract
Proportion Correct
Serial Position
Second Half
Concrete
Abstract
Proportion Correct
Serial Position
First Half
Concrete
Abstract
Mean Number of Intrusions
Serial Position
Second Half
Concrete
Abstract
Mean Number of Intrusions
Serial Position
First Half
Concrete
Abstract
Mean Number of Omissions
Serial Position
Second Half
Concrete
Abstract
Mean Number of Omissions
Serial Position

◀ **Fig. 2** Proportion of concrete and abstract words recalled correctly in order (top row), mean number of intrusion errors (middle row), and mean number of omission errors (bottom row) for the first nine lists (left panels) or the last nine lists (right panels) in Experiment 2. Error bars show the standard error of the means

set. In contrast, researchers who only partially endorse the impurity principle might predict no effect, because item information would no longer be necessary or would be so reduced in importance that only order information would be required.

Experiment 3

Experiment 3 was designed to see whether the frequency effect that is observed when a large open set is used would also be observed when a small closed set was used, as had previously been reported by both Roodenrys and Quinlan (2000) and Quinlan et al. (2017). Experiment 3, then, was identical to Experiment 1, except that word frequency was manipulated instead of concreteness. Roodenrys and Quinlan had a list length of six but a closed pool size of eight, whereas Quinlan et al. had a list length of six and a closed pool size of six. We followed the latter design.

Method

Subjects Sixty different volunteers from ProlificAC were paid £3 (prorated from £8.00 per hour) for their participation. Their mean age was 29.98 years (*SD* = 5.08, range 19–38); 36 of the subjects self-identified as female, and 24 self-identified as male. The subjects were randomly assigned to one of two groups, open pool versus closed pool.

Design The experiment had a 2 Set Size (open vs. closed) × 2 Word Type (low vs. high frequency) × 6 Serial Position mixed factorial design. Set size was a between-subjects factor, whereas word type and serial position were within-subjects factors.

Stimuli The stimuli were 118 low- and 118 high-frequency one-syllable nouns. The words were drawn originally from a much larger pool sampled from Coltheart (1981), and the original pool was reduced in size until the high- and low-frequency words were equated for concreteness, familiarity, imageability, number of letters, number of phonemes, and phonological and orthographic neighborhood. Details are provided in Appendix 1. Importantly, all high-frequency words were higher than all low-frequency words on two different measures of frequency.

Procedure The procedure was identical to that of Experiment 1, except for the stimuli.

Results

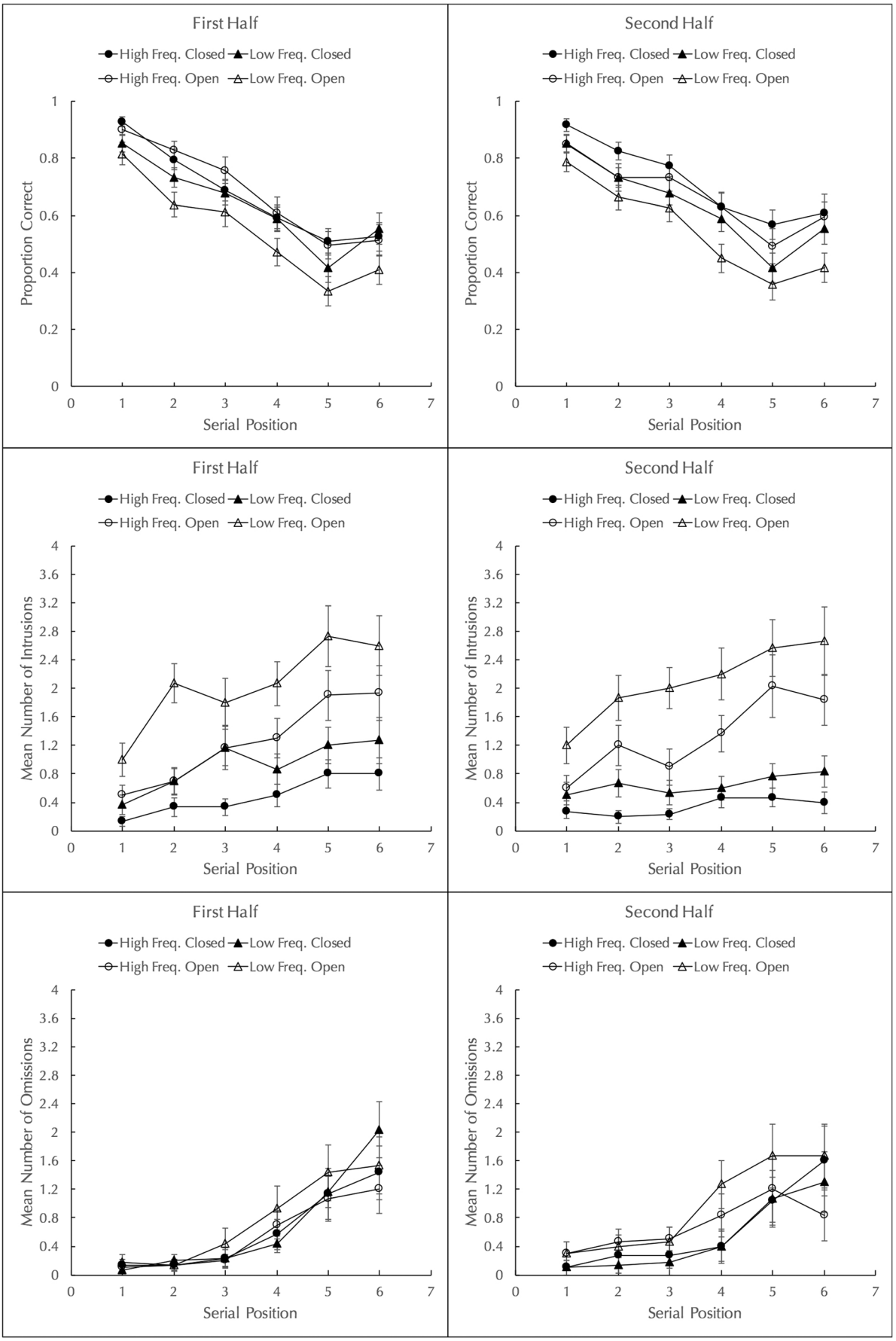
Accuracy analysis The top row of Fig. 3 shows the proportion of high- and low-frequency words recalled correctly and in order as a function of set size and serial position; the left panel shows the data from the first half of the experiment, and the right panel shows the data from the second half of the experiment. Word frequency effects are apparent in both panels for both set sizes.

The proportion of words correctly recalled in order were analyzed with a 2 Set Size (open vs. closed) × 2 Word Type (low vs. high frequency) × 2 Experiment Half (first half vs. second half) × 6 Serial Position mixed factorial ANOVA. Only two main effects were significant: More high-frequency words were recalled correctly (M = .687, SD = .175) than low-frequency words (M = .579, SD = .158), $F(1, 58) = 60.939$, MSE = .070, η_p^2 = .512, $p < .001$, and the main effect of serial position was significant, $F(5, 290) = 109.025$, MSE = .058, η_p^2 = .653, $p < .001$. Performance did not differ between the first half (M = .621, SD = .148) and the second half (M = .645, SD = .187), $F(1, 58) = 2.402$, MSE = .085, η_p^2 = .040, $p > .10$, and also did not differ between the closed group (M = .653, SD = .139) and the open group (M = .613, SD = .173), $F < 1$.

The only significant two-way interactions involved position. Both the experiment half by position, $F(5, 290) = 4.939$, MSE = .021, η_p^2 = .078, $p < .001$, and the word type by position, $F(5, 290) = 2.613$, MSE = .022, η_p^2 = .043, $p < .05$, interactions were significant. The experiment half by set size interaction failed to reach the adopted significance level, $F(1, 58) = 3.317$, MSE = .085, η_p^2 = .054, p = .074. To the extent that this interaction exists, it reflects an improvement in the closed set from the first to the last half (.627 vs. .679),

Table 2 Proportion of each type of response in Experiment 2

		Errors				
		Intrusion	Omission	Repetition	Position	Correct
First half	Abstract	.025	.060	.041	.275	.599
	Concrete	.041	.030	.037	.220	.672
Second half	Abstract	.018	.026	.046	.261	.649
	Concrete	.026	.010	.028	.186	.751


First Half
Second Half
High Freq. Closed
Low Freq. Closed
High Freq. Open
Low Freq. Open
Proportion Correct
Mean Number of Intrusions
Mean Number of Omissions
Serial Position

◀ **Fig. 3** Proportion of high- and low-frequency words recalled correctly in order (top row), mean number of intrusion errors (middle row), and mean number of omission errors (bottom row) for the first eight lists (left panels) or last eight lists (right panels) in Experiment 3. Error bars show the standard error of the means

whereas there was no improvement in the open set (.615 vs. .611). None of the remaining two-way interactions were significant: for word type by set size, $F(1, 58) = 2.360$, $MSE = .070$, $\eta_p^2 = .039$, $p > .10$, and for both position by set size and list half by word type, Fs < 1.

None of the three-way interactions were significant, and all had Fs < 1. The four-way interaction was significant, $F(5, 290) = 2.505$, $MSE = .018$, $\eta_p^2 = .041$, $p < .05$.

The important results from the analysis of the accuracy data are that a word frequency effect was observed and that it did not interact with either set size or experiment half. The lack of a set size effect replicated Experiment 1 of Quinlan et al. (2017) but differed from the significant set size effect reported by Roodenrys and Quinlan (2000). It is not clear what the key difference is between the studies.

Error analysis The proportion of responses that fell in each error category are shown in Table 3, along with the proportion correct. As in Experiment 1, the mean number of repetition errors did not differ as a function of set size, $t(58) = 0.333$, $p > .70$, and their occurrence remained very low (they accounted for less than 2.9% of all responses). The middle row of Fig. 3 shows the mean number of intrusions for each half of the experiment, and the bottom row of Fig. 3 shows the mean number of omissions for each half of the experiment.

The mean number of intrusion errors were analyzed with a 2 Set Size (open vs. closed) × 2 List Type (low vs. high frequency) × 2 Experiment Half (first half vs. second half) mixed factorial ANOVA. There were more intrusion errors in the open condition ($M = 10.050$, $SD = 7.869$) than in the closed condition ($M = 3.600$, $SD = 4.096$), $F(1, 58) = 21.971$, $MSE = 113.612$, $\eta_p^2 = .275$, $p < .001$. There were also more intrusion errors with low-frequency ($M = 8.558$, $SD = 7.677$) than with high-frequency ($M = 5.092$, $SD = 5.888$) lists, $F(1, 58) = 45.482$, $MSE = 15.854$, $\eta_p^2 = .440$, $p < .001$. However, we observed equal number of intrusion errors in the first half ($M = 7.058$, $SD = 6.590$) and in the second half ($M = 6.591$, $SD = 7.431$), $F(1, 58) = 1.294$, $MSE = 10.095$, $\eta_p^2 = .022$, $p > .25$.

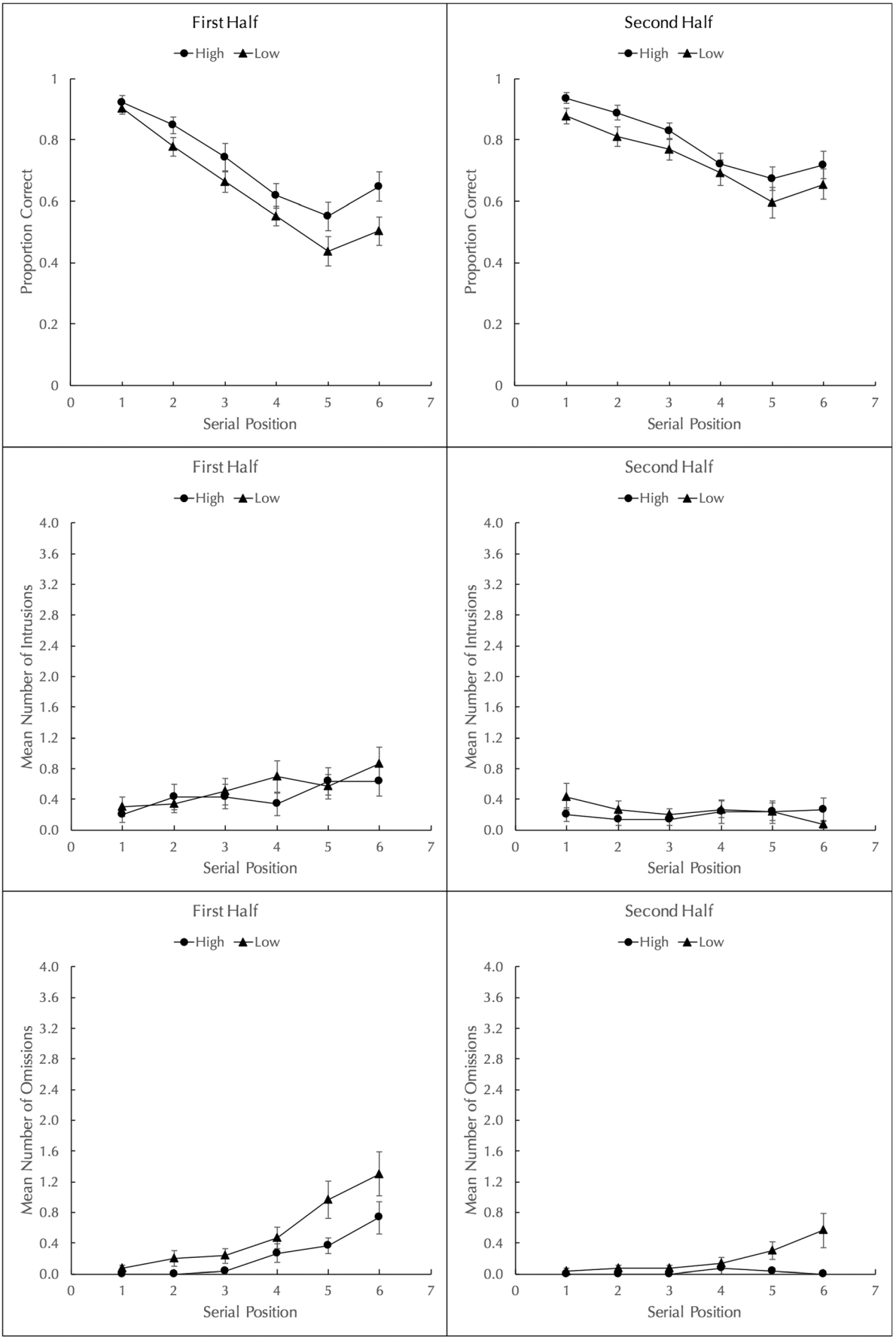
One interaction was significant, List Type × Set Size, $F(1, 58) = 5.450$, $MSE = 15.854$, $\eta_p^2 = .086$, $p < .05$, reflecting a larger difference in intrusion errors in the open condition between the low- and high-frequency lists (12.383 vs. 7.717) than in the closed condition (4.733 vs. 2.467). The Experiment Half × Set Size interaction just failed to reach the adopted significance level, $F(1, 58) = 3.804$, $MSE = 10.095$, $\eta_p^2 = .022$, $p = .056$. To the extent the interaction exists, it reflects a numerical decrease in intrusion errors in the closed group (from 4.233 to 2.967), but an increase in the open group (from 9.883 to 10.217). For all other interactions, Fs < 1.

The mean number of omission errors were also analyzed with a 2 Set Size (open vs. closed) × 2 List Type (abstract vs. concrete) × 2 Experiment Half (first half vs. second half) mixed factorial ANOVA. Unlike with the intrusion data, we found the same number of omission errors in the open condition ($M = 4.483$, $SD = 6.399$) as in the closed condition ($M = 3.650$, $SD = 5.929$), $F < 1$. There were more omission errors with low-frequency ($M = 4.425$, $SD = 6.233$) than with high-frequency ($M = 3.708$, $SD = 6.110$) lists, $F(1, 58) = 4.768$, $MSE = 6.463$, $\eta_p^2 = .076$, $p < .05$. The main effect of experiment half was not significant: The same number of omission errors occurred in the first half ($M = 3.950$, $SD = 5.182$) as in the second half ($M = 4.183$, $SD = 7.040$), $F < 1$.

As with the intrusion data, the only significant interaction in the omission data was List Type × Set Size, $F(1, 58) = 4.768$, $MSE = 6.463$, $\eta_p^2 = .076$, $p < .05$. In the closed-set condition, no difference emerged in the number of omissions for low- and high-frequency lists (3.650 vs. 3.650, respectively) whereas in the open-set condition, more omissions occurred for low-frequency than for high-frequency lists (5.200 vs. 3.767). The remaining interactions were

Table 3 Proportion of each type of response in Experiment 3

			Errors				
			Intrusion	Omission	Repetition	Position	Correct
Closed	First half	Low	.117	.085	.036	.181	.581
		High	.064	.072	.035	.156	.673
	Second half	Low	.081	.066	.023	.192	.638
		High	.042	.076	.017	.144	.720
Open	First half	Low	.256	.096	.040	.062	.547
		High	.157	.070	.033	.056	.684
	Second half	Low	.260	.120	.022	.047	.550
		High	.166	.085	.024	.052	.672


First Half
High
Low
Proportion Correct
Serial Position
Second Half
High
Low
Proportion Correct
Serial Position
First Half
High
Low
Mean Number of Intrusions
Serial Position
Second Half
High
Low
Mean Number of Intrusions
Serial Position
First Half
High
Low
Mean Number of Omissions
Serial Position
Second Half
High
Low
Mean Number of Omissions
Serial Position

◀ **Fig. 4** Proportion of high- and low-frequency words recalled correctly in order (top row), mean number of intrusion errors (middle row), and mean number of omission errors (bottom row) for the first nine lists (left panels) or last nine lists (right panels) in Experiment 4. Error bars show the standard error of the means

Experiment Half × Set Size, $F(1, 58) = 1.797$, $MSE = 16.359$, $\eta_p^2 = .030$ $p > .15$; Experiment Half × Word Type, $F < 1$; and the three-way interaction, $F(1, 58) = 1.308$, $MSE = 5.617$, $\eta_p^2 = .022$, $p > .25$.

Discussion

Experiment 3 resulted in frequency effects in both the open and closed pools. Evidence consistent with the claim that there was little or no uncertainty about the items comes from the mean number of intrusion and omission errors in the second half of the experiment in the closed-set condition: on average, 2.967 intrusion errors and 3.417 omission errors. Despite such low error rates, a word frequency effect was observed, replicating the results of Roodenrys and Quinlan (2000). The results also reinforce those from Experiment 1: Long-term effects can be observed in immediate serial recall even when a small closed set is used, and even when performance is considered over just the second half of the experiment. This pattern of results is consistent with the impurity principle and inconsistent with accounts that question its applicability in edge cases.

Experiment 4

As with Experiment 1, it is possible to argue that the subjects in Experiment 3 might still have had some uncertainty about the items that would appear on each trial, because half the time low-frequency words would appear, whereas the other half of the time high-frequency words would appear. The purpose of Experiment 4 was to address this point by taking advantage of the fact that the frequency effect is readily observable in between-subjects designs (e.g., Morin, Poirier, Fortin, & Hulme, 2006; Saint-Aubin & Poirier, 2005; Stuart & Hulme, 2000). Therefore, in Experiment 4 one group of subjects saw the same six low-frequency words on every trial, and a second group of subjects saw the same six high-frequency words on every trial. According to the impurity principle, a word frequency effect should still be apparent, because serial recall—even with a small closed set—is not a pure, or even relatively pure, test of order memory.

Method

Subjects Sixty different volunteers from ProlificAC were paid £3 (prorated from £8.00 per hour) for their participation. The mean age was 29.33 years ($SD = 5.09$, range 19–38); 29 of the subjects self-identified as female, 28 self-identified as male, and three did not respond to the question. The subjects were randomly assigned to one of two groups, low or high frequency.

Design The experiment had a 2 Word Type (low vs. high frequency) × 6 Serial Position mixed factorial design. Word type was a between-subjects factor, whereas serial position was a within-subjects factor.

Stimuli The stimuli were the same as in Experiment 3.

Procedure The procedure was identical to that of Experiment 2, except that high- and low-frequency words were used.

Results

Accuracy analysis The top row of Fig. 4 shows the proportion of high- and low-frequency words recalled correctly and in order in the first half of the experiment (left panel) and the second half of the experiment (right panel). A frequency effect is apparent in both panels.

The proportion of words correctly recalled in order were analyzed in a 2 Word Type (low vs. high frequency) × 2 Experiment Half (first half vs. second half) × 6 Serial Position mixed factorial ANOVA. Unlike in Experiment 3, all main effects were significant. More high-frequency words were recalled correctly ($M = .759$, $SD = .122$) than low-frequency words ($M = .687$, $SD = .149$), $F(1, 58) = 4.200$, $MSE = .222$, $\eta_p^2 = .068$, $p < .05$. Also, more words were recalled in the second half ($M = .765$, $SD = .152$) than in the first half ($M = .680$, $SD = .149$), $F(1, 58) = 32.261$, $MSE = .039$, $\eta_p^2 = .357$, $p < .001$. Finally, the main effect of serial position was significant, $F(5, 290) = 73.034$, $MSE = .029$, $\eta_p^2 = .557$, $p < .001$.

The only significant interaction was Experiment Half × Position, $F(5, 290) = 6.771$, $MSE = .014$, $\eta_p^2 = .105$, $p < .001$. For all other interactions, Fs < 1.

Error analyses The proportion of responses that fell in each error category are shown in Table 4, along with the proportion correct. The number of repetition errors, which accounted for 3.5% of all responses, did not differ as a function of word type, $t(58) = 0.194$, $p > .80$.

The mean number of intrusion errors were analyzed with a 2 List Type (low vs. high frequency) × 2 Experiment Half (first half vs. second half) mixed factorial ANOVA. More intrusion errors occurred in the first half ($M = 2.967$, $SD = 3.508$) than in the second half ($M = 1.333$, $SD = 2.556$), $F(1, 58) = 25.769$, $MSE = 3.106$, $\eta_p^2 = .308$, $p < .001$. However, the number of intrusion errors did not differ as a function of list type, $F < 1$, with a mean of 2.367 ($SD = 2.940$) for low-frequency, as compared to 1.933 ($SD = 3.384$) for high-frequency, lists. The interaction was not significant, $F < 1$.

Table 4 Proportion of each type of response in Experiment 4

		Errors				
		Intrusion	Omission	Repetition	Position	Correct
First half	Low	.061	.059	.035	.236	.609
	High	.049	.026	.033	.202	.690
Second half	Low	.027	.022	.036	.217	.698
	High	.022	.002	.035	.181	.760

The mean number of omission errors were analyzed with a 2 List Type (abstract vs. concrete) × 2 Experiment Half (first half vs. second half) mixed factorial ANOVA. There were more omission errors in the first half (M = 2.317, SD = 3.244) than in the second half (M = 0.633, SD = 1.687), $F(1, 58) = 23.925$, MSE = 3.553, η_p^2 = .292, $p < .001$. There were also more omission errors for low-frequency (M = 2.200, SD = 3.379) than for high-frequency (M = 0.750, SD = 1.525) lists, $F(1, 58) = 7.099$, MSE = 8.885, $\eta_p^2 = .109$, $p < .05$. The interaction was not significant, $F(1, 58) = 1.241$, MSE = 3.553, $\eta_p^2 = .021$, $p > .25$.

Discussion

Frequency effects were observed in both experiment halves, even though a given subject saw the same six words on every trial. The error analysis suggests that the subjects had learned the six words by the second half of the experiment: The mean number of intrusion errors in the second half was 1.333, and the mean number of omission errors was 0.633. Despite this, a frequency effect was observed. This result is contrary to views that allow for edge cases, such as the pure serial recall hypothesis, which predicts that when the role of long-term memory is minimized by using the same items on every trial, long-term memory and lexical effects would not be apparent. In contrast, the impurity principle states that tasks are not pure, and that immediate serial recall is not a pure test of order memory. Therefore, given a sufficiently large manipulation of frequency, a frequency effect should be observed.

General discussion

In the study of memory, it has long been proposed that tasks are not pure (Crowder, 1993; Jacoby, 1991; Kolers & Roediger, 1984; Restle, 1974), and the impurity principle summarizes this long line of work. Despite its wide endorsement in general, many researchers appear hesitant to fully endorse the principle in certain extreme edge cases. The four experiments reported here assessed one such edge case: Do long-term memory and lexical factors continue to affect memory performance on immediate serial recall tests when a small closed set of items is used?

Experiment 1 revealed a concreteness effect with a closed set, replicating Walker and Hulme (1999). Experiment 2 showed a concreteness effect in a between-subjects design in which a given subject saw the same six words on every trial. Even over the last nine trials, a concreteness effect was observed, and the number of intrusion and omission errors were vanishingly small, suggesting excellent knowledge of the six possible words that would be shown. Experiments 3 and 4 were identical to Experiments 1 and 2, respectively, except that frequency was manipulated, and the results replicated Roodenrys and Quinlan (2000) and Quinlan et al. (2017). In particular, a frequency effect was observed when only the last nine trials were considered, during which intrusion and omission errors were again vanishingly small.

According to the impurity principle view (Surprenant & Neath, 2009b), memory is inherently reconstructive, and individuals use whatever information is available in order to complete a task; if an image of a word or frequency information is available and useful, it can potentially be used. It is because of this that the impurity principle was formulated: Tasks are not pure. The results confirmed that even in this particular edge case, the principle makes the correct prediction.

A variety of models address these findings, but rather than presenting a detailed account of how each and every model fares, we will instead highlight a few models based on the extent to which they either fully or not-quite-fully endorse the impurity principle. Cowan's (1999) embedded processes model holds that working memory is activated long-term memory, and therefore that any item in working memory reflects long-term factors, including concreteness and frequency. Long-term and lexical effects with closed sets are a natural consequence of this architecture. These results are also a natural consequence of the architecture of Nairne's (1990) feature model, a quite different model that denies the existence of time-based decay. All recall is always from secondary memory, which will contain not only phonological information, but also long-term and lexical information. As with the embedded processes model, the feature model views serial recall as inherently impure. To be clear, we are not claiming that either the embedded processes model or the feature model can account for all aspects of the present results. Rather, we are emphasizing that because both models include the impurity principle as a core architectural element, both models are in principle consistent with the results.

These two models can be contrasted with models that allow for exceptions to the impurity principle in edge cases. For example, in Baddeley's (2012) working memory framework, immediate serial recall depends on a rote rehearsal loop with no necessary connection to episodic memory, except as needed to support item representations. If the set of to-be-remembered items is completely known, then there is no need for support from the episodic buffer, so it is possible for concreteness and frequency effects to be absent. This architecture allows for pure serial recall because the episodic buffer is not

required to play a role. Although Hughes et al.'s (2016) framework is very different from the phonological loop version of working memory, it also allows for the same exception. In this case, "pure serial order" tasks are based purely on subvocal speech gestures that give order to the to-be-remembered stimuli.

It is important to note the nature of the two predictions assessed here. We have focused on the strong version of the predictions made by those who do not fully endorse the impurity principle. For example, the strong version of the pure serial recall hypothesis predicts that long-term or lexical effects will *never* be seen when a small, closed pool is used. That is, it predicts that null results will always be obtained. Because of this strong claim, a single positive instance (or in the case of this article, two instances) is sufficient to challenge the claim. A weaker version of the hypothesis might predict merely a reduced effect rather than the absence of an effect, but this also has the consequence of no longer being the pure serial recall hypothesis. Rather, it has changed to a qualitatively different hypothesis that now acknowledges the impurity of the task.

The impurity principle states that because people can potentially use any useful information or processes to help them remember, tasks are not pure. One consequence of this is that the impurity principle predicts that long-term memory and lexical factors can affect immediate serial recall, even when a small closed set is used. One possible problem with assessing this prediction is the construction of the closed stimulus pool. With a pool of only six items, it can be difficult to establish that the words are either statistically identical or statistically different on the dimensions of interest, given the very small number of items compared. Even if this is possible, one of the six items might be unusual or differ in some way that affects the overall results. To minimize the chance of this happening, experiments should use a different randomly selected, small, closed set of items for each subject. In addition, in the larger pool, all items of one type should be higher on the critical dimension (and preferably on multiple measures of the same dimension) than all items of the other type, while still being equated overall on all other dimensions that are likely to affect performance.

Although the proposition is beyond the scope of the present article, we suggest that the impurity principle applies more widely than just to memory. Just as cognitive processes have long been seen as constructive and reconstructive (e.g., Neisser, 1967), they are also subject to impurity. Put another way, if a task as apparently simple as recalling six items in order is not pure, then tasks or processes measuring far more complex concepts, such as executive function or inhibition or intelligence, can hardly be pure, either.

At least as far back as 1885, Ebbinghaus acknowledged that contributions from previous experiences could not be avoided. Despite the fact that early work on span had found the same long-term and lexical influence on performance, the idea of pure memory, whether a pure task or a pure process, has been proposed at various times. The present results add even more data against this recurring idea of pure memory (Crowder, 1993; Restle, 1974).

Author note This research was supported, in part, by grants from the Natural Sciences and Engineering Research Council to each author.

Appendix 1: Descriptive information about the stimuli used in the four experiments

Abstract and concrete one-syllable words used in Experiments 1 and 2

	CNC	FAM	IMG	NLET	NPHN	FREQ	ORTH	OrthZ	OrthF	LgWF	LgCD	CncM	OLD	OLDF	PLD	PLDF
Abstract Words																
Mean	341.68	525.73	410.98	4.42	3.35	75.76	7.64	0.09	92.72	3.16	2.93	2.77	1.53	8.21	1.26	8.55
SD	37.96	51.48	50.85	0.75	0.70	84.86	5.47	0.80	266.94	0.75	0.64	0.61	0.30	0.51	0.29	0.74
Min	234	320	282	3	1	0	0	– 1.23	0	1.43	1.26	1.25	1	6.85	1	6.72
Max	399	600	548	6	5	388.9	24	2.77	2,314.99	4.80	3.90	3.97	2.45	10.02	2.2	10.46
Concrete Words																
Mean	559.11	530.98	563.18	4.35	3.39	68.36	8.14	0.14	73.02	3.18	2.89	4.64	1.51	8.19	1.26	8.53
SD	51.86	54.13	48.94	0.82	0.74	122.07	5.67	0.81	138.64	0.59	0.50	0.29	0.31	0.52	0.30	0.83
Min	408	300	383	3	2	0.54	0	– 1.41	0	1.71	1.54	4	1	6.94	1	6.40
Max	646	600	667	6	5	1,235.84	24	2.41	1,151.71	5.01	3.912	5	2.45	9.60	2.2	11.14
t	**47.36**	0.98	**30.19**	0.84	0.56	0.70	0.88	0.63	0.92	0.21	0.60	**38.52**	0.62	0.27	0.06	0.25
p	**.00**	.33	**.00**	.40	.58	.49	.38	.53	.36	.83	.55	**.00**	.54	.79	.95	.80

Low- and high-frequency one-syllable words used in Experiments 3 and 4

	CNC	FAM	IMG	NLET	NPHN	FREQ	ORTH	OrthZ	OrthF	LgWF	LgCD	CncM	OLD	OLDF	PLD	PLDF
Low-Frequency Words																
Mean	504.42	515.14	522.15	4.19	3.31	5.16	8.78	0.18	76.07	2.29	2.13	4.23	1.49	8.19	1.27	8.44
SD	93.85	32.56	73.15	0.72	0.63	2.60	5.94	0.85	220.68	0.36	0.33	0.76	0.27	0.53	0.27	0.75
Min	262	430	347	3	2	0.59	0	– 1.51	0	1.32	1.26	1.53	1	6.71	1	6.96
Max	634	586	659	5	4	9.94	26	3.13	2,184.92	2.79	2.68	5	1.95	10.02	1.9	10.43
High-Frequency Words																
Mean	504.71	518.71	522.42	4.25	3.33	60.36	8.55	0.16	65.30	3.31	3.02	4.15	1.46	8.22	1.24	8.53
SD	96.45	29.97	75.95	0.66	0.60	82.67	5.63	0.78	108.20	0.33	0.29	0.86	0.28	0.52	0.28	0.77
Min	204	453	302	3	2	10.17	0	– 1.41	0	2.90	2.49	1.55	1	6.96	1	6.98
Max	627	645	635	5	4	640.68	23	2.12	844.39	4.45	3.82	5	2	10.36	1.9	10.67
t	0.02	0.88	0.03	0.66	0.32	**7.25**	0.30	0.13	0.48	**22.46**	**22.23**	0.82	0.61	0.44	0.82	0.90
p	.98	.38	.98	.51	.75	**.00**	.76	.90	.63	**.00**	**.00**	.42	.54	.66	.42	.37

CNC = concreteness, FAM = familiarity, IMG = imageability, NLET = number of letters, NPHN = number of phonemes (from Coltheart, 1981). FREQ = Celex frequency, ORTH = number of orthographic neighbors, OrthF = frequency of the neighbors (from Medler & Binder, 2005). OrthZ is a *z*-score based on ORTH that removes word length as a confounding factor (see Storkel, 2004). LgWF = log frequency, LgCD = log contextual diversity (from Brysbaert & New, 2009). CncM = mean concreteness (from Brysbaert, Warriner, & Kuperman, 2014). OLD = mean Levenshtein distance for the 20 closest orthographic neighbors; OLDF = frequency of the 20 closest orthographic neighbors; PLD = same as OLD, except for phonological neighbors; PLDF = same as OLDF, except for phonological neighbors (from Yarkoni, Balota, & Yap, 2008, via Balota et al., 2007). The bottom two rows show the *t* test and resulting *p* values comparing the abstract and concrete words. The full set of stimuli is available at https://memory.psych.mun.ca/research/stimuli or from the first author.

Appendix 2

All analyses in the article are from uncorrected responses, and therefore intrusion errors include typographical and spelling errors. One reason for analyzing the uncorrected data was that many responses were difficult to interpret: They could be the result of a typing error, or they could simply be the wrong word. A second reason is that correcting the responses would necessarily reduce the number of errors; as such, any correction procedure would be biased against the pure serial recall hypothesis and in favor of the impurity principle.

To assess the effect of not correcting spelling and typing errors, the responses in Experiment 1 were spell-checked. Ambiguous responses were corrected to the first suggestion from the built-in spell checker, except where an adjacent key on the keyboard or the addition of a letter could make the response a valid word. This resulted in a change of 1.15% of responses (132 out of 11,520). The table below shows the changes in the number of response types (raw data minus spell-checked data).

			Errors				
			Intrusion	Omission	Repetition	Position	Correct
Closed	First Half	Abstract	– 5	0	1	0	4
		Concrete	– 12	0	2	0	10
	Second Half	Abstract	– 12	0	1	4	7
		Concrete	– 14	0	1	0	13
Open	First Half	Abstract	– 27	0	3	0	24
		Concrete	– 17	0	0	6	11
	Second Half	Abstract	– 19	0	1	1	17
		Concrete	– 25	0	2	– 1	24

As can be seen, correcting spelling decreased the number of intrusion errors and increased the proportion correct. There was, of course, no effect on omission errors.

A 2 Set Size (open vs. closed) × 2 Word Type (abstract vs. concrete) × 2 Experiment Half (first half vs. second half) × 6 Serial Position mixed factorial ANOVA on the spell-checked accuracy data yielded the same pattern of results as the one reported in the main article. The only potentially important change was the experiment half by word type by set size interaction, $F(1, 58) = 3.906$, $MSE = .046$, $\eta_p^2 = .063$, $p = .053$; in the raw data, $F(1, 58) = 2.875$ and $p = .095$. In both cases, the interaction reflects a pattern in which there was a larger improvement in recall for abstract than for concrete words from the first to the second half of the experiment in the closed group, but a larger improvement for concrete than for abstract words between halves in the open group. Thus, even if this interaction were significant, it would not contradict the conclusions reported in the main article.

All other main effects and interactions remained essentially the same. All main effects were still significant. For the two-way interactions, word type by set size remained $F(1, 58) < 1$. Experiment half by position remained significant, $F(5, 290) = 7.669$, $MSE = 0.017$, $\eta_p^2 = .117$, $p < .001$. The results for the remaining two-way interactions were experiment half by set size, $F(1, 58) = 2.563$, $MSE = .048$, $\eta_p^2 = .042$, $p > .10$; experiment half by word type, $F(1, 58) < 1$; word type by position, $F(5, 290) < 1$; and position by set size, $F(5, 290) = 2.172$, $MSE = .067$, $\eta_p^2 = .036$, $p = .057$. Other than the one three-way interaction noted above, the rest remained $F < 1$, as did the four-way interaction.

The conclusions drawn from the spell-checked analysis and the analysis on the uncorrected responses were identical for all critical comparisons. Because of this, and the two reasons noted earlier, all analyses in the main article are on the uncorrected responses.

References

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

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459. https://doi.org/10.3758/BF03193014

Blankenship, A. B. (1938). Memory span: A review of the literature. *Psychological Bulletin*, *35*, 1–25. https://doi.org/10.1037/h0061086.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–900. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Warriner, A.B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904–911. https://doi.org/10.3758/s13428-013-0403-5

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, *33A*, 497–505. https://doi.org/10.1080/14640748108400805

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62–101). New York, NY: Cambridge University Press.

Crowder, R. G. (1976). *Principles of learning and memory.* Hillsdale, NJ: Erlbaum.

Crowder, R. G. (1993). Systems and principles in memory theory: Another critique of pure memory. In A. F. Collins, S. E. Gathercole, M. A. Conway & P. E. Morris (Eds.), *Theories of memory.* (pp. 139–161). Hillsdale, NJ: Erlbaum.

Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Leipzig, Germany: Duncker & Humboldt.

Hughes, R. W., Chamberland, C., Tremblay, S., & Jones, D. M. (2016). Perceptual–motor determinants of auditory–verbal serial short-term memory. *Journal of Memory and Language*, *90*, 126–146. https://doi.org/10.1016/j.jml.2016.04.006

Hughes, R. W., Marsh, J. E., & Jones, D. M. (2009). Perceptual–gestural (mis)mapping in serial short-term memory: The impact of talker variability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1411–1425. https://doi.org/10.1037/a0017008

Jacobs, J. (1887). Experiments on "prehension" *Mind*, *12*, 75–79.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*, 513–541. https://doi.org/10.1016/0749-596X(91)90025-F

Jalbert, A., Neath, I., & Surprenant, A. M. (2011). Does length or neighborhood size cause the word length effect? *Memory & Cognition*, *39*, 1198–1210. https://doi.org/10.3758/s13421-011-0094-z

Jones, D. M., & Macken, B. (2018). In the beginning was the deed: Verbal short-term memory as object-oriented action. *Current Directions in Psychological Science*, *27*, 351–356. https://doi.org/10.1177/0963721418765796

Jones, G. V. (1988). Images, predicates, and retrieval cues. In M. Denis, J. Englekamp, & J. T. E. Richardson (Eds.), *Cognitive and neuropsychological approaches to mental imagery* (pp. 89–98). Dordrecht, The Netherlands: Martinus Nijhoff.

Kolers, P. A., & Roediger, H. L. III. (1984). Procedures of mind. *Journal of Verbal Learning and Verbal Behavior*, *23*, 425–449. https://doi.org/10.1016/S0022-5371(84)90282-2

Lin, Y., Chen, H., Lai, Y. C., & Wu, D. H. (2015). Phonological similarity and orthographic similarity affect probed serial recall of Chinese characters. *Memory & Cognition*, *43*, 538–554. https://doi.org/10.3758/s13421-014-0495-x

Marschark, M., & Hunt, R. R. (1989). A reexamination of the role of imagery in learning and memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 710–720. https://doi.org/10.1037/0278-7393.15.4.710

Medler, D. A., & Binder, J. R. (2005). MCWord: An on-line orthographic database of the English language. Madison, WI: Medical College of Wisconsin, Language Imaging Laboratory. Retrieved from www.neuro.mcw.edu/mcword/

Monnier, C., & Syssau, A. (2008). Semantic contribution to verbal short-term memory: Are pleasant words easier to remember than neutral words in serial recall and serial recognition? *Memory & Cognition*, *36*, 35–42. https://doi.org/10.3758/MC.36.1.35

Morin, C., Poirier, M., Fortin, C., & Hulme, C. (2006). Word frequency and the mixed-list paradox in immediate and delayed serial recall. *Psychonomic Bulletin & Review*, *13*, 724–729. https://doi.org/10.3758/BF03193987

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

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

Neisser, U. (1967). *Cognitive psychology*. New York, NY: Appleton-Century-Crofts.

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

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology*, *45*, 255–287. https://doi.org/10.1037/h0084295

Quinlan, P. T., Roodenrys, S., & Miller, L. M. (2017). Serial reconstruction of order and serial recall in verbal short-term memory. *Memory & Cognition*, *45*, 1126–1143. https://doi.org/10.3758/s13421-017-0719-y

Restle, F. (1974). Critique of pure memory. In R. L. Solso (Ed.), *Theories of cognitive psychology: The Loyola Symposium* (pp. 203–217). Potomac, MD: Erlbaum.

Roodenrys, S., Hulme, C., Lethbridge, A., Hinton, M., & Nimmo, L. M. (2002). Word-frequency and phonological-neighborhood effects on verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 1019–1034. https://doi.org/10.1037/0278-7393.28.6.1019

Roodenrys, S., & Quinlan, P. T. (2000). The effects of stimulus set size and word frequency on verbal serial recall. *Memory*, *8*, 71–78. https://doi.org/10.1080/096582100387623

Ruiz-Vargas, J. M., Cuevas, I., & Marschark, M. (1996). The effects of concreteness on memory: Dual codes or dual processing? *European Journal of Cognitive Psychology*, *8*, 45–72. https://doi.org/10.1080/095414496383202

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*, 171–177. https://doi.org/10.3758/BF03196364

Saint-Aubin, J., & Poirier, M. (2005). Word frequency effects in immediate serial recall: Item familiarity and item co-occurrence have the same effect. *Memory*, *13*, 325–332. https://doi.org/10.1080/09658210344000369

Schwanenflugel, P. (1991). Why are abstract concepts hard to understand? In P. J. Schwanenflugel (Ed.), *The psychology of word meanings* (pp. 223–250). Hillsdale, NJ: Erlbaum.

Storkel, H. L. (2004). Methods for minimizing the confounding effects of word length in the analysis of phonotactic probability and neighborhood density. *Journal of Speech, Language, and Hearing Research*, *47*, 1454–1468. https://doi.org/10.1044/1092-4388(2004/108)

Stuart, G., & Hulme, C. (2000). The effects of word co-occurrence on short-term memory: Associative links in long-term memory affect short-term memory performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 796–802. https://doi.org/10.1037/0278-7393.26.3.796

Surprenant, A. M., & Neath, I. (2009a). The 9 lives of short-term memory. In A. Thorn & M. Page (Eds.), *Interactions between short-term and long-term memory in the verbal domain* (pp. 16–43). Hove, UK: Psychology Press.

Surprenant, A. M., & Neath, I. (2009b). *Principles of memory*. New York, NY: Psychology Press.

Walker, I., & Hulme, C. (1999). Concrete words are easier to recall than abstract words: Evidence for a semantic contribution to short-term serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *2*, 1256–1271. https://doi.org/10.1037/0278-7393.25.5.1256

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*, 971–979. https://doi.org/10.3758/PBR.15.5.971

Yuille, J. C., & Paivio, A. (1968). Imagery and verbal mediation instructions in paired-associate learning. *Journal of Experimental Psychology*, *78*, 436–441. https://doi.org/10.1037/h0026457