Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale
2016, Vol. 70, No. 2, 125–138

1196-1961/16/$12.00 http://dx.doi.org/10.1037/cep0000088

Order Information is Used to Guide Recall of Long Lists: Further Evidence for the Item-Order Account

Noah D. Forrin and Colin M. MacLeod
University of Waterloo

Differences in memory for item order have been used to explain the absence of between-subjects (i.e., pure-list) effects in free recall for several encoding techniques, including the production effect, the finding that reading aloud benefits memory compared with reading silently. Notably, however, evidence in support of the item-order account (Nairne, Riegler, & Serra, 1991) has derived primarily from short-list paradigms. We provide novel evidence that the item-order account also applies when recalling long lists. In Experiment 1, participants studied and then free recalled 3 different long lists of words: pure aloud, pure silent, and mixed (half aloud, half silent). A Bayesian analysis supported a null pure-list production effect, and subsequent order analyses were largely consistent with the item-order account. These findings indicate that order information is retained in long-term memory and is useful in guiding subsequent free recall. In Experiment 2, a distractor task was inserted between the study and test phases, ensuring that only long-term memory processes were involved in recall: The pattern of results remained consistent with the item-order account. Order information can be retained in long-term memory for long lists, and is useful in guiding subsequent free recall, extending the domain of the item-order account.

Keywords: production effect, memory, recall, item-order account

Supplemental materials: http://dx.doi.org/10.1037/cep0000088.supp

There is a mnemonic benefit of reading aloud over reading silently. First demonstrated experimentally by Hopkins and Edwards (1972), this phenomenon reappeared only periodically over the intervening decades (Conway & Gathercole, 1987; Gathercole & Conway, 1988; MacDonald & MacLeod, 1998) but has received increased attention since MacLeod, Gopie, Hourihan, Neary, and Ozubko (2010) named it the production effect and outlined some of its boundaries.

The production effect in recognition has been documented extensively using within-subject designs (e.g., MacLeod, Gopie, Hourihan, Neary, & Ozubko, 2010), and more recently has been shown to be reliable in between-subjects designs (Fawcett, 2013). In stark contrast, evidence of a production effect in recall has only been found using within-subject designs (Conway & Gathercole, 1987; Cho & Feldman, 2013; Jones & Pyc, 2014; Jonker, Levene, & MacLeod, 2014; Lin & MacLeod, 2012), and not using between-subjects designs. For example, Jones and Pyc (2014) found evidence of a within-subject effect but not of a between-subjects effect in recall. However, they also cautioned that more research would be required before a significant between-subjects production effect in recall could be ruled out conclusively.

Our goal here is to answer their call by conducting experiments with ample statistical power to detect a possible benefit of production in recall. In particular, we examine whether memory for order information—a factor that Jonker, Levene, and Macleod (2014) found influenced the production effect in the free recall of short lists—also plays a key role for long lists where recall cannot rely solely on working memory. We discuss the relevance of order information later in this introduction. First, though, we recap production research that has used recognition tests, as factors found to underlie the production effect in recognition might also be expected to influence the production effect in recall.

The Production Effect in Recognition

Across several recognition experiments, MacLeod et al. (2010) found robust within-subject production effects when "aloud" and "silent" words were studied in a mixed list. However, they did not find any evidence of a pure-list between-subjects production effect (their Experiments 2 and 3; see also Hopkins & Edwards, 1972). The apparent absence of a between-subjects effect led these authors to contend that distinctiveness (Hunt, 2006, 2013; Hunt & Worthen, 2006) is a key component of the production effect.[1] According to Hunt (2006), features that stand out as different in the context of

Noah D. Forrin and Colin M. MacLeod, Department of Psychology, University of Waterloo.

This research was supported by Natural Sciences and Engineering Research Council of Canada Discovery Grant A7459. We thank, Emily Cyr, Sukhdip Grewal, Anaum Nawaz, and Deanna Priori for their assistance in collecting the data.

Correspondence concerning this article should be addressed to Noah D. Forrin, Department of Psychology, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1. E-mail: nforrin@uwaterloo.ca

[1] The memory benefits of bizarreness (McDaniel & Einstein, 1986), generation (Begg, Snider, Foley, & Goddard, 1989), and enactment (Engelkamp & Zimmer, 1997) have also been explained in terms of distinctiveness, and Conway and Gathercole (1987) also advocated a distinctiveness account in their research on production.

similarity are distinctively processed. In a mixed-list production experiment, the similar context is the lexical processing that encompasses all reading, whether done aloud or silently. But the auditory and articulatory processes involved in reading aloud stand out as distinct *relative* to this baseline. Conversely, in a pure-aloud list, aloud words do not benefit from relational (i.e., relative) distinctiveness because audition and articulation no longer stand out as unique processing dimensions.

A recent meta-analysis by Fawcett (2013) revealed, however, that there is in fact a modest ($g = .37$) between-subjects (i.e., pure-list) production effect, a conclusion bolstered by Bodner, Taikh, and Fawcett (2014). This significant pure-list effect made apparent that the relationally distinct processing of aloud information at study likely is not the only factor driving the production effect. Notably, memory for produced words may also be bolstered by distinctiveness at the time of retrieval. Participants may use a distinctiveness heuristic (Schacter, Israel, & Racine, 1999) at test, whereby they strategically attempt to recollect distinct speech information to determine whether a word was studied (Dodson & Schacter, 2001). Retrieval of such speech information is consistent with an item having been studied; failure to retrieve such information leaves open the question of whether an item was previously studied.

Dodson and Schacter (2001) demonstrated that the distinctiveness heuristic can help to lower individuals' false alarm rates to silent lures on a recognition test (i.e., "If I can't remember saying it, I must not have studied it"). Moreover, MacLeod and colleagues (MacLeod et al., 2010; Ozubko & MacLeod, 2010; Ozubko, Major, & MacLeod, 2014) have reported evidence that individuals use a distinctiveness heuristic at test to confirm that words were studied, thereby boosting the hit rates of aloud items. The distinctiveness heuristic is a viable strategy following a pure-aloud list because aloud target information stands out as distinct—and diagnostic of study—against a backdrop of silent lures. But participants may be even more likely to use a distinctiveness heuristic following a mixed study list because the speech already stands out as relationally distinct at the time of study due to being intermixed with silent items. This may increase the likelihood of encoding a record of speech that could then be strategically retrieved at test.

Differing memory strength (Wickelgren, 1969) may also contribute to the production effect: Reading aloud may establish a stronger memory record than does reading silently. Bodner and Taikh (2012; see also Taikh & Bodner, 2016) suggest that there may be a strength-based decision process, analogous to the distinctiveness heuristic, whereby participants evaluate the strength of words on the recognition test, judging those with high strength to have been studied, and hence favoring the stronger (aloud) items. There is, however, evidence that conflicts with this idea: When Ozubko, Major, and MacLeod (2014) equated the strength of aloud and silent items by repeating the silent items, participants still could discriminate aloud and silent items in a study mode judgment task.

Recent research has directly compared the mixed-list and pure-list designs with the goal of shedding light on the factors underlying the production effect in recognition (Bodner, Taikh, & Fawcett, 2014; Forrin, Groot, & MacLeod, 2016). In these experiments, aloud and silent pure-list conditions served as baselines against which mixed-list performance was compared. A memorial benefit of reading aloud in a mixed list versus a pure list would be consistent with relational distinctiveness enhancing memory for aloud words in a mixed list. Conversely, a cost to reading silently in a mixed list versus a pure list would suggest that lazy reading (cf. Begg & Snider, 1987) impaired memory for silent words in a mixed list: In emphasizing the aloud words, participants reduced their attention to silent words.

In terms of hit rates, Bodner et al. (2014) results showed both benefits and costs to reading aloud in a mixed list. However, in terms of d' (a measure of discrimination that incorporates both hits and false alarms) only the costs were apparent. But Forrin, Groot, and MacLeod (2016) have since suggested that this pattern hinged on using a single false alarm rate for both the aloud and silent conditions. When Forrin et al. (2016) revised the Bodner et al. (2014) procedure to allow separate false alarm rates to be obtained for the aloud items and the silent items in the mixed study list, they observed both benefits and costs for d' as well as for hit rates. On this basis, Forrin et al. (2016) suggested that both relational distinctiveness (benefit) and lazy reading (cost) may contribute to the robust mixed-list production effect.

In sum, a rather wider array of mechanisms may underlie the production effect as measured by recognition, despite it appearing to be such a straightforward mnemonic. Both the distinctiveness heuristic and (evaluated) strength may enhance memory for aloud words in both mixed-list and pure-list designs. In mixed lists, relational distinctiveness may provide a further benefit to aloud items and lazy reading may impose a cost on silent items. But what about recall?

The Production Effect in Recall

A number of recent studies have explored the locus of the production effect in recall (Lambert, Bodner, & Taikh, 2016; Jones & Pyc, 2014; Jonker et al., 2014). Although the aforementioned distinctiveness and strength accounts may also help to explain the production effect in recall, Jonker et al. (2014) have argued that another factor may play a principal role: item order. Their proposal derives from the item-order account of design effects, originally proposed by Nairne, Riegler, and Serra (1991) as an explanation for a null between-subjects generation effect in recall that had been reported by several researchers (e.g., Begg & Snider, 1987; Hirshman & Bjork, 1988; McDaniel, Waddill, & Einstein, 1988; Slamecka & Katsaiti, 1987).

Nairne et al. (1991) posited that the item elaboration invoked by generation disrupts the processing of order information—a type of relational information that has been found to facilitate recall from long-term memory (e.g., Nairne & Kelley, 2004; Postman, 1972). Nairne et al. (1991) had participants study short pure lists of words either by generating them or by reading them aloud, and replicated the previously reported null pure-list generation effect. Consistent with their item-order account, they observed superior recognition for the generated lists, but superior order retention for the lists read aloud. Taken together, these results suggested that for pure lists the item-elaboration advantage afforded by the unusual encoding condition—generation—was offset by the superior order processing involved in the common encoding condition—reading aloud (see also Burns, 1996; Burns, Curti, & Lavin, 1993; Mulligan, 2002; Serra & Nairne, 1993).

In a review article, McDaniel and Bugg (2008) extended the item-order account. They contended that unusual encoding tasks—like generation (Slamecka & Graf, 1978) and enactment (Engelkamp & Krumnacker, 1980)—enhance memory for individual items. Conversely, common encoding tasks such as silently reading tend to enhance order information. Another prediction stemming from the item-order account is that there should be a cost to recalling commonly encoded items in a mixed list as opposed to a pure list

(McDaniel & Bugg, 2008). This cost should arise because the order information for common items will be disrupted in a mixed list by the intermittent encoding of unusual items. McDaniel and Bugg (2008) outlined an array of memory phenomena that have shown an empirical pattern consistent with the item-order account (i.e., the lack of a pure list effect coupled with the mixed list cost to the common condition). These phenomena include generation (Hirshman & Bjork, 1988), enactment (Engelkamp & Dehn, 2000), bizarreness (McDaniel, DeLosh, & Merritt, 2000), word frequency (DeLosh & McDaniel, 1996), orthographic distinctiveness (McDaniel, Cahill, Bugg, & Meadow, 2011), and perceptual interference (Mulligan, 1999).

Recently, Jonker et al. (2014) posited that the production effect belongs to this group of phenomena. In two experiments, they employed a mixed versus pure list design in which participants studied several short, eight-item lists. Using an order reconstruction test, they found better order memory for words studied in pure-silent lists compared to words studied in either pure-aloud lists or mixed lists (Experiment 1). Moreover, they found both costs and small benefits to mixed-list production as well as a null pure-list effect (Experiment 2). From these findings, Jonker et al. (2014) argued that the superior order memory for silent items in pure lists was responsible for their pattern of recall data—particularly the costs to the mixed silent condition and the null pure list effect—entirely in keeping with the item-order account.

The absence of a pure-list production effect that Jonker et al. (2014) observed stands in contrast to the significant pure-list effect found in recent recognition experiments (e.g., Bodner et al., 2014; Forrin et al., 2016). Jonker et al. (2014) contended that this difference emerged because order information typically is not useful on a recognition test where the experimenter controls the order in which test items are presented. Thus, order information enhances the memory of the pure-silent condition when the test is recall, but not when it is recognition.[2]

There is one important caveat, however, to this item-order research (including Jonker et al., 2014): It ordinarily has involved the study of short lists, whereas the phenomena it aims to explain typically involve the study of long lists. This invites the question: Would the pattern of production recall data obtained by Jonker et al. (2014) replicate if participants studied a long list rather than a series of short lists? The Jones and Pyc (2014) study partially addresses this question in two experiments in which participants studied 40 words in either a pure-aloud list, a pure-silent list, or a mixed list, prior to a free recall test. Consistent with Jonker et al. (2014)—and with the item-order account—Jones and Pyc (2014) found a recall cost to having read silent items in a mixed list as well as no difference between pure-aloud and pure-silent lists.

Jones and Pyc's (2014) results suggest that production is ineffective in boosting recall. Although they noted that their data were consistent with the item-order account, they did not examine whether participants in their pure-silent condition used order information to a greater extent than did participants in their pure-aloud and mixed conditions—a result that would have substantiated the item-order account. The purpose of our present research, therefore, was to determine whether the pattern of results obtained by Jones and Pyc (2014), if replicated, could be explained by the item-order account. In doing so, we examined a question that has received scant research attention: Namely, does the item-order account apply to the study of long lists?

Is Order Information Retained From Studying Long lists?

We are aware of only one experiment that directly tested the item-order account using long lists. Burns (1996) had participants study long, pure lists of words (40 in his Experiment 1, and 32 in Experiments 2 and 3). All words in a list were either read aloud (a relatively common process) or generated (a relatively uncommon process). Study was followed by a distraction phase and then free recall and/or an order reconstruction task. Across three experiments that assessed order reconstruction and order retention in free recall, Burns found consistent evidence that participants had superior order memory after studying pure aloud lists versus generated lists, even when they did not anticipate a recall test (Experiment 3).

In short, consistent with the item-order account, Burns (1996) demonstrated that participants have superior order retention for relatively more "common" (read aloud) than "uncommon" (generated) pure-lists of words, using a traditional long-list free recall paradigm, replicating the results of experiments that used short lists (e.g., Nairne et al., 1991). Given Burns' results that individuals retain order from long pure-aloud lists, we also expected to find that individuals retain order from long pure-silent lists. Silent reading is a more common type of encoding that entails less item-specific processing than does reading aloud and ought to, in line with the item-order account, encourage an even greater degree of order encoding (as was the case in the short-list production research of Jonker et al., 2014).

Although Burns (1996; see also Kahana, 1996) found evidence that order information is retained after studying long lists, he did not directly test whether participants used order information to facilitate their free recall. This relation between the retention of order information and free recall was the focus on research by Mulligan and Lozito (2007), who reasoned that significant correlations must be shown between measures of order retention and free recall to more conclusively demonstrate that participants are harnessing order information to help guide their free recall. Mulligan and Lozito (2007) had participants study multiple word lists of varying lengths—either eight words (their Experiment 1A), 16 words (Experiment 2A), or 24 words (Experiment 2B)—followed by a distraction phase and free recall. They found a significant positive correlation between the order measure of input-output correspondence (I-O correspondence; Asch & Ebenholtz, 1962) and free recall only for the short, eight-word lists. This correlation was progressively smaller as list length increased, and was nonsignificant for the 16- and 32-word lists.[3]

Thus, Mulligan and Lozito's (2007) results suggest that individuals' reliance on order information systematically decreases as list-length increases (see also Jahnke, 1965; Mandler

[2] Indeed, as Nairne et al. (1991) argued in their original formulation of the item-order account, recognition performance is largely driven by item-specific information, which explains why several mnemonics show pure list effects in recognition—including generation (e.g., Serra & Nairne, 1993), enactment (e.g., Engelkamp & Dehn, 2000), and production (Fawcett, 2013).

[3] These correlations were still in the expected positive direction, and did reach significance for a group of participants who studied 16-word lists and were instructed to focus on order information during study.

& Dean, 1969), suggesting that Jones and Pyc's (2014) pattern of results may not, in fact, be reflective of the item-order account. With this in mind, it is worth noting that Mulligan and Lozito (2007)—like Burns (1996)—found that participants who studied the long lists of words still retained order information at greater-than-chance levels. Where their results diverged from Burns (1996) was in showing that order information appeared to be largely ineffective for improving free recall following long-list study. One possibility is that participants who studied long lists relied more on other memory strategies, which overwhelmed the influence of order information.

Another possibility, however—investigated in our experiments—was that the I-O correspondence measure is not very sensitive to the types of order information that individuals rely on when studying long lists of words. Perhaps individuals focus more on recalling clusters of words (regardless of the direction in which they were encoded) when studying long lists. If this were the case, other order measures such as the bidirectional pair frequency measure (Anderson & Watts, 1969; see also Sternberg & Tulving, 1977) used by Burns (1996) or the study distance measure used by Jonker et al. (2014; see Kahana, 1996) may be more sensitive for measuring participants' reliance on types of order information that aid their subsequent recall. For this reason, we analysed our recall data in terms of all three of these order measures (which we describe in the Method section of our Experiment 1).

The Present Research

We sought to replicate Jones and Pyc's (2014) results. Like them, we employed a mixed versus pure list design in which participants studied long word lists. Our major goal was to directly assess whether our results were consistent with the item-order account. We analysed the extent to which participants' item recall order corresponded to the order in which the items were presented at study. A higher correspondence in the pure-silent condition relative to the pure-aloud or mixed conditions would constitute evidence for the item-order account.

Although we expected to find costs to reading silently in a mixed list—both Jonker et al. (2014) and Jones and Pyc (2014) had found robust costs—we were less certain of whether we would find benefits. Jones and Pyc did not find a significant benefit, but in both of their experiments there was a trend in this direction; moreover, combining and analysing the data from both of their experiments revealed a significant benefit, $t(76) = 2.21$, $p = .03$, $d = 0.52$. Jonker et al. (2014, Experiment 2) found a significant benefit. A significant mixed-list benefit in recall would also be consistent with the item-order account. According to McDaniel and Bugg (2008; pp. 239–240), order memory for unusual items is better in a mixed list than in a pure list, due to the common items in the mixed list promoting relational encoding (though not to the same extent as in a pure common list). To the extent that aloud items may be encoded as relatively "unusual" in a mixed list, and silent items as comparatively "common," the same pattern may emerge for the production effect. Notably, however, Jonker et al. (2014) did not find evidence of superior order encoding for mixed-aloud versus pure-aloud items—though the rest of their order memory results were consistent with the item-order account—casting some doubt on this prediction.

We were particularly interested in examining whether a pure-list production effect would emerge in recall. Although two previous studies have shown nonsignificant pure-list effects (Jones & Pyc, 2014; Jonker et al., 2014), it would perhaps be premature to conclude that there is no pure-list production effect in recall based on these two papers alone. After all, we now know that there is a reliable between-subjects production effect in recognition (Fawcett, 2013), contrary to early production experiments that did not find evidence of this effect (Hopkins & Edwards, 1972; MacLeod et al., 2010). With this in mind, we aimed for high statistical power in the present experiments by running large samples (120 participants in Experiment 1 and 150 participants in Experiment 2) in a within-subject, blocked design (see the Method section). A post hoc power analysis using the statistical software G*Power (Erdfelder, Faul, & Buchner, 1996) showed that our experimental design had high statistical power (.95 in Experiment 1 and .98 in Experiment 2) to detect even a small effect (Cohen's $d = .3$). We also used a Bayesian analysis to test for null effects.

Experiment 1 used a free recall test that immediately followed study (i.e., without an intervening distractor phase). Our rationale here was to precisely replicate the mixed- versus pure-list study design that we have used previously for recognition memory experiments (Forrin et al., 2016), including the same blocked design, word list, and immediate test. Controlling for these factors meant that the different pattern of results obtained here could be attributed to the nature of the memory test (recall vs. recognition), as opposed to the addition of a distractor task. Experiment 2 was a conceptual replication in which we implemented a distractor phase between the study phase and the free recall test—following the convention with research testing the item-order account (e.g., Burns, 1996; Burns et al., 1993; Mulligan, 2002; Mulligan & Lozito, 2007; Nairne et al., 1991; Serra & Nairne, 1993)—to minimize the influence of short-term memory (STM) on free recall.

To foreshadow, both experiments replicated Jones and Pyc's (2014) pattern of recall. We found a cost to reading silently in a mixed list versus a pure list, as well as a nonsignificant difference between pure-aloud and pure-silent lists. Extending Jones and Pyc's (2014) results, our order analyses also supported their assumption that the item-order account can explain the production effect in long list free recall. However, our order results differed from those obtained Lambert, Bodner, and Taikh (2016), who also examined the production effect in long list recall. They obtained a comparable pattern of results for free recall (cf. their Experiments 1 and 4), but did not find evidence of an order memory advantage for pure-silent lists. We will discuss these diverging results in our General Discussion.

Experiment 1

Method

Participants. One-hundred and 20 undergraduate participants from the University of Waterloo completed this experiment in exchange for course credit.

Stimuli. The word pool comprised 240 words obtained from the MRC Psycholinguistic Database. The words were five to 10

letters long and had frequencies of greater than 30 per million (Thorndike & Lorge, 1944). This word pool was identical to that used in Experiments 1 and 2 of Forrin et al. (2016).

Apparatus. Stimuli were displayed on a 17" LCD monitor and presented using E-Prime software (Psychology Software Tools Inc., Pittsburgh, PA).

Procedure. We employed a within-subject blocked design in which participants studied three different 40-word lists—pure aloud, pure silent, and mixed—in a counterbalanced order. This resulted in 20 participants being randomly assigned to each of the six block orders. Each study list was immediately followed by a free recall test. Although told that a memory test would follow each study list, participants were not told specifically that it would be a free recall test. After being given the free recall test following the first study list, however, they may have correctly inferred that free recall tests would also follow the subsequent two study lists. Thus, participants could have employed different strategies while studying the second and third lists—strategies that could have aided their subsequent recall. We examine this possibility in our analysis of block order effects (see Appendix 1 of the Online Supplemental section).

In the mixed study list, font color indicated whether each word should be read aloud (blue font) or silently (white font). Previous research (e.g., Bodner et al., 2014; MacLeod et al., 2010) has demonstrated that font colour and study modality do not interact, so these factors were not counterbalanced here. Each word was presented in 16 point Courier New lowercase font, centered against a black background. Presentation duration was 3 s and there was a 500-ms interstimulus interval.

A self-paced free recall test immediately followed each study list. Participants recalled by typing each word and pressing ENTER after it. When they pressed ENTER, the word that they had just typed disappeared from the screen. Participants were told that they could take as long as they wanted and that there was no penalty for typing incorrect words. They indicated when they were finished recalling words by typing the word "finished," at which point they proceeded immediately to the next study list.

Results

Overview. The within-subject blocked design used in this experiment accords a high level of statistical power. A blocked design is also typically used in short-list item-order research, including research involving the production effect (Jonker et al., 2014). However, a drawback of this design is possible block-order effects. Therefore, block-order analyses are presented and interpreted in Appendix 1 of the Online Supplement.

To address potential concerns regarding block-order effects, we also conducted first-block-only analyses of our free recall, intrusion, and order retention data. Because these analyses involved only the first block of each participant's data, they could not have been contaminated by block order effects. We present these first-block analyses in Appendix 2 of the Online Supplement. Encouragingly, the first-block results were identical to the full analyses collapsed across all three blocks, which are the primary data reported in the Results sections. It appears, therefore, that counterbalancing block order had the desired effect of neutralizing possible block order effects.

Exclusion criteria. We applied two exclusion criteria before analysing these data. First, we did not include the order data from any recall tests in which participants recalled three or fewer words. We did this because order measures are not informative with so few recalled items and, moreover, they tend to yield outliers that contaminate order analyses. Of the 360 total blocks of recall data (each of the 120 participants studied and was tested on the three different lists), our exclusion criterion resulted in the removal of 19 blocks (5.27%) of the order data. Four blocks were excluded from the mixed condition, six from the pure-aloud condition, and nine from the pure-silent condition.

We then checked the data for extreme outliers that were at least three standard deviations from the means for each variable. These outliers constituted 1.04% of the recall data, 1.99% of the intrusion data, and 0.56% of the order data. We present the results of the outlier-free analyses in the body of this Results section and include footnotes to highlight any differences that emerged when we analysed the data with the outliers included.

Recall. Table 1 shows the average proportion of correctly recalled words in each of the four cells of the experiment (mixed aloud, mixed silent, pure aloud, and pure silent). Responses that were the plural version of a studied word were scored as correct. To assess the production effect in recall, we conducted a Study Modality (aloud vs. silent) × List Type (mixed vs. pure) repeated measures ANOVA, in which both factors were manipulated within-subject. Unsurprisingly, there was a main of effect of Study Modality, $F(1, 115) = 30.07$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.21$, with reading aloud leading to greater recall than reading silently. There was also a main effect of List Type, $F(1, 115) = 14.78$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.11$, indicating greater recall of words in the pure lists than in the mixed list.

Critically, there was a significant Study Modality × List Type interaction, $F(1, 115) = 48.00$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.29$, replicating previous research (Jones & Pyc, 2014; Jonker et al., 2014). There was a robust production effect for the mixed list, $F(1, 117) = 59.20$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.34$. For the pure lists, however, there was no production effect, $F(1, 116) = 0.01$, $MSE = 0.01$, $p = .91$, $\eta^2 = 0.00$.

To test the null hypothesis for the pure-list effect, we used the Bayesian approximation procedure recommended by Wagenmakers (2007). We estimated the posterior odds from the ANOVA sum of

Table 1
Proportion of Correct Recall (With SEs) and Number of Intrusions in Experiments 1 and 2

List type/condition	Aloud items	Silent items	Intrusions
Experiment 1			
Mixed list	.25 (.01)	.15 (.01)	1.47 (.16)
Pure list	.23 (.01)	.24 (.01)	1.03 (.11)/1.42 (.13)
Experiment 2			
Mixed list	.21 (.01)	.13 (.01)	1.26 (.12)
Pure list	.21 (.01)	.21 (.01)	1.08 (.10)/1.42 (.12)
Combined			
Mixed list	.23 (.01)	.14 (.01)	1.36 (.10)
Pure list	.22 (.01)	.22 (.01)	1.06 (.07)/1.42 (.09)

Note. The intrusions for the pure-list condition refer to the pure-aloud and pure-silent groups, respectively.

squares, using a calculator provided by Masson (2011). We then converted the posterior odds into *p*BIC, which quantifies the degree of support favoring the null relative to the alternative hypothesis on a scale from 0 to 1 (where 1 indicates full support for the null). This analysis yielded "positive" evidence in favor of the null hypothesis, *p*BIC = 0.91, according to Raftery's (1995) system for labelling the strength of the evidence. This Bayesian analysis supports the conclusion that in recall there is no pure-list production effect.

There was a large cost associated with studying words silently in a mixed list versus in a pure list, $F(1, 119) = 64.69$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.35$. There was also a significant—though much smaller—benefit of studying words aloud in a mixed list versus in a pure list, $F(1, 115) = 4.45$, $MSE = 0.01$, $p = .04$, $\eta^2 = 0.04$.[4]

Intrusions. The overall number of intrusions was calculated for each list type of the experiment: mixed, pure aloud, and pure silent (see Table 1). Because participants did not indicate the study mode of each recalled word, only one overall intrusion rate was obtained in the mixed condition. Participants made significantly fewer intrusions when recalling words from the pure-aloud list than from the pure-silent list, $t(112) = 2.93$, $p = .004$, $d = 0.49$. Participants also made fewer intrusions when recalling words from the pure-aloud list versus the mixed list, $t(114) = 2.29$, $p = .02$, $d = 0.27$. The number of intrusions was similar in the mixed list and the pure-silent list, $t(115) = 0.14$, $p = .89$, $d = 0.02$. Thus, it appears that pure-aloud items may have held one type of memorial advantage: fewer intrusions (relative to mixed and pure-silent items). We address this result in the General Discussion.

Order data. We applied three order measures to our recall data for each list type. Each measure represented a different approach to assessing the consistency between study presentation order and recall order. First, we examined I-O correspondence (Asch & Ebenholtz, 1962)—the proportion of consecutively recalled words that were presented in the same direction during study. I-O correspondence is a measure commonly used in item-order memory research (e.g., Burns et al., 1993; Engelkamp, Jahn, & Seiler, 2003; Jonker et al., 2014; Mulligan & Lozito, 2007; Nairne et al., 1991). Second, we used a bidirectional pair frequency measure (Anderson & Watts, 1969; see also Sternberg & Tulving, 1977). This is a measure of the proportion of successively recalled items that were studied in adjacent study positions (i.e., in either a forward or a backward direction); Burns (1996) had used this measure to demonstrate the retention of order information of words in long lists. Third, we calculated the *distance* between the study list positions of words that were recalled consecutively. This measure takes into account the fact that order information may guide recall even when words are not recalled adjacently to words that were presented adjacently during study. For example, assume that "table" and "computer" were recalled consecutively. If "table" was the 18th word on the study list and "computer" was the 15th, then their distance (as an absolute value) would be 3. Thus, for each consecutively recalled pair of words, we calculated the distance between their study list positions, and we then obtained the mean of these distances (see Jonker et al., 2014). A low mean suggested that a participant was relatively reliant on order information, perhaps tending to recall words in clusters.

Table 2 shows the means for our three order measures: I-O correspondence, pair frequency, and distance. We had two hypotheses for these order data, both of which were derived from the mixed-list versus pure-list item-order account predictions outlined in McDaniel and Bugg (2008, Table 2). Our first hypothesis was that participants would retain more order information from pure-silent lists (the "common" encoding type) than from either pure-aloud lists (the relatively "uncommon" encoding type) or mixed lists. Our second hypothesis was that participants would retain more order information from mixed lists versus pure-aloud lists, although Jonker et al. (2014) did not obtain this result.

Table 2
Means (With SEs) for Three Different Order Measures in Experiments 1 and 2

List type/condition	I-O Correspondence	Pair Frequency	Distance
Experiment 1			
Mixed	.52 (.01)	.10 (.01)	13.06 (.40)
Pure aloud	.51 (.01)	.15 (.01)	12.66 (.43)
Pure silent	.55 (.02)	.21 (.02)	10.38 (.35)
Experiment 2			
Mixed	.52 (.01)	.11 (.01)	11.70 (.36)
Pure aloud	.54 (.01)	.14 (.01)	12.02 (.38)
Pure silent	.61 (.01)	.20 (.01)	10.52 (.36)
Combined			
Mixed	.52 (.01)	.11 (.01)	12.35 (.27)
Pure aloud	.52 (.01)	.14 (.01)	12.31 (.28)
Pure silent	.58 (.01)	.20 (.01)	10.46 (.25)

Consistent with Nairne et al. (1991) item-order account, participants had superior retention of order information after studying pure-silent lists than after studying pure-aloud lists or mixed lists. However, contrary to McDaniel and Bugg's (2008) claim that individuals retain more order information from mixed lists than from pure-uncommon lists, we did not observe superior order memory for mixed lists versus pure-aloud lists.

The increased reliance on order information in the pure-silent condition was least evident with the I-O correspondence measure. A repeated measures ANOVA, with list type (mixed vs. pure aloud vs. pure silent) as the within-subject factor, was nonsignificant, $F(2, 204) = 1.87$, $p = .16$, $\eta^2 = 0.02$. Paired sample *t* tests revealed I-O correspondence was greater for the pure-silent list than for the pure-aloud list, $t(104) = 2.18$, $p = .03$, $d = 0.26$. Although also numerically greater for the pure-silent list than for the mixed list, this difference in correspondence was not reliable, $t(108) = 1.44$, $p = .15$, $d = 0.19$. There was also a nonsignificant difference in I-O correspondence between the pure-aloud list and the mixed list, $t(108) = 0.57$, $p = .57$, $d = 0.08$.

The pair frequency measure revealed a tendency for participants to rely on order information particularly when recalling the pure-silent list. A one-way repeated measures ANOVA revealed a reliable effect of list type, $F(2, 202) = 13.20$, $MSE = 0.02$, $p < .001$, $\eta^2 = 0.12$. The proportion of consecutively recalled words that were study pairs was significantly higher for the pure-silent list than for either the pure-aloud list, $t(104) = 2.60$, $p = .01$, $d = 0.37$, or the mixed list, $t(106) = 5.55$, $p < .001$, $d = 0.77$. Contrary to our prediction, the proportion of consecutively recalled words that were studied consecutively was higher for the pure-aloud list than for the mixed list, $t(107) = 3.08$, $p = .003$, $d = 0.42$.

[4] When we included outliers, the benefit was not reliable, $F(1, 119) = 2.34$, $p = 0.13$, $\eta^2 = 0.02$.

The distance measure also provided evidence that participants relied more on order information to guide their recall of the pure-silent list compared to the pure-aloud list or the mixed list. The one-way ANOVA was again significant, $F(2, 206) = 13.47$, $MSE = 13.89$, $p < .001$, $\eta^2 = 0.12$. The study list distance of consecutively recalled items was significantly smaller for the pure-silent list than for either the pure-aloud list, $t(104) = 4.75$, $p < .001$, $d = 0.57$, or the mixed list, $t(107) = 5.02$, $p < .001$, $d = 0.65$. The study list distance of consecutively recalled words was not reliably different for the pure-aloud list versus the mixed list, $t(109) = 0.44$, $p = .66$, $d = 0.05$.

We also tested whether retention of order was greater than chance for each of the three study list conditions. The I-O correspondence measure was the most straightforward in this regard because the chance proportion of words recalled in a forward direction is clearly 0.50. A one-sample t test revealed that the I-O correspondence rate in the pure-silent condition was significantly higher than chance, $t(110) = 3.01$, $p = .003$, $d = 0.57$. I-O correspondence was not significantly higher than chance in the mixed condition, $t(115) = 1.21$, $p = .23$, $d = 0.23$, or in the pure-aloud condition, $t(111) = 0.50$, $p = .62$, $d = 0.10$.

For the distance measure, chance was calculated using the formula E = (w + 1)/3, where E is the expected distance score and w is the number of words on the study list. According to this formula, the expected distance score for a 40-item study list is 41/3 = 13.67. A one-sample t test revealed that participants' distance measure scores in the pure-silent condition were significantly lower than chance, $t(109) = -9.52$, $p < .001$, $d = 1.82$, signifying that participants had a strong tendency to cluster words in this condition. Participants' distance scores were not significantly lower than chance in the mixed condition, $t(115) = -1.55$, $p = .12$, $d = 0.29$, but they were significantly lower than chance in the pure-aloud condition, $t(113) = -2.36$, $p = .02$, $d = 0.44$, consistent with Burns' (1996) finding that a small amount of order information is retained after reading long pure-aloud lists.

In sum, using three different order measures, we found evidence of superior order retention for words studied in a pure-silent list compared to words studied in a pure-aloud list or a mixed list.

Relations between order measures and recall. Although this evidence of an order memory advantage in the pure-silent condition is consistent with the sizable mixed-list cost of production, this claim would be bolstered by demonstrating that order retention is positively correlated with recall, in particular for the pure-silent group (a result that was not found in Mulligan & Lozito, 2007, using the I-O correspondence measure). Assuming that participants do not harness order information to the same extent when recalling words following mixed lists and pure-aloud lists, the correlations between the three order measures and recall should be nonsignificant (or weaker) in the these two conditions.

To test whether the order memory advantage in the pure-silent condition improved recall, we correlated participants' free recall with their order memory scores for each of the three order measures. These correlations are shown in Table 3. Importantly, there was a significant correlation in the pure-silent condition between the proportion of words correctly recalled and the distance order measure, $r(108) = -0.25$, $p = .008$. Participants who had smaller distances between the original study positions of words that they recalled consecutively tended to have better free recall. The pair frequency measure was also significantly correlated with recall, $r(108) = 0.20$, $p = .04$.[5] The I-O correspondence measure was correlated in the expected direction with the recall, $r(109) = 0.10$, $p = .28$, but this correlation was not significant. In contrast, in the pure-aloud and mixed conditions, none of the three order measures were significantly correlated with recall (all ps > 0.05).[6]

Table 3
Correlation Coefficients Between the Three Order Measures and Recall for Each Condition of Experiment 1

Condition	Recall	I-O	Pair	Distance
Silent				
Recall	—			
I-O	.10	—		
Pair	.20*	.06	—	
Distance	−.25**	−.13	−.49***	—
Aloud				
Recall	—			
I-O	.05	—		
Pair	.14	.07	—	
Distance	−.08	−.22	−.54***	—
Mixed				
Recall	—			
I-O	.04	—		
Pair	.15	−.01	—	
Distance	−.08	−.03	−.34***	—

* $p < .05$. ** $p < .01$. *** $p < .001$.

Discussion

In this experiment, we assessed the production effect in free recall by comparing mixed versus pure long lists, using a large sample and a within-subject blocked design to ensure high experimental power. We found a robust mixed-list production effect in recall, but no pure-list production effect. There was also a large cost: Reading silently in a mixed list impaired recall relative to reading silently in a pure list.

This pattern of results was identical to that obtained by Jones and Pyc (2014; see also Lambert et al., 2016). Order analyses suggested that differential retention of order may underlie this pattern of results, consistent with the item-order account. Across three order measures, we found evidence that participants relied on order information more when recalling words that they had studied in a pure-silent list than when recalling words that they had studied in a pure-aloud list or a mixed list (though this evidence was weaker for the I-O correspondence measure). This superior order retention for pure-silent lists could account for the null pure-list production effect as well as for the cost of reading mixed lists silently. Correlations between the order measures and recall also yielded some evidence that participants may

[5] When outliers were included, this correlation was not statistically significant, $r(109) = 0.15$, $p = 0.11$.

[6] Unexpectedly, when the outliers were included, significant correlations emerged between pure-aloud condition recall and the I-O correspondence, $r(112) = 0.19$, $p = 0.05$, pair frequency, $r(112) = 0.23$, $p = 0.01$, and distance measures, $r(112) = -0.19$, $p = 0.04$. These measures were not significantly correlated in Experiment 2 or in the combined Experiment 1 and 2 analyses (either with or without the outliers removed). Therefore, there was inconsistent evidence of a relation between order retention and recall in the pure-aloud condition.

use order information to facilitate their free recall, especially for pure-silent lists.

Consistent with Jonker et al. (2014), there was also a benefit in free recall of reading aloud in a mixed list compared to a pure-aloud list. However, the item-order account would not appear to explain this benefit, because participants did not have better order memory for mixed lists versus pure-aloud lists. This benefit may have arisen because relational distinctiveness enhanced the recall of mixed aloud words, as we have argued is the case in a recognition task (see Forrin et al., 2016).

Experiment 2

Overall, Experiment 1 was consistent with previous research showing that order information is retained from studying long lists (Burns, 1996; Mulligan & Lozito, 2007). Further, our results suggest that order information can guide recall following the study of long lists—particularly when those lists are read silently. Without some form of distraction between study and recall, however, it is possible that STM factors could have contributed to these findings. Given that order information clearly is retained in STM (Bjork & Healy, 1974; Jonker et al., 2014), our goal in Experiment 2 was to eliminate the contribution of STM to participants' free recall. We did so in the traditional way (see Glanzer & Cunitz, 1966; Postman & Phillips, 1965), by adding a distractor task between the study and free recall phases of the experiment (consistent with the designs of Jones & Pyc, 2014, and Lambert et al., 2016). We used a distractor task commonly used in the item-order literature, which involved participants identifying digits as odd or even (e.g., Nairne et al., 1991). The addition of this distractor task allowed us to more conclusively test whether order information is better retained in long-term memory—without a possible STM contribution—when words are studied in a pure-silent list compared to a pure-aloud list or a mixed list.

Method

Participants. One-hundred and 50 undergraduate participants from the University of Waterloo took part in exchange for course credit. There were 25 participants in each of the six block orders. Two participants were excluded because they did not comply with the task instructions, leaving 148 participants.

Stimuli and apparatus. The same word list and apparatus were used as in Experiment 1.

Procedure. Participants again were given three study-test blocks (one for each list type: pure aloud, pure silent, and mixed), and block order was counterbalanced. The procedure was identical to that of Experiment 1, with the addition of a distractor task inserted between each study and test phase. In the distractor task, participants were presented with a series of digits (randomly chosen from the range of 1–9), one at a time in the centre of the screen. Participants were instructed to respond whether each digit was even or odd by pressing the "e" or "o" key; no feedback was provided. Each number appeared on the screen for 1 s with a 100-ms interstimulus interval. The distractor task took 30 s to complete. This distractor task was identical to that used by Nairne et al. (1991) in their original investigation of the item-order account, and has been commonly used in other item-order experiments, even those that did not find evidence of order information guiding recall following long study lists (Mulligan & Lozito, 2007).

Results

Exclusion criteria. Of the 444 blocks of order data (each of the 148 participants studied and were tested on the three different lists), our exclusion criterion resulted in the removal of 47 blocks (10.59%) of order data (21 blocks from the mixed condition, 10 from the pure aloud condition, and 16 from the pure silent condition). Recall was somewhat poorer in Experiment 2 (likely due to the distractor task), more blocks of order data were removed relative to Experiment 1. Removal outliers that were at least three standard deviations from the mean of each variable resulted in the exclusion of 0.51% of the recall data, 2.50% of the intrusion data, and 0.93% of the order data.

Recall. The mean proportions of correctly recalled words for each condition are shown in Table 1, along with the intrusion rates. As with Experiment 1, we conducted separate analyses for the full data and for the first-block data (see Appendix 3 of the Online Supplement for the block order analyses, and Appendix 4 of the Online Supplement for analyses of the first-block data). The results were largely consistent across these two analytic approaches although, not surprisingly, the full data again showed stronger support for the item-order account than did the first-block data. These differences are addressed in the General Discussion.

The ANOVA revealed a main effect of study modality, $F(1, 144) = 23.90$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.14$, with reading aloud leading to superior recall relative to reading silently. There was also a main effect of list type, $F(1, 144) = 46.06$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.25$, indicating greater recall of words in the pure lists than in the mixed list. Most important, there was a significant Study Modality × List Type interaction, $F(1, 144) = 33.51$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.19$. There was a robust production effect for the mixed list, with participants recalling a significantly higher proportion of words read aloud than read silently, $F(1, 146) = 46.01$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.24$. For the pure lists, however, there was again no trace of a production effect, $F(1, 145) = 0.31$, $MSE = 0.01$, $p = .58$, $\eta^2 = 0.002$. As in Experiment 1, we obtained "positive" support for the null (pBIC = .91).

There was a large cost associated with studying words silently in a mixed list versus in a pure list, $F(1, 146) = 79.77$, $MSE = 0.01$, $p < .001$, $\eta^2 = 0.36$. Conversely, there was not a significant benefit of studying words aloud in a mixed list versus in a pure list, $F(1, 146) = 0.02$, $MSE = 0.01$, $p = .88$, $\eta^2 = 0.00$. The Bayesian approximation procedure yielded "positive" evidence in favor of the null hypothesis, pBIC = 0.92, supporting the conclusion that there was no production benefit to reading words aloud in a mixed list versus a pure list.

Intrusions. Participants committed significantly fewer intrusions when recalling words from the pure-aloud list than from the pure-silent list, $t(141) = 2.25$, $p = .03$, $d = 0.23$. This small effect, also present in Experiment 1, suggests that a pure-list production effect may exist in terms of lower intrusions rates. Unlike Experiment 1, however, participants did not make fewer intrusions when recalling words from the pure-aloud list versus the mixed list, $t(140) = 1.22$, $p = .23$, $d = 0.12$. There also was no reliable difference between the number of intrusions in the mixed list and the pure-silent list, $t(143) = 0.95$, $p = .34$, $d = 0.10$.

Order analyses. Table 2 shows the means for our three order measures. First, we examined whether participants had higher I-O correspondence following the pure-silent list compared to the mixed list and the pure-aloud list. A repeated measures ANOVA, with list type as the within-subject factor, was significant, $F(2, 206) = 9.30$, $p < .001$, $MSE = 0.02$, $\eta^2 = 0.08$. Correspondence was greater for the pure-silent list compared to the pure-aloud list, $t(122) = 4.11$, $p < .001$, $d = 0.54$, and compared with the mixed list, $t(109) = 3.86$, $p < .001$, $d = 0.53$. There was not a reliable difference in correspondence between the pure-aloud list and the mixed list, $t(116) = 0.90$, $p = .37$, $d = 0.12$.

Next, for the pair frequency measure, a repeated measures ANOVA again revealed a reliable effect of list type, $F(2, 208) = 10.93$, $MSE = 0.02$, $p < .001$, $\eta^2 = 0.10$. The proportion of consecutively recalled words that were study pairs was significantly higher for the pure-silent list than for either the pure-aloud list, $t(122) = 3.82$, $p = .001$, $d = 0.46$, or the mixed list, $t(110) = 4.05$, $p < .001$, $d = 0.53$. The proportion of consecutively recalled words that were also studied consecutively was higher for the pure-aloud list than for the mixed list, $t(116) = 2.58$, $p = .01$, $d = 0.31$, a result that runs counter to the item-order account.

The results of the distance measure were also consistent with the item-order account, although not as strongly. The repeated measures ANOVA was marginally significant, $F(2, 212) = 2.79$, $MSE = 16.02$, $p = .06$, $\eta^2 = 0.03$. The distance measure for the pure-silent list was significantly smaller than for the pure-aloud list, $t(122) = 2.73$, $p = .01$, $d = 0.35$, and for the mixed list, $t(109) = 1.84$, $p = .07$, $d = 0.24$. The distance measure was not reliably different for the pure-aloud list and the mixed list, $t(118) = 0.45$, $p = .66$, $d = 0.06$.[7]

We then examined whether order memory was better than chance in each condition. In the pure-silent condition, participants' average I-O correspondence rate was significantly higher than chance, $t(129) = 8.76$, $p < .001$, $d = 1.54$. I-O correspondence was also significantly higher than chance in the pure-aloud condition, $t(136) = 2.68$, $p = .01$, $d = 0.46$, but not in the mixed condition, $t(124) = 1.62$, $p = .11$, $d = 0.29$. Participants' distance scores were significantly smaller than chance in the pure-silent condition, $t(127) = 8.78$, $p < .001$, $d = 1.56$, in the pure-aloud condition, $t(137) = 4.37$, $p < .001$, $d = 0.75$, and in the mixed condition, $t(126) = 5.50$, $p < .01$, $d = 0.98$. These smaller-than-chance distance scores indicate that participants' free recall tended to be clustered (in terms of study list order) in all three conditions.

In sum, extending the results of Experiment 1, we found that participants retained order information from long study lists when a distractor phase followed study to eliminate the potential impact of STM. Participants still showed superior retention of order information in the pure-silent condition, although this advantage in terms of clustering (as assessed by the distance measure) decreased relative to Experiment 1. This decreased advantage in the pure-silent condition appears to have occurred because participants now showed a tendency to cluster recalled items in the mixed and pure-aloud conditions as well, perhaps because the distractor phase mitigated the influence of STM factors such as recency effects. We examined these differences more closely in a combined analysis (see Appendix 5 of the Online Supplement).

Relations between order measures and recall. Next, we correlated each order memory measure with free recall to test the prediction that superior order retention in the pure-silent condition facilitated free recall. These correlations are shown in Table 4. In the pure-silent condition, the distance measure correlated negatively with recall, as anticipated, although this correlation was only marginal, $r(125) = -0.15$, $p = .09$.[8] Neither the pair frequency measure nor the I-O correspondence measure correlated significantly with recall. In the pure-aloud condition, none of the order measures correlated significantly with the proportion of words correctly recalled. Last, in the mixed condition, unexpectedly, the pair frequency measure was significantly positively correlated with the proportion of words correctly recalled, $r(123) = 0.21$, $p = .02$. The distance and I-O correspondence measures were not significantly correlated with the recall of mixed-list words.

Table 4
Correlation Coefficients Between the Three Order Measures and Recall for Each Condition of Experiment 2

Condition	Recall	I-O	Pair	Distance
Silent				
Recall	—			
I-O	−.11	—		
Pair	.11	.30***	—	
Distance	−.15†	−.13	−.46***	—
Aloud				
Recall	—			
I-O	−.05	—		
Pair	.06	.25**	—	
Distance	−.01	−.04	−.37***	—
Mixed				
Recall	—			
I-O	.11	—		
Pair	.21*	.05	—	
Distance	−.08	−.08	−.41***	—

† $p < .1$. * $p < .05$. ** $p < .01$. *** $p < .001$.

In summary, these analyses demonstrated that order retention in the pure-silent condition was related to recall, in particular for the distance measure. Surprisingly, the distance measure was a stronger predictor of pure-silent list recall in the first-block data (see Appendix 4 of the Online Supplement), despite the fact that a pure-silent list-order memory benefit was not evident for this measure—indeed, participants did not show a greater-than-chance tendency to cluster their recall.

Discussion

In Experiment 2, we followed the study phase with a distractor task intended to eliminate STM contributions to free recall. The pattern was again consistent with the item-order account: There was a robust cost of reading silently in a mixed list, and no pure-list effect. Unlike Experiment 1, however, the benefit to

[7] When outliers were included, the results of the distance measure were less consistent with the item-order account. The repeated measures ANOVA was non-significant, $F(2, 212) = 1.64$, $MSE = 17.42$, $p = 0.23$, $\eta^2 = 0.01$. However, participants still had smaller distance scores for their recalled words after studying the pure-silent list vs. the pure-aloud list, $t(125) = 2.16$, $p = 0.03$, $d = 0.28$, even though there was no difference between the pure-silent list and the mixed list, $t(112) = 0.76$, $p = 0.45$, $d = 0.10$.

[8] With outliers included, this correlation became significant, $r(129) = -0.24$, $p = 0.007$.

reading aloud in a mixed list was nonsignificant, consistent with previous work (Jones & Pyc, 2014; Jonker et al., 2014) showing that there is, at most, a small benefit in recall—one that is overshadowed by a sizable cost.

Once again, our results were only partly consistent with the item-order account predictions outlined by McDaniel and Bugg (2008). In line with their predictions, our order measures again showed fairly consistent evidence that order memory was used to guide recall to a greater extent following the pure-silent list than following the pure-aloud list or the mixed list. As with Experiment 1, two of the order measures—I-O correspondence and pair frequency—clearly showed superior order memory for the pure-silent list. The distance order measure also showed this pattern in the full data, though it was reduced in magnitude relative to Experiment 1 and was nonsignificant in the first-block data. (We address the differences between the full data and the first-block data in the General Discussion.) However, contrary to one of the item-order predictions outlined by McDaniel and Bugg (2008), there was again no evidence of superior order memory for mixed lists versus pure-aloud lists.

General Discussion

In two experiments, we found a significant mixed-list production effect in long-list free recall and, conversely, positive Bayesian evidence of a *null* pure-list production effect. The significant difference in these two production effects reflected a robust mixed-list cost to reading silently, accompanied by at most a small benefit of reading aloud. This is the same pattern of results obtained in previous research (Jones & Pyc, 2014; Jonker et al., 2014; Lambert et al., 2016).

Contrasting the Production Effect in Recall and Recognition

Notably, this pattern of results for recall diverges from that obtained for mixed-list versus pure-list designs that have examined the production effect for recognition (Bodner et al., 2014; Forrin et al., 2016). There are two main differences that we will address in turn. The first is that a pure-list production effect is nonexistent in recall, despite there being a significant pure-list production effect in recognition (Bodner et al., 2014; Fawcett, 2013; Forrin et al., 2016). As previously argued (Jones & Pyc, 2014; Jonker et al., 2014), the absence of the effect in recall seems to be due to the advantage of superior order encoding in the pure-silent condition offsetting the advantage of superior item encoding in the pure aloud condition, consistent with the item-order account (McDaniel & Bugg, 2008; Nairne et al., 1991). On a recognition test, however, because the experimenter determines the order of the test, only the advantage of item encoding comes into play, resulting in a pure-list production effect.

That said, we did find evidence of a certain type of advantage to pure-list production in recall: Fewer intrusions occurred for pure-aloud lists than for pure-silent lists (except in the first-block analysis of Experiment 2; see Appendix 4 of the Online Supplement). Just as the distinctiveness heuristic has been posited to underlie the lower pure-aloud false alarm rate on a recognition test (Dodson & Schacter, 2001), this heuristic may similarly reduce the pure-aloud intrusion rate on a free recall test. Previous research has demonstrated that individuals may use the distinctiveness heuristic to avoid recalling critical lures in a DRM paradigm (McCabe & Smith, 2006). The present results suggest that the distinctiveness heuristic may have the benefit of mitigating intrusions more broadly (cf. Lambert et al., 2016, who did not find any significant differences in intrusion rates).

The second way in which the pattern of mixed-list versus pure-list production effect results differs across recall and recognition is in terms of the benefit of reading aloud in a mixed list. The benefit effect found here was small (Experiment 1) and unreliable (Experiment 2), and has been comparably small (Jonker et al., 2014) and unreliable (Jones & Pyc, 2014; Lambert et al., 2016) in other experiments. Conversely, Forrin et al. (2016) found a sizable benefit of production in recognition. The mixed-list benefit of production in recognition has been presumed (MacLeod et al., 2010) to derive from distinctiveness—both relational distinctiveness at encoding and a speech distinctiveness heuristic at test. Perhaps, then, there is little to no mixed-list benefit of production in recall because distinctiveness does not enhance recall to the same extent that it bolsters recognition.

According to McDaniel and Bugg (2008), free recall of uncommon items is better in a mixed list versus a pure list—a benefit effect—because mixed lists hold an order memory advantage compared to pure lists of uncommon items (e.g., DeLosh & McDaniel, 1996; McDaniel et al., 1995; Serra & Nairne, 1993). McDaniel and Bugg (2008) contend that this order memory advantage arises from a spillover effect from the common items on the mixed list. However, the present results and those of Lambert et al. (2016) did not yield evidence of better order memory for mixed lists relative to pure-aloud lists. Moreover, Jonker et al. (2014), who found a reliable benefit effect in short-list recall, also did not find superior order memory for mixed lists versus pure-aloud lists. These results call into question McDaniel and Bugg's (2008) claim that a benefit arises in recall from an order memory advantage. Instead, we posit that relational distinctiveness may account for any benefit of production there appears to be in recall. Moreover, we propose that this benefit may tend to be larger for more distinctive processes (e.g., generated or bizarre items), a conjecture that the literature would seem to support (see Serra & Nairne, 1993, and McDaniel et al., 1995, respectively, for robust benefit effects).

Conversely, the mixed-list cost of reading silently in recall is consistent and robust (Experiments 1 and 2; Jones & Pyc, 2014; Jonker et al., 2014; Lambert et al., 2016). Thus, in contrast to the aforementioned differences in the pattern of production effects across recall and recognition, both tests appear to yield a significant cost (for a cost in recognition, see Bodner et al., 2014; Forrin et al., 2016). The mixed list cost in recognition has been found to be consistent with a lazy reading account (see Bodner et al., 2014). Conversely, Lambert et al. (2016) present evidence that lazy reading does not underlie the cost in recall. Our results support the hypothesis that this cost is at least partially driven by superior order memory for pure-silent lists (cf. Lambert et al., 2016).

Reconciling the Differences Between Our Study and Lambert et al. (2016)

Both we and Lambert et al. (2016) found evidence of a significant cost in recall to reading silently in mixed list. The cost that

we observed was consistent with the item-order account (i.e., participants retained more order information from pure-silent lists than from mixed lists), whereas the cost observed by Lambert et al. (2016) does not appear to have resulted from superior order encoding in pure-silent lists. What might explain these divergent results?

Arguably, the most parsimonious explanation would rest on the fact that Lambert et al. (2016) instructed participants not to use any memory strategies during study, and removed any participants who did so. Their tendency to find that order information was not retained from pure-silent lists—and did not guide free recall could therefore have arisen simply because their participants were prohibited from encoding this information. Despite these instructions, Lambert et al. (2016) still found evidence in their Experiment 4 of an order memory advantage for pure-silent versus mixed lists (in terms of the pair frequency measure in particular), consistent with the robust cost in that experiment. We submit that Lambert et al.'s (2016) Experiment 4 results constitute strong evidence that participants retain order information from long pure-silent lists, insofar as participants persisted in using this strategy to some degree despite instructions that prohibited its use.

A second factor that may have contributed to our experiments yielding much stronger evidence of pure-silent list order retention than their experiments is the fact that we used a within-subject blocked design, whereas their participants only studied and were tested on one list type. For our second and third blocks, participants would likely have anticipated the recall test. Participants' knowledge of the recall test may have encouraged them to use more effective encoding and particularly retrieval strategies to facilitate their recall—including using order information—which is reflected in the pattern of improved recall across blocks (see Appendixes 1 and 3 of the Online Supplement). This assumption is also supported by the result that participants tended to show a more consistent order memory advantage for pure-silent lists in the full data set than in the first-block data, particularly in our Experiment 2 (see Appendix 4 of the Online Supplement).[9]

A third factor that may account for our stronger evidence of participants retaining order information from long pure-silent lists is the nature of the distractor phase. In our experiments recall immediately followed study (Experiment 1) or an easy distractor phase (Experiment 2), whereas Lambert et al.'s (2016) free-recall test followed a longer and more difficult distractor phase. Their distractor phase may have reduced participants' reliance on order information (see Burns et al., 1993, for evidence that difficult distractor tasks disrupt relational information more than easy distractor tasks do).

Although the above factors may account for why we found stronger evidence of pure-silent list order retention than did Lambert et al. (2016), they nonetheless also found ample evidence of a cost. Thus, it appears that the differential retention of order information (as per the item-order account) is not the only mechanism driving the cost to recall of reading silently in a mixed list. What other mechanisms may have come into play? A prime candidate would seem to be output order interference: Recall of silent items studied in mixed lists may be impaired by participants' proclivity to recall aloud items first.[10] Future research could test this possibility by instructing participants to recall the silent items first after studying mixed lists. If output interference is a factor, this should mitigate the cost in recall.

We also contend, however, that Lambert et al.'s (2016) results were not entirely inconsistent the notion that the cost in recall can be explained by the item order account—in particular, when comparing their Experiments 1 and 4, which most closely matched the designs of our experiments. In their Experiment 1, they did not find any evidence of order retention for pure-silent lists, and their cost was relatively small (4%) and nonsignificant. In contrast, their Experiment 4 results showed significant evidence of pure-silent list order retention, and, correspondingly, a larger (8%) and significant cost.

Does Order Information Guide Free Recall of Long lists?

Importantly, our evidence that participants *retain* order information from long lists—particularly those of common (pure silent) items—is entirely consistent with previous long-list item-order research. Indeed, this was demonstrated by Burns (1996) for pure-aloud lists (up to 40 words long), which involve more "unusual" encoding than pure-silent lists. Mulligan and Lozito (2007) also showed greater-than-chance order retention of 24 word pure-silent lists (cf. Lambert et al., 2016). But there is a crucial difference worth noting: Mulligan and Lozito (2007) did not find the significant relation between long-list order memory and recall that we found here. What could account for this difference?

First, it may indeed be the case, as Mulligan and Lozito (2007) clearly showed, that order information guides free recall to a lesser extent for long lists than for short lists. For long lists, the influence of order information may be fairly weak and may require higher experimental power to uncover. In line with this possibility, the correlations that we found between pure-silent order memory and recall were the strongest and were more consistent for our combined analyses (see Table E5 of the Online Supplement).

Second, perhaps study direction information (the order information capture by the I-O correspondence measure) simply does not guide individuals' free recall of long lists. Indeed, replicating Mulligan and Lozito's (2007) results (see also Engelkamp et al., 2003), we did not find evidence of a significant relation between the I-O correspondence measure and free recall in either of our experiments or in our combined results. We did, however, find evidence of significant relations between recall and the other two order measures (pair frequency and distance) and recall. These correlations were significant in the combined data—both full data (see Table 5) and first-block (see Table E5 of the Online Supplement). It appears that participants may be able to retrieve information about words that were studied in close proximity (i.e., in pairs or clusters) to guide their free recall of long lists.

[9] Note that our blocked design was not unusual for item-order research: Short-list item-order research typically uses several study-test blocks for each list type (e.g., Jonker et al., 2014). Our results also are consistent with Engelkamp, Jahn, and Seiler (2003) research showing that recall increases across blocks in a short-list free recall paradigm, as does participants' reliance on order information.

[10] We thank Reviewer 2 for raising this point.

Table 5
Correlation Coefficients Between the Three Order Measures and Recall for Each Condition of the Combined Data From Experiments 1 and 2

Condition	Recall	I-O	Pair	Distance
Silent				
Recall	—			
I-O	−.01	—		
Pair	.15*	.17**	—	
Distance	−.20**	−.12	−.48***	—
Aloud				
Recall	—			
I-O	−.03	—		
Pair	.10	.17**	—	
Distance	−.03	−.12†	−.44***	—
Mixed				
Recall	—			
I-O	.07	—		
Pair	.18**	.02	—	
Distance	−.01	−.06	−.37***	—

† $p < .1$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Summary and Conclusions

The present research constitutes novel evidence that the item-order account may also encompass a long-list free recall paradigm. Whether participants were given an immediate recall test (Experiment 1) or a delayed test following a distractor task (Experiment 2) they tended to show superior order retention after studying a pure-silent list than after studying a pure-aloud list or a mixed list. Moreover, participants' order memory was significantly correlated with their recall in the pure-silent condition for two of the three order measures.

Our results replicated Jones and Pyc's (2014) finding that there is no pure-list production effect in free recall in terms of the proportion of words correctly recalled. We also replicated Jones and Pyc's (2014) finding that there is a robust cost to reading silently in a mixed list without a reliable corresponding benefit to reading aloud (see also Lambert et al., 2016). In so doing, we demonstrated that this pattern of results was largely consistent with Nairne et al. (1991) item-order account. Interestingly, regardless of whether participants study words in short lists (Jonker et al., 2014) or in a long list (the present experiment), they rely more on order information to guide recall of words in a pure-silent list than in a pure-aloud list or a mixed list. It would seem that remembering order can enhance recall regardless of the length of the list, at least when nothing is done to disrupt the encoding of order. The work that we have reported here has, then, broadened the domain of the item-order account of design effects in free recall.

Résumé

Les différences au niveau de la mémoire quant à lordre des items ont été utilisées pour expliquer l'absence d'effets entre sujets (c'est-à-dire présentés en liste pure) en rappel libre pour plusieurs techniques de codage, y compris l'effet de production, la conclusion suggérant que la lecture à voix haute favoriserait la mémorisation comparativement à la lecture silencieuse. Il est à noter, toutefois, que les éléments qui étayent ce postulat de l'ordre des items (Nairne, Riegler, & Serra, 1991) sont principalement dérivés de paradigmes de listes courtes. Nous présentons de nouvelles preuves suggérant que le postulat de l'ordre des items s'applique également lors du rappel de longues listes. Dans l'expérience 1, les participants ont étudié puis effectué un rappel libre de trois différentes longues listes de mots : exclusivement à voix haute, exclusivement silencieusement, puis un mélange des deux (la moitié lus à haute voix, la moitié lus silencieusement). Une analyse bayésienne n'a pas démontré d'effet de production en lien avec les items présentés en liste pure, et les analyses subséquentes de l'ordre des items étaient en grande partie conformes avec le postulat de l'ordre des items. Ces résultats indiquent que l'information relative à l'ordre est conservée dans la mémoire à long terme et est utile dans le guidage de rappel libre subséquent. Dans l'expérience 2, une tâche de distraction est insérée entre les phases d'étude et d'essai, faisant en sorte que seuls les processus de mémoire à long terme sont impliqués dans le rappel : les résultats sont restés conformes avec le postulat de l'ordre des items. Les informations de commande peuvent être conservées dans la mémoire à long terme pour les longues listes et sont utiles dans le guidage de rappels libres subséquents, élargissant ainsi l'étendue du postulat de l'ordre des items.

Mots-clés : effet de production, mémoire, rappel, postulat de l'ordre des items.

References

Anderson, R. C., & Watts, G. H. (1969). Bidirectional associations in multi-trial free recall. *Psychonomic Science, 15,* 288–289. http://dx.doi.org/10.3758/BF03336303

Asch, S. E., & Ebenholtz, S. M. (1962). The process of free recall: Evidence for non-associative factors in acquisition and retention. *The Journal of Psychology, 54,* 3–31. http://dx.doi.org/10.1080/00223980.1962.9713093

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

Begg, I., Snider, A., Foley, F., & Goddard, R. (1989). The generation effect is no artifact: Generating makes words distinctive. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 977–989. http://dx.doi.org/10.1037/0278-7393.15.5.977

Bjork, E. L., & Healy, A. F. (1974). Short-term order and item retention. *Journal of Verbal Learning and Verbal Behavior, 13,* 80–97. http://dx.doi.org/10.1016/S0022-5371(74)80033-2

Bodner, G. E., & Taikh, A. (2012). Reassessing the basis of the production effect in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 1711–1719. http://dx.doi.org/10.1037/a0028466

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

Burns, D. J. (1996). The item-order distinction and the generation effect: The importance of order information in long-term memory. *The American Journal of Psychology, 109,* 567–580. http://dx.doi.org/10.2307/1423395

Burns, D. J., Curti, E. T., & Lavin, J. C. (1993). The effects of generation on item and order retention in immediate and delayed recall. *Memory & Cognition, 21,* 846–852. http://dx.doi.org/10.3758/BF03202752

Cho, K. W., & Feldman, L. B. (2013). Production and accent affect memory. *The Mental Lexicon, 8,* 295–319. http://dx.doi.org/10.1075/ml.8.3.02cho

Conway, M. A., & Gathercole, S. E. (1987). Modality and long-term memory. *Journal of Memory and Language, 26,* 341–361. http://dx.doi.org/10.1016/0749-596X(87)90118-5

DeLosh, E. L., & McDaniel, M. A. (1996). The role of order information in free recall: Application to the word-frequency effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1136–1146. http://dx.doi.org/10.1037/0278-7393.22.5.1136

Dodson, C. S., & Schacter, D. L. (2001). "If I had said it I would have remembered it:" Reducing false memories with a distinctiveness heuristic. *Psychonomic Bulletin & Review, 8,* 155–161. http://dx.doi.org/10.3758/BF03196152

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

Engelkamp, J., Jahn, P., & Seiler, K. H. (2003). The item-order hypothesis reconsidered: The role of order information in free recall. *Psychological Research, 67,* 280–290. http://dx.doi.org/10.1007/s00426-002-0118-1

Engelkamp, J., & Krumnacker, H. (1980). Imaginale und motorische Prozesse beim Behalten verbalen Materials [Imagery and motor processes in memory of verbal material]. *Zeitschrift für Experimentelle und Angewandte Psychologie, 27,* 511–533.

Engelkamp, J., & Zimmer, H. D. (1997). Sensory factors in memory for subject-performed tasks. *Acta Psychologica, 96,* 43–60. http://dx.doi.org/10.1016/S0001-6918(97)00005-X

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers, 28,* 1–11. http://dx.doi.org/10.3758/BF03203630

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

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

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

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior, 5,* 351–360. http://dx.doi.org/10.1016/S0022-5371(66)80044-0

Hirshman, E., & Bjork, R. A. (1988). The generation effect: Support for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 484–494. http://dx.doi.org/10.1037/0278-7393.14.3.484

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

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 1–25). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195169669.003.0001

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

Hunt, R. R., & Worthen, J. B. (2006). *Distinctiveness and memory*. New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195169669.001.0001

Jahnke, J. C. (1965). Primacy and recency effects in serial-position curves of immediate recall. *Journal of Experimental Psychology, 70,* 130–132. http://dx.doi.org/10.1037/h0022013

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

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

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition, 24,* 103–109. http://dx.doi.org/10.3758/BF03197276

Lambert, A. M., Bodner, G. E., & Taikh, A. (2016). *Evaluating the basis of the production effect in recall*. Manuscript submitted for publication.

Lin, O. Y. H., & MacLeod, C. M. (2012). Aging and the production effect: A test of the distinctiveness account. *Canadian Journal of Experimental Psychology, 66,* 212–216. http://dx.doi.org/10.1037/a0028309

MacDonald, P. A., & MacLeod, C. M. (1998). The influence of attention at encoding on direct and indirect remembering. *Acta Psychologica, 98,* 291–310. http://dx.doi.org/10.1016/S0001-6918(97)00047-4

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

Mandler, G., & Dean, P. J. (1969). Seriation: Development of serial order in free recall. *Journal of Experimental Psychology, 81,* 207–215. http://dx.doi.org/10.1037/h0027767

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43,* 679–690. http://dx.doi.org/10.3758/s13428-010-0049-5

McCabe, D. P., & Smith, A. D. (2006). The distinctiveness heuristic in false recognition and false recall. *Memory, 14,* 570–583. http://dx.doi.org/10.1080/09658210600624564

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

McDaniel, M. A., Cahill, M., Bugg, J. M., & Meadow, N. G. (2011). Dissociative effects of orthographic distinctiveness in pure and mixed lists: An item-order account. *Memory & Cognition, 39,* 1162–1173. http://dx.doi.org/10.3758/s13421-011-0097-9

McDaniel, M. A., DeLosh, E. L., & Merritt, P. S. (2000). Order information and retrieval distinctiveness: Recall of common versus bizarre material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1045–1056. http://dx.doi.org/10.1037/0278-7393.26.4.1045

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

McDaniel, M. A., Einstein, G. O., De Losh, E. L., May, C. P., & Brady, P. (1995). The bizarreness effect: It's not surprising, it's complex. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21,* 422–435.

McDaniel, M. A., Waddill, P. J., & Einstein, G. O. (1988). A contextual account of the generation effect: A three-factor theory. *Journal of Memory and Language, 27,* 521–536. http://dx.doi.org/10.1016/0749-596X(88)90023-X

Mulligan, N. W. (1999). The effects of perceptual interference at encoding on organization and order: Investigating the roles of item-specific and relational information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 54–69. http://dx.doi.org/10.1037/0278-7393.25.1.54

Mulligan, N. W. (2002). The generation effect: Dissociating enhanced item memory and disrupted order memory. *Memory & Cognition, 30,* 850–861. http://dx.doi.org/10.3758/BF03195771

Mulligan, N. W., & Lozito, J. P. (2007). Order information and free recall: Evaluating the item-order hypothesis. *Quarterly Journal of Experimen-*

tal Psychology: Human Experimental Psychology, 60, 732–751. http://dx.doi.org/10.1080/17470210600785141

Nairne, J. S., & Kelley, M. R. (2004). Separating item and order information through process dissociation. *Journal of Memory and Language, 50,* 113–133. http://dx.doi.org/10.1016/j.jml.2003.09.005

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

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

Ozubko, J. D., Major, J., & MacLeod, C. M. (2014). Remembered study mode: Support for the distinctiveness account of the production effect. *Memory, 22,* 509–524. http://dx.doi.org/10.1080/09658211.2013.800554

Postman, L. (1972). A pragmatic view of organization theory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 3–48). New York, NY: Academic Press.

Postman, L., & Phillips, L. W. (1965). Short-term temporal changes in free recall. *The Quarterly Journal of Experimental Psychology, 17,* 132–138. http://dx.doi.org/10.1080/17470216508416422

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology 1995* (pp. 111–196). Cambridge, MA: Blackwell.

Schacter, D. L., Israel, L., & Racine, C. (1999). Suppressing false recognition in younger and older adults: The distinctiveness heuristic. *Journal of Memory and Language, 40,* 1–24. http://dx.doi.org/10.1006/jmla.1998.2611

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

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 592–604. http://dx.doi.org/10.1037/0278-7393.4.6.592

Slamecka, N. J., & Katsaiti, L. T. (1987). The generation effect as an artifact of selective displaced rehearsal. *Journal of Memory and Language, 26,* 589–607. http://dx.doi.org/10.1016/0749-596X(87)90104-5

Sternberg, R. J., & Tulving, E. (1977). The measurement of subjective organization in free recall. *Psychological Bulletin, 84,* 539–556. http://dx.doi.org/10.1037/0033-2909.84.3.539

Taikh, A., & Bodner, G. E. (2016). *Evaluating the basis of the between-group production effect in recognition.* Manuscript submitted for publication.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York, NY: Columbia University, Teachers College.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14,* 779–804. http://dx.doi.org/10.3758/BF03194105

Wickelgren, W. A. (1969). Associative strength theory of recognition memory for pitch. *Journal of Mathematical Psychology, 6,* 13–61. http://dx.doi.org/10.1016/0022-2496(69)90028-5

Received July 7, 2015
Accepted March 4, 2016 ■