Memory & Cognition (2024) 52:1408–1421
https://doi.org/10.3758/s13421-024-01551-6

The drawing effect: Evidence for costs and benefits using pure and mixed lists

Mark J. Huff[1] · **Jacob M. Namias**[1] · **Peyton Poe**[1]

Accepted: 7 March 2024 / Published online: 22 March 2024

Abstract
Drawing a referent of a to-be-remembered word often results in better recognition and recall of this word relative to a control task in which the word is written, a pattern dubbed the *drawing effect*. Although this effect is not always found in pure lists, we report three experiments in which the drawing effect emerged in both pure- and mixed-lists on recognition and recall tests, though the effect was larger in mixed lists. Our experiments then compared drawing effects on memory between pure- and mixed-list contexts to determine whether the larger mixed-list drawing effect reflected a benefit to draw items, a cost to write items, or a combination. In delayed recognition and free-recall tests, a mixed-list benefit emerged for draw items in which memory for mixed-list draw items was greater than pure-list draw items. This mixed-list drawing benefit was accompanied by a mixed-list writing cost compared to pure-list write items, indicating that the mixed-list drawing effect does not operate cost-free. Our findings of a pure-list drawing effect are consistent with a memory strength account, however, the larger drawing effect in mixed lists suggest that participants may also deploy a distinctiveness heuristic to aid retrieval of drawn items.

Keywords Recognition · Recall · Drawing effect · Design effect · Distinctive encoding

Introduction

The method in which studied information is initially encoded is critical for successful retrieval. For decades, memory researchers have investigated effective methods to facilitate long-term retention. Encoding strategies are often ascribed to one of two categories – deep/elaborative tasks, which are effective at improving retention, or shallow tasks, which are less effective (Craik & Lockhart, 1972). One mechanism for the benefit of elaborative encoding involves the recruitment of additional semantic, motor, or perceptual features associated with to-be-remembered items that can serve as effective retrieval cues at test (McDaniel & Bugg, 2008). Indeed, many of these elaborative tasks produce robust memory improvements when compared to shallow tasks or more "neutral" tasks such as intentional encoding without specific task instructions (e.g., silent reading). Examples of elaborative tasks include production (Conway & Gathercole, 1987; Hopkins & Edwards, 1972; MacLeod & Bodner, 2017), enactment (Engelkamp & Krumnacker, 1980; Engelkamp & Dehn, 2000; Roberts et al., 2022), generation (Bertsch et al., 2007; Slamecka & Graf, 1978), survival processing (Nairne et al., 2007), bizarreness (McDaniel & Einstein, 1986), and drawing an image of a word's referent versus writing (Fernandes et al., 2018; Namias et al., 2022; Wammes et al., 2016, 2017, 2018).

While the growing list of encoding tools that can improve memory is important for both basic and applied practice, a common feature of these encoding tasks is that the relative improvement is contextually moderated. Specifically, presenting study items in a *mixed-list* context, in which participants complete an elaborative task within the context of a more impoverished task, often yields a larger memory improvement for elaboratively encoded items compared to when these tasks are placed in a *pure-list context*. For instance, MacLeod et al. (2010) reported that correct recognition was only reliably greater when production was completed in a mixed- than a pure-list context. A similar pattern was also reported by Bodner et al. (2014), who reported a larger difference in recognition discriminability for produced (vs. silent) items in mixed than pure lists (cf. Fawcett, 2013). In free recall, Jones and Pyc (2014) did not find a pure-list

✉ Mark J. Huff
mark.huff@usm.edu

[1] The University of Southern Mississippi, 118 College Drive #5025, Hattiesburg, MS 39406, USA

production effect, and Jonker et al. (2014) reported a diminished production effect for pure lists relative to mixed lists.

Outside of the production effect, differences in magnitude of the memory improvement between mixed- and pure-list conditions are common. This *design effect* (McDaniel et al., 1988) has been found using other deep encoding tasks including the generation effect (Begg & Snider, 1987; Bertsch et al., 2007; Slamecka & Graf, 1978) and studying pictures versus words (Schacter et al., 1999). In a recent meta-analysis, Huff et al. (2015) reported greater hit rates in mixed lists than pure lists across several elaborative encoding tasks including generation, production, picture encoding, and unique font tasks, indicating that design effects are not a biproduct of an individual task type and manifest more broadly. Design effects have also been used as evidence for different memory processes that contribute to benefits of elaborative encoding. Specifically, in a mixed-list design, elaborative encoding can result in the creation of distinctive memory traces that participants can then use to monitor for at test, allowing them to retrieve distinctive memory items more effectively over non-distinctive memory items (e.g., *a distinctiveness heuristic*; Schacter et al., 1999). Separately, in pure-list designs, elaborative tasks lack a non-distinctive comparison, thus any memory improvements would be due to enhancement of the *strength* of the memory trace. Instances in which both mixed-list and pure-list memory improvements occur are interesting because they indicate that elaborative encoding is enhancing both strength and distinctiveness, which can operate in tandem. For instance, an elaborative task may enhance the strength of encoded items, but when participants are presented with non-distinctive items in mixed lists, they can also deploy a distinctiveness heuristic at test.

Aside from test-based distinctiveness processes, design effects may also suggest that participants use different encoding processes in pure- and mixed-list contexts. MacLeod et al. (2010) suggested that mixed-list production encoding increases distinctive processing of aloud items because they are contrasted to silently read words. Further, McDaniel and Bugg (2008) suggested that this elaborative encoding contrast can produce a tradeoff between the encoding of distinctive, item-specific information and encoding of relational interitem associations. Specifically, elaborative tasks may operate to strengthen encoding of individual items, but this encoding comes at the expense of processing associations between items, including serial order information that could be used as a retrieval cue (see too, Hege & Dodson, 2004, for a similar impoverished relational encoding account). In mixed lists, relational encoding is more likely to become disrupted because participants actively switch between two encoding tasks. However, item-specific information will be enhanced, leading to a greater memory difference between elaborative and non-elaborative encoding tasks. In contrast, pure lists that are studied using non-elaborative encoding may benefit from relatively intact relational encoding (because encoding tasks do not change), which reduces the difference between the two encoding types.

Although elaborative encoding tasks are often simple to deploy, the utility of these tasks can be accompanied by both costs and benefits to memory due to tradeoffs in encoding processes. A *mixed-list benefit* refers to higher memory for items studied using an elaborative encoding task in a mixed list compared to a pure list. A *mixed-list cost* refers to lower memory for items studied using a non-elaborative encoding task in a mixed-list compared to a pure-list context. Using a production effect paradigm, Bodner et al. (2014) compared recognition responses for items that were encoded either aloud (i.e., elaborative) or silently in both mixed and pure lists. Consistent with a design effect, a robust production effect was found for mixed lists, and this difference was greater than the production effect found for pure lists. Importantly, when aloud and silent items were compared across mixed and pure lists, hits were higher for aloud items in mixed lists than pure lists – a mixed-list benefit – but hits were lower for silent items in mixed lists than pure lists – a mixed-list cost. This pattern is important because it indicates that mixed-list production does not operate cost-free and can produce a memory deficit for the impoverished silent encoding task (see too, Begg & Roe, 1988; Slamecka & Katsaiti, 1987, for design effects and mixed-list costs/benefits in generation).

Akin to production and generation, the *drawing effect* refers to a similar memory improvement in which drawing an image of a word's referent enhances memory relative to silent reading or writing of a word (Fernandes et al., 2018; Namias et al., 2022; Wammes et al., 2016). Drawing individual images has been hypothesized to promote the processing of item-specific information through three processes: Elaboration, motor action, and pictorial processing (Fernandes et al., 2018; Namias et al., 2022). These processes may produce a concomitant reduction in relational/interitem associations that may contribute to design effect costs and benefits in production and generation. Consistent with this possibility, Wammes et al. (2016; Experiments 6 and 7) compared two pure-list groups who either drew images of a word's referent or wrote the word repeatedly for a similar study duration, and a separate mixed-list group who completed both tasks on the same list. A drawing effect on free recall emerged in both pure-list and mixed-list comparisons, but the mixed-list drawing effect was larger than the pure-list drawing effect, consistent with previous design patterns.

More recently, Jonker et al. (2019) evaluated the design effect between pure and mixed lists using study lists that were either short (eight words) or long (20 or 34 words). Drawing was compared to a silent-reading control task. Mixed lists were again found to produce larger drawing

effects than pure lists when lists were long, but a null and even a *reversed* drawing effect was found on pure-short lists. The authors attributed this reversed drawing effect to greater memory for relational item-order information for short lists as participants were better able to reconstruct the list sequencing presented at study, which restricted memory improvements due to drawing. Additionally, the authors computed estimates of memory sequencing by computing interitem associations (likelihood of recalling words that were presented consecutively at study) and recall distance (mean serial position distance between consecutively recalled words). Memory sequencing was found to be greater in pure silent reading lists, providing supportive evidence that drawing, whether completed in a pure list or a mixed list, disrupted memory sequencing information. This pattern is consistent with the notion that drawing operates to strengthen item-specific information but results in a decrement to relational interitem information.

In the present study, we similarly examined the drawing effect in pure- and mixed-list designs with the goal of directly assessing the costs and benefits associated with drawing in different contexts. We note that Wammes et al. (2016) and Jonker et al. (2019) report discrepant findings regarding a pure-list drawing effect. Specifically, Wammes et al. reported a reliable pure-list drawing effect that was smaller than the drawing effect found in mixed lists, but Jonker et al. found no pure-list drawing effect in two experiments and a reversed drawing effect in another. We further test for the presence of a pure-list drawing effect across three experiments in both recognition and recall but, given evidence that the drawing effect is diminished in pure lists versus mixed lists, we test for the presence of mixed-list costs and benefits by directly comparing draw and write items between both designs.

The costs and benefits of drawing on mixed versus pure lists have not yet been compared in the literature, but researchers have examined costs and benefits in the generation and production effects. For example, Begg and colleagues examined whether the robust mixed-list generation effect was due to a "true" benefit of generation (i.e., a benefit of generation in a mixed list vs. a pure list) or due to a cost to read items (i.e., a cost to reading intact items in a mixed list vs. a pure list; Begg & Roe, 1988; Begg & Snider, 1987). Begg and Snider argued that generation creates a memory criterion of identifying memory items as independent entities. In a mixed-list context, this criterion can encourage cursory encoding of read items (i.e., "lazy reading") in which identification of generate items is less effortful. Indeed, comparisons between mixed-list and pure-list contexts indicated that the robust mixed-list generation effect reflected costs to read items and not benefits to generate items (but see Begg et al., 1989, for evidence of both costs and benefits using related pairs and categorized word lists). This pattern was echoed in production effect studies (Bodner et al., 2014; Jones & Pyc, 2014), in which a mixed-list production effect was reflective of a cost to items that were read silently in mixed lists rather than of a benefit to aloud items. Only when silent and aloud items were blocked consecutively in a mixed list in the Bodner et al. study was this cost eliminated.

Our study therefore evaluated whether the mixed-list drawing effect reflected enhanced memory for draw items and/or a memory decrement for write items compared to pure-list groups. In Experiment 1, we evaluated the costs and benefits of drawing on an immediate recognition test by comparing a mixed-list group who completed drawing and writing tasks on a single list of words (Experiment 1A), to pure-list groups who completed only either drawing or writing on a list (Experiment 1B). Across experiments, we anticipated a reliable design effect in which the drawing effect would be larger for mixed than pure lists. Additionally, we expected that this design effect would emerge even if a pure-list drawing effect occurred (cf. Wammes et al., 2016). This prediction is based on both a strength-based account, in which drawing facilitates the encoding of items through recruitment of semantic, motor, or perceptual features (Fernandes et al., 2018), and a distinctiveness account, in which mixed lists show an enhanced drawing benefit due to the presence of distinctive and non-distinctive information (e.g., MacLeod et al., 2010). Finally, we expected that the larger drawing effect in mixed lists would reflect a pattern of costs and benefits when compared to pure lists. Specifically, mixed-list drawing would not produce a benefit over pure-list drawing, but that a cost would emerge for mixed-list writing relative to pure-list writing. This prediction is based on patterns reported in both the generation and the production effect literatures (cf. Begg & Snider, 1987; Hopkins & Edwards, 1972), and is consistent with a cursory-reading account of the design effect (Bodner et al., 2014). Collectively, although drawing was anticipated to be beneficial for enhancing memory for items that were drawn, drawing would not operate cost-free and would negatively affect memory for written items in mixed lists.

Experiment 1A: Drawing versus writing mixed-list immediate recognition

Method

Participants

A total of 32 participants were recruited from undergraduate Psychology courses at The University of Southern Mississippi and received partial course credit for participation. All were fluent English speakers with normal or corrected-to-normal vision. Due to a computer error, data from one

Table 1 Lexical and semantic characteristics for study words used in Experiments 1–3

Variable/descriptives	*M*	*SD*	*Min*	*Max*
Number of letters	6.39	1.42	5.00	12.00
SUBTLEX frequency	2.66	0.66	0.70	4.51
Concreteness	588.65	30.84	504.00	645.00
Age of acquisition	5.70	1.47	3.23	11.21
Body-object interaction	5.41	0.96	2.30	6.88

Characteristics are taken from the English Lexicon Project (Balota et al., 2007) which included SUBLTEX word frequency (Brysbaert & New, 2009), Concreteness (Brysbaert et al., 2014), Age of Acquisition ratings (Kuperman et al., 2012), and Body-Object Interaction ratings (Pexman et al., 2019)

participant were not recorded, leaving 31 participants available for analysis. Mean participant age was 22.65 years (*SD* = 8.25; *range* = 18–58), a mean of 13.74 years of formal education was reported (*SD* = 2.38; *range* = 12–20), and 54.84% of participants identified with the female gender. A sensitivity analysis using *G*POWER 3* (Faul et al., 2007) indicated that the sample had adequate statistical power (0.80) to detect medium effect sizes of Cohen's $d = 0.52$ or larger, two-tailed, using a repeated-measures design with a *t*-distribution. This sample size is consistent with or even exceeds sample sizes used in other drawing effect studies (e.g., Jonker et al., 2019; Wammes et al., 2016).

Materials

Eighty unrelated words were selected from the English Lexicon Project (Balota et al., 2007) with the specification that the words were high in concreteness to aid in drawing (see Table 1 for semantic and lexical characteristics of the selected words). Words were divided into two, 40-item sublists that were matched on word frequency using the SUBTLEX norms (Brysbaert & New, 2009) and concreteness (Brysbaert et al., 2014). Participants were presented with one 40-item sublist at study. The two sublists were counterbalanced across participants. The two 40-item sublists were further segmented into two 20-item lists (also matched on frequency and concreteness) that were used for draw or write items. One set of 20 items was presented in blue font, and the other set of 20 items was presented in red font. The colors corresponded to words that would be drawn or written. The 20-item sublists that were associated with each color and task were counterbalanced across participants such that, across counterbalanced versions, all study words were presented in all of the possible color/task combinations. An 80-item recognition test was then created that included the 40 studied items, 20 of which corresponded to draw words and 20 of which corresponded to write items, and the 40 items from the non-studied sublist, which served as distractors. All participants therefore saw the same recognition test items, but the items that were studied (vs. non-studied) depended upon which sublist was studied. Recognition test items were displayed in black font and presented in a newly randomized order for each participant.

Procedure

Participants were tested individually with an experimenter present using a computer running SuperLab 6 software (Cedrus, 2022). Following informed consent, participants completed a brief demographics questionnaire that was followed by experimental instructions. Participants were instructed that they would study a series of words on the computer screen that would be displayed individually and presented in either red font or blue font. Encoding instructions were closely modeled after those used by Wammes et al. (2016). Specifically, participants were informed that for half of the words, they would draw an image of the word's referent into a box presented on a sheet of paper for 10 s. Time was measured by the experimenter, beginning as soon as the participant's pencil touched the paper. Participants were directed to start their drawings as soon as the next word appeared on the computer screen. The experimenter would prompt the participant to stop drawing when the 10-s period ended and the next word appeared. For the other half of the words, participants were asked to write each individual study word on a separate sheet of paper with lines and to write one word per line. Participants were told they would have up to 10 s to write each word out, but the experimenter would prompt the participant to stop writing the word and advance to the next word when the time ended. Depending on the counterbalanced version the participant was assigned, draw words would be presented in blue font and write words in red font or vice versa. During each encoding trial, participants were only presented with the individual word presented in blue or red font on the computer screen. However, participants were provided with an instruction sheet on the table in front of them reminding them which color corresponded to either the draw or the write task. The experimenter advanced to the next word at the end of the 10-s study interval.

Immediately following the study phase, all participants were presented with an 80-item old/new recognition test. Participants were instructed to classify test items as "old" if they were presented on the study list, and "new" if they were not presented on the study list. Recognition responses were measured using the keyboard, with participants pressing the "O" key for old and the "N" key for new. The experiment took approximately 20 min to complete and was followed by debriefing and allocation of course credit.

Experiment 1B: Drawing versus writing pure-list immediate recognition

Method

Participants

Sixty-four University of Southern Mississippi undergraduates participated for partial course credit. All participants reported normal or corrected-to-normal vision. Participants were randomly assigned to either the draw group ($n = 32$) or the write group ($n = 32$). Data from one participant in the write group were unavailable due to a computer error. Mean age was 19.90 years ($SD = 4.65$; *range* = 18–51) with 12.92 mean years of reported formal education ($SD = 1.22$; *range* = 12–16), and 76.19% identified with the female gender. A sensitivity analysis indicated that the sample had adequate statistical power (0.80) to detect effect sizes of Cohen's $d = 0.71$ or larger, two-tailed, using an independent-samples design with a *t*-distribution.

Materials and procedure

All materials and procedures in Experiment 1B were the same as Experiment 1A with the following exceptions. Specifically, the two 20-item sublists used in 1A to separate mixed draw and write items were combined into a single 40-item study list in which participants only completed either the draw task for all words or the write task for all words. Each study list consisted of items that were presented in both blue and red font color as in Experiment 1A; however, participants were told to draw or write each word regardless of the font color the word was presented in. We used both colors in pure lists to match the color presentations that were used in Experiment 1A. Colors may have served as an additional retrieval cue for items, and therefore using the same color presentations kept these retrieval cues constant between the mixed and pure lists. Font colors and task types were again counterbalanced across participants and the recognition test was identical to Experiment 1A with the exception that studied test words only corresponded to either draw or write tasks.

Results: Experiments 1A and 1B

For all results reported, a $p < .05$ criterion was used. For brevity, null-hypothesis-significance testing *p*-values are not reported for significant comparisons. Instead, we supplement our standard NHST analyses with Bayesian hypothesis testing. The Bayesian test computes a Bayes factor (BF), a numerical value that quantifies the predictive capacity of the null hypothesis model (H_0) compared to the alternative hypothesis model (H_1). For reported BFs, reported subscripts correspond to the favored hypothesis, either H_1 over H_0 (BF_{10}) or H_0 over H_1 (BF_{01}). Several interpretive criteria for BFs have been proposed; we follow interpretive values reported by van Doorn et al. (2021). For null hypothesis evidence, BF_{10}s less than 1/10 suggest strong evidence for the null, BF_{10}s between 1/10 and 1/3 suggest moderate evidence for the null, and BF_{10}s between 1/3 and 1 suggest weak evidence for the null. For alternative hypothesis evidence, BF_{10}s greater than 10 suggest strong evidence for the alternative, BF_{10}s between 3 and 10 suggest moderate evidence, and BF_{10}s between 1 and 3 suggest weak evidence. We caution, however, against applying these interpretive values as all-or-none cutoffs for making data conclusions, and that BFs should not be conflated with general estimates of effect size. We include effect size estimates by computing Cohen's *d* for each comparison. Proportions of correct recognition of studied items (hits) and false alarms to non-studied distractors as a function of draw and write instructions in Experiments 1A and 1B are reported in Table 2 (top panel).

Recognition responses for draw and write items were first compared within each experiment.[1] Starting with correct recognition, a robust drawing effect was found in the mixed lists (Experiment 1A), in which correct recognition of draw items exceeded that of write items (.97 vs. .69, for draw and write means, respectively), $t(30) = 9.27$, $d = 1.66$, $BF_{10} = 4.02 \times 10^7$. A corresponding drawing effect emerged in the pure-list comparison (Experiment 1B), in which correct recognition of draw items exceeded write items (.97 vs. .82), $t(61) = 6.75$, $d = 1.70$, $BF_{10} = 1.42 \times 10^6$. In both experiments, correct recognition for draw items was at ceiling and false alarm rates were at floor. False alarm rates could not be segregated for draw and write items in Experiment 1A as distractors could not be yoked to a specific study task ($M = .01$). However, in Experiment 1B, false alarm rates were lower in the draw group than in the write group (.01 vs. .06), $t(61) = 3.45$, $d = 0.87$, $BF_{10} = 30.49$.

Turning to a cost/benefit analysis, correct recognition for draw items did not differ in the Experiment 1A mixed list compared to the Experiment 1B pure list (.97 vs. .97), $t < 1$, $BF_{10} = 0.26$, showing neither a benefit nor a cost between design types. This pattern may have been due to a restricted range from ceiling performance of draw items in both design types. For write items, however, correct recognition in Experiment 1B pure lists was greater than correct recognition in Experiment 1A mixed

[1] Because hit and false alarm rates were perfect for many participants (1.00 or 0.00, respectively), we focus our analyses on raw recognition rates versus a hits-minus-false-alarms correction or d' using signal detection. A signal-detection analysis requires a correction for asymptotic values (e.g., MacMillan & Creelman, 1988), and most participants would have required at least one corrected score to compute d' due to ceiling hits or floor false alarms which would have distorted the dataset.

Table 2 Mean (± 95% CI) proportions of "old" recognition responses to studied list items (hits) and distractors (false alarms) in Experiment 1 (immediate recognition test) and Experiment 2 (delayed recognition test) and proportions of correct recall and mean number of intrusions recalled in Experiment 3

	Experiment 1A Within Recognition		Experiment 1B Between Recognition		Design effect (W/in-Btw)		Drawing effect (Draw-Write)	
Item type/group	Draw	Write	Draw	Write	Draw	Write	Within	Between
N	31		32	31				
Studied items (Hits)	.97 (.01)	.69 (.06)	.97 (.01)	.82 (.04)	.00	-.12*	.28	.16
Distractors (False alarms)	.01 (.01)		.01 (.01)	.06 (.02)	.00	-.04	—	-.04
	Experiment 2A Within Recognition		Experiment 2B Between Recognition		Design effect (W/in-Btw)		Drawing effect (Draw-Write)	
Item type/group	Draw	Write	Draw	Write	Draw	Write	Within	Between
N	31		31	31				
Studied items (Hits)	.91 (.04)	.56 (.07)	.84 (.05)	.68 (.04)	.07*	-.12*	.35	.16
Distractors (False alarms)	.16 (.04)		.06 (.03)	.31 (.05)	.09	-.15	—	-.24
	Experiment 3A Within Recall		Experiment 3B Between Recall		Design effect (W/in-Btw)		Drawing effect (Draw-Write)	
Item type	Draw	Write	Draw	Write	Draw	Write	Within	Between
N	33		32	32				
Correct recall	.57 (.05)	.16 (.03)	.45 (.04)	.27 (.04)	.12*	-.10*	.41	.19
# Intrusions recalled	0.21 (.14)		0.56 (.28)	0.47 (.26)	-0.35	-0.26	—	0.09

The ***Design effect*** refers to the between-subject minus within-subject means for draw and write items. Positive means refer to memory *benefits* of within-subject designs, negative means refer to memory *costs* of within-subject designs. * = benefits and costs to correct recognition/recall, $p < .05$. The ***Drawing effect*** refers to the benefits of draw items over write items in the within-subject and between-subject experiments

lists (.82 vs. .69), $t(60) = 3.11$, $d = 0.79$, $BF_{10} = 12.89$, revealing a *cost* to writing when completed in a mixed-list context. Thus, there was a larger drawing effect found in correct recognition in a mixed list compared to a pure list (.28 vs. .16), which reflected a recognition cost to mixed-list write items.

A cost/benefit analysis could not be computed for false alarms as only a single false alarm rate was available for mixed-list recognition. However, for completeness, we compared false alarm rates between experiments. The mixed-list (Experiment 1A) false alarm rate was equivalent to the false alarm rate for the draw group in Experiment 1B (.01 vs. .01), $t < 1$, $BF_{10} = 0.28$, but lower than the false alarm rate for the write group in Experiment 1B (.01 vs. .06), $t(60) = 3.56$, $d = 0.90$, $BF_{10} = 40.06$. The presence of draw items, in either mixed or pure lists, therefore reduced false alarms relative to writing only, but these patterns were at floor.

Discussion

Experiments 1A and 1B were designed to evaluate the effects of drawing and writing on recognition in mixed- and pure-list designs. A robust mixed-list drawing effect was found on correct recognition in Experiment 1A. A drawing effect was also found on pure lists in Experiment 1B, an important finding given pure-list drawing effects have not been found consistently (cf. Jonker et al., 2019; Wammes et al., 2016). While drawing yielded recognition improvements in both list types, a design effect consistent with those in generation and production (e.g., Bertsch et al., 2007; Bodner et al., 2014) emerged: The better recognition for drawing than for writing was greater for mixed than for pure lists, though correct recognition following drawing was at ceiling in both list types.

A cost/benefit analysis was then used to identify the loci of the design effect by directly comparing draw and write items between mixed- and pure-list contexts. Although drawing produced equivalent correct recognition rates in mixed and pure lists (i.e., neither a benefit nor a cost), correct recognition of write items was lower in mixed lists than in pure lists. This pattern indicates that the design effect in drawing is at least partially attributable to a cost to mixed-list writing, suggesting that participants are engaging in cursory reading. Performance for draw items in both mixed- and pure-list contexts was robust and at ceiling (.97 in both contexts), making it impossible to evaluate whether mixed-list drawing may also result in costs, or even benefits.

To further evaluate costs and benefits to drawing mixed and pure lists, we conducted a separate experiment that implemented a 48-h retention interval between the study and test phases. The purpose of this delay was twofold. First, given ceiling performance on correct recognition of draw items on an immediate test, a longer retention interval would likely produce forgetting of the study list, which would pull performance off ceiling. Indeed, Gardiner and Java (1991) reported

that delays as short as 24 h were successful at reducing correct recognition approximately 25% relative to an immediate test, and that forgetting was more likely to affect recollection-based processes than familiarity using a remember/know paradigm. We note that Gardiner and Java's experiments used a much larger set of study items than Experiment 1 above (72 vs. 40), and, therefore, we used a longer 48-h retention interval to increase the likelihood of forgetting. Secondly, the use of a delay provides an additional evaluation of the drawing effect in pure and mixed lists. Although reliable drawing effects have been reported following a delay (see Fernandes et al., 2018), to our knowledge, no study has evaluated delay effects on drawing in both pure and mixed lists or whether design effects may reflect a pattern of costs and benefits.

Experiment 2A: Drawing versus writing mixed-list delayed recognition

Participants

Thirty-one University of Southern Mississippi students participated for partial course credit. All were fluent English speakers with normal or corrected-to-normal vision. Mean age was 20.90 years (*SD* = 5.09; *range* = 18–41) with a mean of 13.52 years of formal education reported (*SD* = 1.90; *range* = 12–20). Of the sample, 74.19% identified with the female gender. A sensitivity analysis indicated that the sample had adequate statistical power (0.80) to detect medium effect sizes of Cohen's $d = 0.52$ or larger, two-tailed, using a repeated-measures design with a *t*-distribution.

Materials and procedure

All materials and procedures as those used in Experiment 1A were identical except for a delay that was inserted between the study and recognition test phases. Specifically, immediately following the study phase, participants were dismissed from the experimental session with the instruction to return to the same research site to complete the test phase 48 h later. Following the same 80-item old/new recognition test after the delay, participants were debriefed and compensated with course credit.

Experiment 2B: Drawing versus writing pure-list delayed recognition

Participants

Sixty-two University of Southern Mississippi students participated for partial course credit. All were fluent English speakers with normal or corrected-to-normal vision. Participants were randomly assigned to either the draw group ($n = 31$) or the write group ($n = 31$). Mean age was 21.04 years (*SD* = 4.44; *range* = 18–42), participants reported a mean of 13.63 years of formal education (*SD* = 1.85; *range* = 12–19), and 62.90% identified with the female gender. A sensitivity analysis indicated that the sample had adequate statistical power (0.80) to detect effect sizes of Cohen's $d = 0.72$ or larger, two-tailed, using an independent-samples design with a *t*-distribution.

Materials and procedure

Materials and procedures were the same as those used in Experiment 2A with the exception that participants completed either the draw task or the write task for all study items. Recognition testing was similarly completed after a 48-h delay.

Results: Experiments 2A and 2B

Proportions of study item hits and false alarms to distractors as a function of draw and write instructions for Experiments 2A and 2B are reported in Table 2 (middle panel).

As in Experiment 1, recognition for draw and write items were first compared within each experiment. For correct recognition, a large drawing effect was again found in the Experiment 2A mixed lists (.91 vs. .56, for draw and write means, respectively), $t(30) = 9.49$, $d = 1.70$, $BF_{10} = 6.66 \times 10^7$, and a smaller, but potent, drawing effect was found in the Experiment 2B pure lists (.84 vs. 68), $t(60) = 5.13$, $d = 1.30$, $BF_{10} = 4606.53$ – patterns found in Experiment 1. False alarms were higher following a delay in Experiment 2A mixed lists ($M = .16$) and were lower for the Experiment 2B draw group than the write group (.06 vs. .31), $t(60) = 8.12$, $d = 2.06$, $BF_{10} = 4.97 \times 10^{14}$.

A cost/benefit analysis was then conducted to test for design effects in drawing with performance now off ceiling. Importantly, and unlike Experiment 1, correct recognition for draw items was greater in Experiment 2A mixed lists than Experiment 2B pure lists (.91 vs. .84), $t(60) = 2.24$, $d = 0.57$, $BF_{10} = 2.06$ – a mixed-list drawing benefit. For write items, correct recognition in Experiment 2A mixed lists was lower than Experiment 2B pure lists (.56 vs. .68), $t(60) = 2.92$, $d = 0.74$, $BF_{10} = 4.10 \times 10^7$ – a mixed-list writing cost that was similarly found in Experiment 1. Collectively, the larger drawing benefit that was found in a mixed-list design versus a pure-list design (.35 vs. .16) reflected a combination of benefits to drawing and costs to writing in a mixed-list context.

False alarm rates to non-studied test items were also compared between experiments. False alarms were higher in the Experiment 2A within group than the Experiment 2B draw group (.16 vs. .06), $t(60) = 4.08$, $d = 1.04$, $BF_{10} = 168.85$, but lower in the Experiment 2A within group than the Experiment 2B write group (.16 vs. .31), $t(60) = 4.73$, $d = 1.20$, $BF_{10} = 1258.63$. Thus, false alarm rates also appeared to produce both costs and benefits depending on the experimental design used. Specifically, and consistent with Experiment 1, the presence of draw items reduced false alarms, especially in the Experiment 2B pure draw group, in which write items were absent.

Discussion

Experiments 2A and 2B further evaluated the drawing effect in recognition in pure and mixed-list designs, but with a 48-h retention interval. Consistent with Experiment 1, a robust drawing effect was found that included a design effect in which the drawing improvement over writing was larger in a mixed- than pure-list context. The delay was successful at reducing recognition performance and under this condition, a costs/benefits analysis revealed that the larger mixed-list drawing effect was due to both lower correct recognition for write items in mixed than pure lists – a mixed-list cost – and higher correct recognition for draw items in mixed than pure lists – a pure-list benefit.

We discuss this pattern of costs and benefits further in the *General discussion*, but first we sought to replicate and extend these findings using free-recall test, which typically produces more moderate memory performance than recognition (Schonfield & Robertson, 1966) and should also keep performance off ceiling. Relative to recognition, free recall is a recollection-heavy task (Yonelinas, 2002), which provides less support to aid retrieval. Additionally, most drawing effect studies have used free recall (Fernandes et al., 2018; Meade et al., 2018), including Jonker et al. (2019), who did not find a pure-list drawing effect. Therefore, evaluating the costs and benefits of drawing on free recall improves cohesion in the literature including a design comparison that more closely aligns with previous work.

In addition to standard analyses on proportions of correct recall, the use of free-recall testing allows for a series of qualitative analyses to assess retrieval dynamics between writing and drawing tasks. In our first analysis, memory sequencing information will be analyzed by computing lag conditional response probability (lag-CRP) functions using serial position information for studied items. Lag-CRPs plot the probability of correct recall for items presented sequentially at study based on the lag (i.e., distance) over serial positions (Howard & Kahana, 1999). Lag-CRPs typically indicate that recall is highest in adjacent lag (-1 or +1, referring to recalled words presented immediately before or after a given word, respectively), and decreases over greater lags. A notable feature of lag-CRPs is that they are typically asymmetrical – showing greater recall in the forward than backward direction (Polyn et al., 2009; Wahlheim & Huff, 2015). These functions are therefore informative regarding memory sequencing, and we extend lag-CRPs to pure lists in Experiment 3B to compare memory order for drawing and writing tasks. Lag-CRPs are applied only in pure lists because mixed lists randomly alternate between drawing and writing tasks and this task switching may affect memory sequencing. If drawing disrupts memory sequencing, as reported by Jonker et al. (2019), a reduction in lag probabilities would be expected for drawing, but only for adjacent positions.

A second analysis will compute the mean serial position difference between adjacently recalled items at test (i.e., recall distance), an analysis used by Jonker et al. (2019) to estimate memory order. While both lag-CRPs and recall distance compute the likelihood that individuals order recall by serial position, lag-CRPs provide a more granular measure by computing mean order probabilities for the five preceding and five subsequent lags. Despite these computational differences, the two metrics should yield similar conclusions regarding memory order for drawing and writing.

Finally, we include plots of serial position functions for writing and drawing tasks (Murdock, 1962). While serial position curves are not necessarily informative regarding item sequencing, differences in the shape of these curves might provide information regarding task differences in primacy and recency effects. For example, Huff and Bodner (2019) reported that relative to a read-only control task, deep encoding tasks resulted in larger recency effects, suggesting that deep processing increases the availability of information in short-term/working memory. If drawing operates similarly to other deep tasks, similar recency patterns may be expected. Standard analyses of mean recall rates and intrusions are therefore supplemented with retrieval dynamics to further examine organizational processes following drawing and writing.

Experiment 3A: Drawing versus writing mixed-list recall

Method

Participants

Thirty-three University of Southern Mississippi students participated for partial course credit. All were fluent English speakers with normal or corrected-to-normal vision. Mean age was 21.48 years ($SD = 8.81$; *range* = 18–60), a mean of 12.79 years of formal education was reported ($SD = 1.90$; *range* = 12–20), and 66.67% of the sample identified with the female gender. A sensitivity analysis indicated that the sample had adequate statistical power (0.80) to detect medium effect sizes of Cohen's $d = 0.50$ or larger, two-tailed, using a repeated-measures design with a *t*-distribution.

Materials and procedure

All materials and procedures were the same as those used in Experiment 1A except for the memory test. Specifically, after participants completed the study phase, they were instructed that they would complete a 5-min free-recall test. Participants were given a sheet of paper with 40 blank spaces and were asked to recall the words that they studied in any order and without cost for incorrect spellings. Participants were neither encouraged nor discouraged from guessing and no mention of the two study instructions occurred to minimize the likelihood that an additional retrieval cue would be provided. Participants were timed by the experimenter and cued when to end their recall.

Experiment 3B: Drawing versus writing pure list recall

Method

Participants

Sixty-four University of Southern Mississippi undergraduates participated for partial course credit. All participants reported normal or corrected-to-normal vision. Participants were randomly assigned to either the draw group ($n = 32$) or the write group ($n = 32$). Mean age was 20.08 years ($SD = 3.35$; *range* = 18–35) with 13.22 mean years of reported formal education ($SD = 1.59$; *range* = 11–17), and 79.69% identified with the female gender. A sensitivity analysis indicated that the sample had adequate statistical power (0.80) to detect effect sizes of Cohen's $d = 0.71$ or larger, two-tailed, using an independent-samples design with a *t*-distribution.

Materials and procedure

Materials and procedures were the same as those used in Experiment 3A with the exception that participants completed either the draw task or the write task for all 40 items.

Results

Correct recall and intrusions

Mean proportions of correct recall were computed as the total number of words correctly recalled, divided by the total number of words presented in each study condition (e.g., in Experiment 3A, recalled draw and write words were each divided by 20). Correct recall was scored using a relatively liberal criterion in which words that were misspelled, including pluralizations, were counted as correct. The same scoring criterion was applied to both experiments. Correct recall of studied items and mean numbers of intrusions falsely recalled per list as a function of draw and write instructions in Experiments 3A and 3B are reported in Table 2 (bottom).

Analyses of retrieval dynamics for free-recall responses for draw and write tasks were also conducted, which include plots of lag-CRPs (Howard & Kahana, 1999), mean serial position recall distance between adjacent items recall (Jonker et al., 2019), and serial-position curves (Murdock, 1962; see Crowder, 1976, for review) to provide additional information regarding task effects on memory sequencing. Recall dynamics were unavailable for one participant in the pure draw group due to a computer error which failed to record serial positions from the encoded list. Proportions of correct recall and intrusions were still available, and therefore, we only excluded this participant for the recall dynamics analyses.

Proportions of correct recall for draw and write items were first compared within each experiment. A robust drawing effect was again found in both a mixed-list context (Experiment 3A; .57 vs. .16, $t(32) = 14.81$, $d = 2.58$, $BF_{10} = 8.39 \times 10^{12}$), and a pure-list context (Experiment 3B; .45 vs. .27; $t(62) = 6.95$, $d = 1.74$, $BF_{10} = 3.24 \times 10^{6}$). Mean numbers of extra-list intrusions recalled did not differ between the draw and write groups in Experiment 3B (.56 vs. .47), $t < 1$, $BF_{10} = 0.28$.

A cost/benefit analysis was then conducted that first compared correct recall proportions for mixed-list draw items (Experiment 3A) to pure-list draw items (Experiment 3B). Consistent with Experiment 2, a significant *benefit* to mixed-list drawing was found, in which recall of Experiment 3A draw items was higher than Experiment 3B draw items (.57 vs. .45), $t(63) = 3.82$, $d = 0.95$, $BF_{10} = 83.59$. This drawing benefit was also accompanied with a *cost* to write items. Specifically, recall of Experiment 3A write items was significantly lower than that of Experiment 3B write items (.16 vs. .27), $t(63) = 4.18$, $d = 1.04$, $BF_{10} = 239.61$. Thus, the design effect, which indicated a larger drawing improvement in mixed lists than in pure lists (.41 vs. .19), reflected a combination of benefits to draw items and costs to write items.

Extra-list intrusions were also compared between experiments. Overall, intrusions were rare, but mean number of intrusions was lower in the Experiment 3A draw group than the Experiment 3B draw group (.21 vs. .56), $t(63) = 2.23$, $d = 0.55$, $BF_{10} = 1.99$, and marginally lower than in the Experiment 3B write group (.21 vs. .47), $t(63) = 1.69$, $p = .10$, $d = 0.42$, $BF_{10} = 0.85$.

Memory for order

Figure 1 plots lag-CRP functions for Experiment 3B as a function of task type. CRPs were analyzed using a 2(Task: Drawing vs. Writing) × 10(Lag) mixed analysis of variance (ANOVA). An effect of lag was found, $F(9, 558) = 4.36$, $\eta_p^2 = .07$, $BF_{10} = 33.80$, confirming a pattern of stronger recall

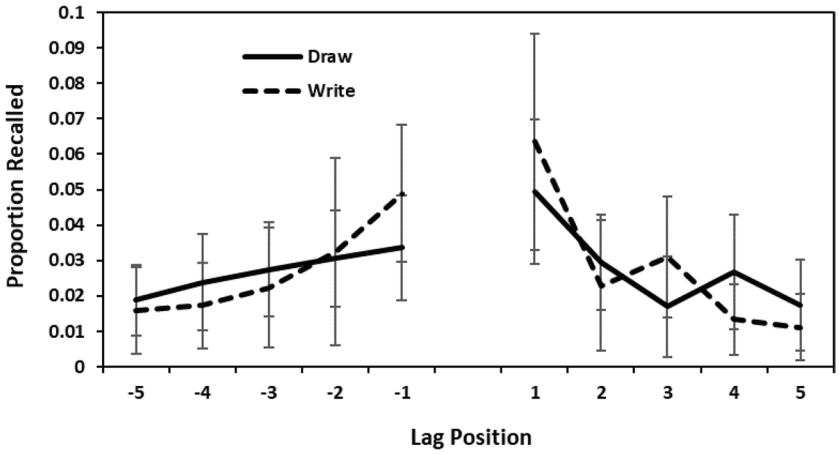

Fig. 1 Experiment 3B: Conditional response probabilities as a function of lag (-5 to +5) for draw and write pure lists. Error bars are 95% confidence intervals

in adjacent lag positions, particularly in the forward (lag +1) direction. However, the effect of task, $F < 1$, $BF_{10} = 0.46$, and the interaction, $F < 1$, $BF_{10} = 0.01$, were unreliable. This pattern indicates that memory sequencing effects were invariant for drawing and writing tasks.

Mean serial position recall distance was also computed for draw and write groups. Consistent with lag-CRPs, mean recall distance was equivalent between the draw group (M = 13.11 positions; $SD = 2.39$) and the write group ($M = 13.72$ positions; $SD = 4.29$), $t < 1$, $BF_{10} = 0.32$, providing further evidence that recall ordering was not sensitive to task type.

Turning to serial position effects, Fig. 2 plots correct recall proportions for each of the 40 serial positions for the draw and write tasks in Experiment 3B. A 2(Task: Drawing vs. Writing) × 40(Serial Position) mixed ANOVA was used. Consistent with overall proportions of correct recall above, a pure-list drawing effect was found with greater recall for draw than write lists, $F(1, 62) = 32.06$, $\eta_p^2 = .34$, $BF_{10} = 8.63 \times 10^{15}$. Differences were also found across serial position, $F(39, 2418) = 4.45$, $\eta_p^2 = .07$, $BF_{10} = 4.70 \times 10^4$, which confirmed the presence of primacy and recency effects. The interaction was at the conventional level of significance, $F(39, 2418) = 1.40$, $p = .050$, $\eta_p^2 = .02$, $BF_{10} = 0.02$. Follow-up comparisons indicated that this interaction reflected a slightly larger recency effect for the draw group than the write group, though this difference was small and often did not survive adjustments for multiple comparisons. We therefore interpret this interaction with caution given this correction.

Discussion

The results of Experiments 3A and 3B are quite clear. First, a drawing effect on recall was found in both mixed lists and pure lists, replicating the recognition drawing effects in Experiments 1 and 2. A design effect was also found in which the drawing improvement was larger in mixed than pure lists, and this pattern extended to recall. Second, a costs/benefits analysis indicated that the larger mixed-list drawing effect reflected contributions of both costs and benefits. Recall of mixed-list write words was lower than recall of pure-list write words – a cost to writing that was also present in Experiments 1 and 2. Importantly, recall of mixed-list draw words was greater than recall of pure-list draw words, a drawing benefit that was also present in delayed recognition. Collectively, the design effect found with drawing in free recall also reflects a contribution of a cost to mixed-list write words and a benefit to mixed-list draw words when compared to pure lists.

Finally, we examined memory for order sequencing in pure lists by evaluating recall dynamics of drawing and writing using lag-CRPs, recall distance, and serial position curves. Although lag-CRPs indicated that participants order their recall to adjacent serial positions, particularly in the forward direction (cf. Howard & Kahana, 1999; Wahlheim & Huff, 2015), there were no differences in sequencing between draw and write tasks. This invariance in memory sequencing was also found when computing mean recall distance. These patterns contrast with Jonker et al. (2019), who reported that a characteristic of drawing is that it disrupts memory sequencing. We discuss this discrepancy in greater detail in the *General discussion* but note here that there are several differences between our study and Jonker et al., including the type of comparison control group used, which may have contributed to these differences. Regarding serial position effects, both writing and drawing tasks produced standard serial position patterns, though recall rates were greater across positions for draw lists, particularly in later positions, though this effect was modest.

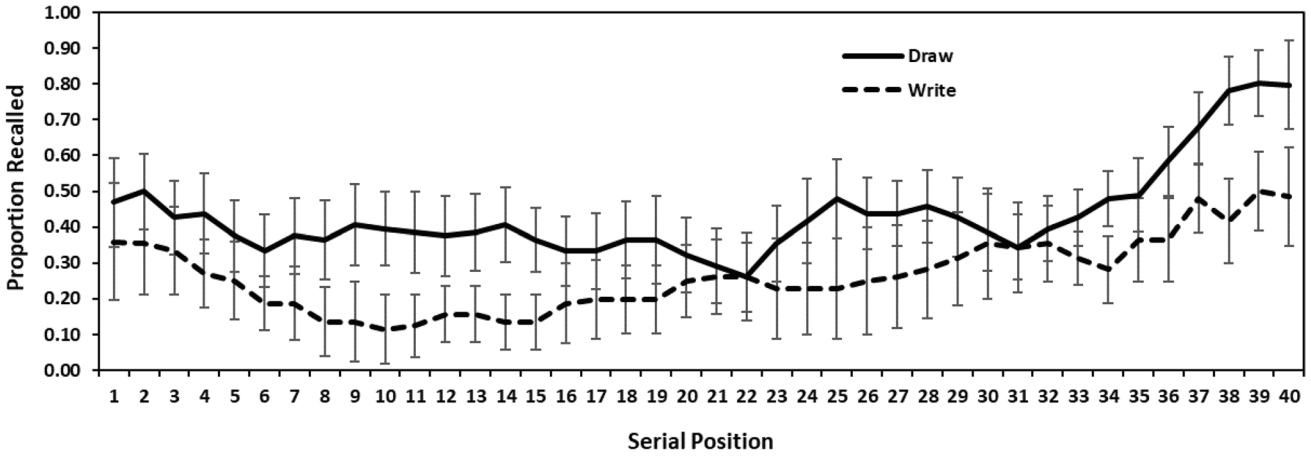

Fig. 2 Experiment 3B: Serial position curves as a function of encoding position for draw and write pure lists. Data were averaged across adjacent positions. Error Bars are 95% confidence intervals

General discussion

Our experiments provide several new contributions to our understanding of the drawing effect in recognition and recall. First, a reliable pure-list drawing effect was found in both recognition and recall test types across three experiments, an effect that has not been found consistently (e.g., Jonker et al., 2019). Second, a reliable design effect was also found in both recognition and recall in which the drawing improvement over writing was greater in a mixed-list design than in a pure-list design. This pattern is particularly notable from our experiments as the design effect was robust despite the presence of a pure-list drawing effect. Third, we examined the loci of the design effect in mixed-list drawing by conducting a cost/benefit analysis that directly compared drawing and writing tasks across design types (e.g., Bodner et al., 2014; Hopkins & Edwards, 1972). In immediate recognition (Experiment 1), our results indicated that the larger drawing effect on mixed lists included a cost to write items, but neither a benefit nor a cost to draw items, though a ceiling effect on recognition restricts this conclusion. When recognition was off ceiling due to a delayed test (Experiment 2), the larger mixed-list drawing effect reflected both a cost to mixed-list write items and a benefit to mixed-list draw items. This pattern was also found in free recall (Experiment 3), where again, the larger drawing effect on mixed lists was due to a combination of both costs to write items and benefits to draw items. Finally, we examined memory for order by plotting lag-CRPs and serial position curves for pure-list recall. Our analyses indicated that memory for order was task-invariant, with writing and drawing tasks producing equivalent orderings in serial recall. Collectively, these patterns suggest that drawing may not disrupt memory for order, at least as indexed by these recall measures.

The finding that drawing leads to a benefit in a pure design is consistent with a memory strength account for drawing (e.g., Bodner et al., 2014; Icht et al., 2014; MacLeod & Bodner, 2017, for application to the production effect). Drawing images of to-be-remembered items increases the memory strength of these items, which is beneficial in both mixed and pure lists. In contrast, a distinctiveness account (e.g., MacLeod et al., 2010) would suggest that draw items are distinctive relative to write items, which requires a mixed-list context containing non-distinctive items for comparison. At test, participants can then apply a heuristic to aid retrieval (e.g., retrieving that the item was drawn must indicate that the word was studied). The finding that drawing improves recognition and recall in both mixed and pure lists suggests that participants do not necessarily require the retrieval of context information for draw items, suggesting that drawing is facilitating the overall strength of the memory trace.

While evidence of a pure-list drawing effect supports a strength account, this account alone cannot explain the design effect in which the drawing effect was larger in mixed than pure lists, and why there was a mixed-list benefit for draw items. One possibility that has been applied to the production effect (e.g., MacLeod & Bodner, 2017) is that strength and distinctiveness heuristic processes are not mutually exclusive. Specifically, drawing may facilitate the strength of the memory trace, but study contexts with distinctive and non-distinctive item types may also allow for the distinctiveness heuristic to operate at test. For instance, drawing may improve the memorability of a study item, but the presence of non-distinctive items may further aid retrieval by providing participants with additional retrieval cues that are useful for discerning items that were studied. The distinctiveness heuristic may therefore operate in tandem with memory strength and provide an additional "boost" to increase the size of the drawing effect in mixed lists.

Turning to the mixed-list drawing effect in recognition and recall, this pattern was driven by a combination of both

costs and benefits. As reviewed above, mixed-list costs have been found in the production effect (Bodner et al., 2014; Hopkins & Edwards, 1972) and described as "lazy reading" of the silent-read words when placed in the context of aloud words. Borrowing from this characterization, we similarly provide evidence for "lazy writing" in which participants appear to be engaging in cursory processing of written words in the context of drawing, resulting in a memory cost. Because of this cost to writing words, the best strategy for maximizing recognition and recall may be to generate drawings for all items rather than for a subset.

Although mixed lists have consistently produced a drawing effect in our experiments and in the extant literature, pure-list effects are less consistent, occurring in some studies (Experiments 1B and 2B here; Namias et al., 2022; Wammes et al., 2016), but not in others (Jonker et al., 2019, Experiments 1–3). While our experiments were developed with the goal of resolving these discrepancies, it is important to highlight several methodological differences that may have contributed to these divergent findings. One such difference, which was explicitly noted by Jonker et al., is list length. In Wammes et al. and in our experiments, the lists were relatively long (40 and 66 items) compared to Jonker et al. (20 and 34 items). While memory performance may be lower overall on longer than shorter lists, there is little difference between the 34-item lists in Jonker et al. and our use of 40-item lists. Moreover, Namias et al., who found a reliable pure-list drawing effect, had participants study eight-item lists that consisted of related words taken from the Deese/Roediger-McDermott paradigm (DRM; Roediger & McDermott, 1995). Thus, a pure-list drawing effect may occur across a range of list lengths.

Another possibility is that the control comparisons differed. Our experiments and Wammes et al. (2016) used a write-control comparison, and Jonker et al. (2019) used a silent-reading control. Although the comparison baseline is important given the drawing effect refers to the relative difference in memory performance between drawing and the control task, there are two findings that challenge this possibility. First, Namias et al. (2022; Experiment 1) reported a pure-list drawing effect in correct recognition of DRM lists items compared to silent reading. Second, writing operates as a type of production effect and improves memory relative to silent reading (Forrin et al., 2012). If writing enhances memory, then one would expect that a writing comparison would reduce the size of the drawing effect relative to a silent-reading control. If anything, a silent-reading control task should increase the likelihood of detecting a pure-list drawing effect compared to a writing control.

A final possibility is that the drawing effect may be sensitive to the number of consecutive study/test cycles that are completed. In our experiments and in Wammes et al., participants studied and were tested on a single list, and in Namias et al. (2022), participants studied ten consecutive DRM lists but only completed a single recognition test for these lists. In contrast, Jonker et al. (2019) had participants study several lists that were individually followed by a test. These repeated study/test cycles may have produced a shift in encoding strategies over consecutive lists thereby reducing or eliminating the drawing effect. Consistent with this possibility, deWinstanley and Bjork (2004) reported that when participants completed an initial study/test cycle using a generation task, the generation effect was eliminated on a subsequent study/test cycle. Moreover, the elimination of the generation effect was not contingent upon participants showing a generation effect initially (Burnett & Bodner, 2014). Jonker et al. did not report memory differences for initial study lists versus subsequent lists, but if drawing operates similarly to generation, the potency of drawing, particularly on pure lists, may be reduced or eliminated over consecutive study/test cycles. Moreover, completion of several study/test cycles may result in the buildup of proactive interference, particularly if the list items are not discernable (e.g., all unrelated vs. taken from different semantic categories; MacLeod, 1975; Wickens et al., 1963). The lack of a pure-list drawing effect could therefore be due to repeated study/test cycles encouraging a shift in encoding strategies, the buildup of proactive interference, or some combination of the two.

In addition to examining drawing and design effect differences, we also examined memory-sequencing effects via lag-CRPs and recall serial position distance. Jonker et al. (2019) reported evidence indicating that drawing leads to a disruption in memory sequencing relative to a silent-reading control. In Experiment 3, our recall measures provided no evidence indicating that memory sequencing was disrupted for drawing over writing. Instead, these measures were more consistent with the notion that drawing of individual images encouraged item-specific encoding of information (e.g., Namias et al., 2022), but it is possible that item-specific processing may not always affect inter-item associations. Indeed, Huff and Bodner (2019; Online Supplemental Materials) computed lag-CRPs for participants who studied and then completed recall tests on DRM lists using either item-specific encoding, relational encoding, or silent-reading instructions manipulated between-subjects (Experiment 1). Although overall lag-CRPs showed strong sequencing effects for items presented in the nearest serial positions, which diminished over lags, there were no differences in sequencing across the three encoding groups, suggesting that recall sequencing measures may be less sensitive to encoding processes. Future studies will need to further evaluate organizational processing effects on recall, including whether some of the procedural differences we list above (e.g., repeated study/test cycles) are a contributing factor.

Finally, given that drawing is a relatively simple encoding task to deploy, there are several obvious applied implications, particularly for educational settings. For instance, drawing could be used to depict both concrete materials (e.g., anatomical features of the human body in an anatomy and physiology class) and more abstract concepts such as hypothetical models or plotting the similarities or differences of different concepts, which may not lend themselves easily to drawing (e.g., flow charts, Venn diagrams, etc.). While the benefits of drawing are well documented and have even been extended directly to educational settings (Fernandes et al., 2018), our study does suggest some caution regarding when drawing should be used. In particular, educational settings naturally mimic mixed-list contexts in which learners will draw some concepts at study and write or silently read others. As we show, this approach will produce benefits to draw items and costs to write items, suggesting that individuals may need to apply drawing strategically to avoid unwanted costs. For instance, generating drawings for concepts that are particularly difficult to learn and writing concepts that are already partially learned may minimize costs associated with writing while maximizing drawing benefits. The application of drawing strategically by the participant is an interesting prospect for future research given that the experimenter pre-selects which items will be drawn regardless of participants' beliefs about which items are easier or are more challenging to remember. Thus, while drawing is a quality task for enhancing memory, to promote best learning practices, a better understanding is needed regarding which contexts may help or harm memory.

Conclusion

Overall, across three experiments we found evidence that drawing images of to-be-remembered words improves immediate recognition, delayed recognition, and recall in both mixed- and pure-list contexts over writing. Drawing consistently produced a design effect in which the drawing effect was larger in a mixed list than a pure list, and this effect was due to a combination of costs and benefits. Mixed-list drawing produced a benefit relative to pure-list drawing (in delayed recognition and recall), and mixed-list writing produced a cost relative to pure-list writing (in recognition, delayed recognition, and recall). Secondary analyses found that drawing does not appear to reduce memory sequencing effects in free recall relative to writing. Drawing therefore appears to increase memory for items that are drawn in both mixed- and pure-list designs, but drawing benefits are accompanied by writing costs.

Authors' note Data collected were used for partial fulfillment of the Honor's thesis requirements for P.P. Subject level data reported are available on our Open Science Framework page (osf.io/5wx4m).

References

Balota, D.A., Yap, M.J., Hutchison, K.A. et al. The English lexicon project. *Behavior Research Methods, 39*, 445–459.

Begg, I., & Roe, H. (1988). On the inhibition of reading by generating. *Canadian Journal of Psychology, 42*, 325–336.

Begg, I., & Snider, A. (1987). The generation effect: Evidence for generalized inhibition. *Journal Experimental Psychology: Learning, Memory, and Cognition, 13*, 553–563.

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.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*, 201–210.

Bodner, G. E., Taikh, A., & Fawcett, J. M. (2014). Assessing the costs and benefits of production in recognition. *Psychonomic Bulletin & Review, 21*, 149–154.

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*, 977–990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*, 904–911.

Burnett, A. N., & Bodner, G. E. (2014). Learnin' 'bout my generation? Evaluating the effects of generation on encoding, recall, and metamemory across study-test experiences. *Journal of Memory and Language, 75*, 1–13.

Conway, M. A., & Gathercole, S. E. (1987). Modality and long-term memory. *Journal of Memory and Language, 26*, 341–361.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior, 11*, 671–684.

Crowder, R. G. (1976). *Principles of Learning and Memory*. Lawrence Erlbaum.

deWinstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & Cognition, 32*, 945–955.

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.

Engelkamp, J., & Krumnacker, H. (1980). Image- and motor-processes in the retention of verbal materials. *Zeitschrift für Experimentelle und Angewandte Psychologie, 27*, 511–533.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). GPower 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Fawcett, J. M. (2013). The production effect benefits performance in between-subjects designs: A meta-analysis. *Acta Psychologica, 142*, 1–5.

Fernandes, M. A., Wammes, J. D., & Meade, M. E. (2018). The surprisingly powerful influence of drawing on memory. *Current Directions in Psychological Science, 27*(5), 302–308.

Forrin, N. D., MacLeod, C. M., & Ozubko, J. D. (2012). Widening the boundaries of the production effect. *Memory & Cognition, 40*, 1046–1055.

Hege, A. C. G., & Dodson, C. S. (2004). Why distinctive information reduces false memories: Evidence for both impoverished relational-encoding and distinctiveness heuristic accounts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(4), 787–795.

Hopkins, R. H., & Edwards, R. E. (1972). Pronunciation effects in recognition memory. *Journal of Verbal Learning & Verbal Behavior, 11*, 534–537.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of experimental psychology. Learning, Memory, and Cognition*, *25*(4), 923-941.

Huff, M. J., & Bodner, G. E. (2019). Item-specific and relational processing both improve recall accuracy in the DRM paradigm. *Quarterly Journal of Experimental Psychology, 72*, 1493–1506.

Huff, M. J., Bodner, G. E., & Fawcett, J. M. (2015). Effects of distinctive encoding on correct and false memory: A meta-analytic review of costs and benefits and their origins in the DRM paradigm. *Psychonomic Bulletin & Review, 22*, 349–365.

Icht, M., Mama, M., & Algom, D. (2014). The production effect in memory: Multiple species of distinctiveness. *Frontiers in Psychology, 5*, 886.

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.

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.

Jonker, T. R., Wammes, J. D., & MacLeod, C. M. (2019). Drawing enhances item information but undermines sequence information in memory. *Journal of experimental psychology. Learning, Memory, and Cognition, 45*, 689–699.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*, 978–990.

MacLeod, C. M. (1975). Release from proactive interference: Insufficiency of an attention account. *The American Journal of Psychology, 88*, 459–465.

MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science, 26*, 390–395.

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.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection Theory: A User's Guide*. Cambridge University Press.

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

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.

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.

Meade, M. E., Wammes, J. D., & Fernandes, M. A. (2018). Drawing as an encoding tool: Memorial benefits in younger and older adults. *Experimental Aging Research, 44*, 369–396.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64*, 482–488.

Nairne, J. S., Thompson, S. R., & Pandeirada, J. N. S. (2007). A990. daptive memory: Survival processing enhances retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 263–273.

Namias, J. M., Huff, M. J., Smith, A., & Maxwell, N. P. (2022). Drawing individual images benefits recognition accuracy in the Deese-Roediger-McDermott paradigm. *Quarterly Journal of Experimental Psychology, 75*, 1571–1582.

Pexman, P. M., Muraki, E., Sidhu, D. M., Siakaluk, P. D., & Yap, M. J. (2019). Quantifying sensorimotor experience: Body-object interaction ratings for more than 9,000 English words. *Behavior Research Methods, 51*, 453–466.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review, 116*, 129–156.

Roberts, B. R. T., MacLeod, C. M., & Fernandes, M. A. (2022). The enactment effect: A systematic review and meta-analysis of behavioral, neuroimaging, and patient studies. *Psychological Bulletin, 148*, 397–434.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 803–814.

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.

Schonfield, D., & Robertson, B.-A. (1966). Memory storage and aging. *Canadian Journal of Experimental Psychology, 20*, 228–236.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 592–604.

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.

van Doorn, J., & ven den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T. … Wagenmakers, E.-J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Behavior Research Methods, 28*, 813–826.

Wahlheim, C. N., & Huff, M. J. (2015). Age differences in the focus of retrieval: Evidence from dual-list free recall. *Psychology and Aging, 30*, 768–780.

Wammes, J. D., Meade, M. E., & Fernandes, M. A. (2016). The drawing effect: Evidence for reliable and robust memory benefits in free recall. *Quarterly Journal of Experimental Psychology, 69*, 1752–1776.

Wammes, J. D., Meade, M. E., & Fernandes, M. A. (2017). Learning terms and definitions: Drawing and the role of elaborative encoding. *Acta Psychologica, 179*, 104–113.

Wammes, J. D., Meade, M. E., & Fernandes, M. A. (2018). Creating a recollection-based memory though drawing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*, 734–751.

Wickens, D. D., Born, D. G., & Allen, C. K. (1963). Proactive inhibition and item similarity in short-term memory. *Journal of Verbal Learning and Verbal Behavior, 2*, 440–445.

Yonelinas, A. P. (2002). The nature of recollection and familiarity. A review of 30 years of research. *Journal of Memory and Language, 46*, 441–517.

Open practices statement Data for Experiments 1, 2, and 3 are available at osf.io/5wx4m. Study materials are available upon request.