Memory & Cognition
https://doi.org/10.3758/s13421-020-01044-2

Pure-list production improves item recognition and sometimes also improves source memory

Glen E. Bodner[1] · Mark J. Huff[2] · Alexander Taikh[3]

Abstract
Relative to reading silently, reading words aloud (a type of "production") typically enhances item recognition, even when production is manipulated between groups using pure lists. We investigated whether pure-list production also enhances memory for various item details (i.e., source memory). Screen side (Experiment 1), font size (Experiment 2), or reading versus generating from anagrams (Experiments 3–4) were the sources varied within-subject, and aloud versus silent reading was varied across groups. Thus, the manipulation of source was apparent to participants, whereas the manipulation of production was not. Traditional measures and multinomial modeling established that the aloud groups generally showed improved item recognition—and showed improved source memory when steps were taken to enhance the salience of the source manipulation (Experiment 4). In summary, reading an entire list of items improves item recognition and can also improve memory for some types of source details.

Keywords Production effect · Recognition · Source memory · Distinctiveness · Strength

It would be ideal if encoding strategies enhanced memory both for whether an item was studied (item recognition) and for contextual details about the item and its encoding (source memory; Johnson, Hashtroudi, & Lindsay, 1993). In reality, however, item–context trade-offs often occur, such that processing of the identity of items comes at the expense of processing of item details and context (e.g., Mulligan, 2004). Our study tested whether one encoding strategy that has recently become of interest—and that has been found to robustly improve item recognition—also improves source memory. This encoding strategy is simply reading the items aloud during study.

Reading aloud is among the simplest encoding strategies learners can use to enhance later recognition. But until recently, it was largely overlooked by memory researchers. MacLeod, Gopie, Hourihan, Neary, and Ozubko (2010) remedied this oversight, and termed the memorial benefits of reading aloud relative to reading silently the *production effect.* Other effective forms of production include mouthing, spelling, writing, and typing (e.g., Forrin, MacLeod, & Ozubko, 2012). There is now a substantial literature showing that production reliably improves item recognition, at least within a set of boundary conditions (for a brief review, see MacLeod & Bodner, 2017). An important characteristic of the production effect on item recognition (established only after MacLeod et al.'s, 2010, initial delineation) is that it is not limited to mixed-list designs—it also occurs in pure-list designs (Bodner, Jamieson, Cormack, McDonald, & Bernstein, 2016; Bodner, Taikh, & Fawcett, 2014; Fawcett, 2013; Fawcett & Ozubko, 2016; Forrin, Groot, & MacLeod, 2016; Taikh & Bodner, 2016). This finding indicates that production does not merely reflect trade-offs between encoding of the aloud versus silent items within a list, it genuinely enhances memory for the aloud items. This important characteristic of the production effect helps to constrain potential accounts.

An early basis posited for the production effect is that produced items stand out in memory as distinctive relative to nonproduced items (Conway & Gathercole, 1987; MacLeod et al., 2010). The pure-list production effect clearly challenges relative distinctiveness as the sole basis of the production effect. Nevertheless, relative distinctiveness accounts for at least

✉ Glen E. Bodner
glen.bodner@flinders.edu.au

[1] College of Education, Psychology and Social Work, Flinders University, Adelaide, SA 5001, Australia

[2] School of Psychology, The University of Southern Mississippi, Hattiesburg, MS, USA

[3] Department of Psychology, University of Alberta, Edmonton, AB, Canada

Published online: 12 May 2020

two key findings. First, relative distinctiveness can account for why the production effect is typically larger within mixed lists than across pure lists (see Fawcett, 2013; Forrin et al., 2016; cf. Bodner et al., 2016). Second, it can account for the finding that production increases both recollection and familiarity for items in mixed lists (based on confidence ratings or remember/know judgments), but increases only the familiarity of items in pure lists (Fawcett & Ozubko, 2016). This pattern supports a *dual-process account* of the production effect (see also Ozubko et al., 2012a). The dual-process account's explanation for the pure-list effect is analogous to the *memory-strength account*, which ascribes the pure-list effect to stronger memory representations being formed for produced items at encoding (Bodner & Taikh, 2012; Bodner et al., 2014; Taikh & Bodner, 2016; cf. MacLeod et al., 2010; Ozubko & MacLeod, 2010). Because our study focuses on pure-list production, we jointly refer to these variants as the *strength account.*

Alternatively, the pure-list effect may be due to application of a global strategic decision-based process adopted during the recognition test. By a *distinctiveness-heuristic account*, participants base their recognition decisions on whether they can recollect having produced the items during encoding. This account has often been used to explain how distinctive encoding tasks reduce the Deese–Roediger–McDermott (DRM) associative-memory illusion (e.g., Dodson & Schacter, 2001; for a review, see Huff, Bodner, & Fawcett, 2015). Participants are said to apply a decision rule of the form "if I can recollect X, then the item must have been studied, otherwise it must be new", or at least they behave as though they have applied an explicit rule (see Taikh & Bodner, 2016).

To distinguish strength and distinctiveness-heuristic accounts of the pure-list production effect, Taikh and Bodner (2016) gauged whether the pure-list effect survived when attempts were made to minimize use of a distinctiveness heuristic. To this end, across study trials, the pure-aloud and pure-silent groups experienced a salient within-subjects source manipulation of either font size (small vs. large), generation (generate vs. read), or imagery (imagine referent vs. imagine uppercase). In this design, the source manipulations were salient to participants, whereas the production manipulation was not. Taikh and Bodner reasoned that the aloud groups would therefore be more likely to form a distinctiveness heuristic based on the salient source dimension than based on the nonsalient production dimension. If so, then the pure-list production effect should be eliminated. The results were mixed. Consistent with the strength account, the pure-list production effect persisted with font size and generation manipulations. But, consistent with the distinctiveness-heuristic account, the pure-list production effect was eliminated when the within-subjects factor was imagery. Thus, the pure-list production effect appears to have multiple bases.

Importantly for present purposes, although Taikh and Bodner (2016) collected old/new recognition judgments, they did not collect source judgments about item details (e.g., whether a recognized word had been studied in a small or large font). By doing so in the present study, we better tested the distinctiveness and strength accounts. By the strength account, there is no reason to expect that strengthening the representations of items through production should improve recollection of source information. It has been found that intentional binding of item and source dimensions at encoding can lead to accurate familiarity-based source judgments (e.g., Addante, Ranganath, & Yonelinas, 2012; Diana, Yonelinas, & Ranganath, 2008), but our paradigm did not involve "unitizing" the production task with the source dimensions. In contrast, if pure-list production enhances item recollection and/or reliance on a recollection-based distinctiveness heuristic, then source judgment accuracy might be higher in the aloud groups. If production facilitates access to encoding episodes, it might in turn facilitate source decisions about item details such as font size. Consistent with this possibility, studies have found evidence that episodic details about items and contexts can be bound together at encoding (Boywitt & Meiser, 2012a, 2012b; Meiser & Bröder, 2002; Starns & Hicks, 2005).

Only a few studies have examined how production affects source memory (Conway & Gathercole, 1987; Gathercole & Conway, 1988; Ozubko, Gopie, & MacLeod, 2012b; Ozubko, Major, & MacLeod, 2014)—and all of these studies used a mixed-list design with source judgments about the production factor itself (i.e., silent vs. aloud). In these studies, source accuracy was greater for aloud than for silent items. However, the effects of encoding tasks on item recognition and source memory may be confounded (see Murnane & Bayan, 1996), such that greater item recognition drives greater source accuracy. Furthermore, where memory strength is known to differ across source options, as in the case of aloud versus silent encoding, accurate source judgments might be made on the basis of item familiarity or strength rather than recollection. To address this issue, Ozubko et al. (2014, Experiment 3) used Batchelder and Riefer's (1990) multinomial-model approach to generate independent parameter estimates of item recognition, source memory, and response bias. Source memory was more accurate for aloud than silent items even when the strength of aloud and silent items was matched (via repetition of silent items). This finding is consistent with mixed-list production enhancing relative distinctiveness (e.g., MacLeod et al., 2010) and/or recollection (Fawcett & Ozubko, 2016), but it does not speak to the role of recollection in the pure-list production effect.

The effects of other encoding strategies on source memory have been studied. Mulligan (2004) compared the effects of reading versus generating items via cues (e.g., open–c____; close) on memory for extrinsic versus intrinsic item details (i.e., details that are part of the item's context vs. perceptual

details about the items themselves). Generation improved item recognition (i.e., the generation effect; Slamecka & Graf, 1978), but did not improve source memory for extrinsic item details (left vs. right screen side, background color, cue-word color). In contrast, generation *impaired* source memory for an intrinsic item detail (font color). This dissociation between item recognition and source memory held whether generation was varied within or between subjects. Mulligan interpreted this pattern using a processing account in which generation impairs perceptual encoding of items, but has no influence on the encoding of extrinsic/context information. A host of studies have since debated the robustness of this account (vs. other accounts, such as an item–context trade-off account) as a function of various encoding and retrieval conditions (McCurdy, Leach, & Leshikar, 2017; Mulligan, 2011; Mulligan, Lozito, & Rosner, 2006; Nieznański, 2011, 2012, 2014; Overman, Richard, & Stephens, 2017; Riefer, Chien, & Reimer, 2007). Importantly, unlike for generation (in which the item is not presented), production would not be expected to lead to impaired perceptual processing during encoding.

The benefit of survival-related processing of items (vs. a control task) on memory in a mixed list has also been shown to extend to source memory about the encoding task (Kroneisen & Bell, 2018; Misirlisoy, Tanyas, & Atalay, 2019). Whether survival processing enhances source memory for item details, and whether it does so in a pure-list design, have not yet been explored. As a third example, Wammes et al. (2018) found mixed evidence that the memory advantage from drawing (vs. writing) items at study in a mixed-list design extended to source memory for the encoding task.

In sum, our experiments are the first to examine whether reading an entire list of items aloud enhances memory for source details. In Experiment 1, the source dimension was screen side, an extrinsic item detail (Mulligan 2004, 2011; cf. Mather, 2007). In Experiment 2, we varied font size, an intrinsic item detail. In Experiments 3 and 4, we varied generation, an extrinsic processing-related detail. We report both measure-based analyses of item recognition and source memory and multinomial-model-based analyses. We expected to replicate the pure-list production effect on item recognition. Enhanced source memory for item details would suggest that production may sponsor use of a recollection-based distinctiveness heuristic, whereas enhanced source memory would not be expected if production merely strengthens items' representations in memory.

Experiment 1: Screen side

In Experiment 1, the within-subjects source factor was left versus right screen side at study (e.g., Mulligan, 2004). Separate groups read all of the study list words aloud or silently. At test, half made only recognition judgments (as in Taikh & Bodner, 2016), and half made recognition and source judgments indicating which screen side recognized words had appeared on at study. Thus, we used a 2 (production: silent vs. aloud) × 2 (test type: no-source vs. source) × 2 (source: left vs. right) design; production and test type were varied between groups, and screen side was varied within each group. Test type was included in Experiment 1 to test whether making source judgments influences the pure-list production effect on recognition. For example, production might lead the aloud group to focus on trying to recollect screen side rather than production at test, which might reduce the pure-list production effect (see Taikh & Bodner, 2016). On the other hand, if the pure-list production effect simply increases item strength (Fawcett & Ozubko, 2016), then the effect should be similar in the source and no-source groups.

Method

Participants University of Calgary undergraduates participated for course credit in one of Experiments 1–4a. Participants in Experiment 1 were randomly assigned to the no-source/silent, no-source/aloud, source/silent, or source/aloud group (32 per group). A sensitivity analysis using G*POWER (Erdfelder, Faul, & Buchner, 1996), indicated that this sample size had sufficient power (.80) to detect medium-sized or larger effects (Cohen's $d > .48$) for main effects and interactions.

Stimuli and design The stimuli were 100 nouns, 5–8 letters in length, used previously (e.g., Taikh & Bodner, 2016), which were split into four sets. At study, one set was shown on the left and another on the right, randomly mixed. At test, these 50 studied words were randomly mixed, with the remaining 50 words serving as lures. Set assignment was counterbalanced.

Procedure The procedure followed Taikh and Bodner (2016), except (1) the source manipulation was screen side (after Mulligan, 2004), and (2) the source groups made source judgments. Participants were tested individually in lab via computer with the experimenter present. They were asked to study a list of words for an unspecified memory test. They were not asked to encode screen side. Each study list word was presented for 2 s, in black 36-point Arial font, on either the left or right side of the screen, followed by a .5-s blank screen. The aloud group read each word aloud, and the silent group read each word silently without mouthing/whispering. The test phase immediately followed. Words were presented individually in the center of the screen in black 36-point Arial font, along with the cues OLD and NEW. Via key press, participants indicated whether the word was old or new. In the source groups, when a word was deemed old, the cues LEFT and RIGHT then appeared on the screen and participants indicated via key press whether the word had appeared on the left or right side of the screen at study (guessing if necessary).

Results

Measure-based analyses of item recognition and source memory For significant comparisons, partial eta-squared (η_p^2) is reported as a measure of effect size for analyses of variance (ANOVAs) and Cohen's *d* for *t* tests. All nonsignificant comparisons were further tested using a Bayesian estimate of the strength of evidence supporting the null hypothesis (Masson, 2011; Wagenmakers, 2007). This analysis compares a model that assumes an effect with a model that assumes a null effect. The analysis produces a Bayesian estimate termed p_{BIC} (Bayesian information criterion), which is the probability that a null effect is retained. The p_{BIC} estimate is sensitive to sample size, thereby increasing confidence in null effects reported. Our p_{BIC} analysis therefore supplements null effects found using traditional null-hypothesis-significance testing.

We used *d′*, a signal-detection measure of discrimination, for our measure-based analysis of whether production enhanced item recognition. Floor false alarms and ceiling hit rates were adjusted using a 1/2*N* correction (Macmillan & Creelman, 1991). Table 1 provides the mean hit and false-alarm rates and *d′*s for each experiment. Mean *d′* was analyzed using a 2 (production: silent vs. aloud) × 2 (test type: no-source vs. source) × 2 (source: left vs. right) mixed-factor ANOVA. As expected, there was a pure-list production effect, reflecting better discrimination in the aloud (vs. silent) groups (2.22 vs. 1.89), $F(1, 124) = 6.58$, $MSE = 1.08$, $p = .01$, $\eta_p^2 = .05$. The test type effect reflected poorer discrimination in the source (vs. no-source) groups (1.92 vs. 2.19), $F(1, 124) = 4.30$, $MSE = 1.08$, $p = .04$, $\eta_p^2 = .03$. Critically, the pure-list production effect was of similar magnitude in the no-source and source groups ($F < 1$, $p_{BIC} = .88$). Thus, the addition of source judgments did not influence the pure-list production effect. Discrimination was not significantly different for words presented on the left versus right side of the screen (2.09 vs. 2.02), $F(1, 124) = 3.21$, $MSE = 0.12$, $p = .08$, $\eta_p^2 = .03$. The other interactions were not significant (Fs < 1.79; ps > .19; p_{BIC}s > .82).

Figure 1 provides the mean proportion of each source judgment for each item type. For our measure-based analysis of whether production enhances source memory, following Mulligan (2004), an identification-of-origin (IO) score was calculated for each participant in the source groups by dividing the number of correct source judgments for studied words by the number of correctly recognized words. Table 2 provides the mean source accuracy for aloud and silent source groups for each experiment. Accuracy was above chance (.5)

Table 1. Experiments 1–4b: Mean (±95% CI) proportions of hits, false alarms, and discrimination (*d′*)

Experiment/group/source	Hits		False alarms		Discrimination (*d′*)	
	Silent	Aloud	Silent	Aloud	Silent	Aloud
Exp. 1: Screen side						
No-source group						
Left	.76 (.06)	.77 (.06)	.14 (.04)	.09 (.02)	2.04 (.29)	2.41 (.25)
Right	.71 (.06)	.78 (.06)	.14 (.04)	.09 (.02)	1.88 (.31)	2.43 (.27)
Source group						
Left	.68 (.08)	.72 (.04)	.13 (.04)	.10 (04)	1.87 (.27)	2.06 (.22)
Right	.64 (.08)	.71 (.04)	.13 (.04)	.10 (04)	1.76 (.29)	2.00 (.18)
Exp. 2: Font size						
Small	.64 (.06)	.71 (.06)	.12 (.02)	.11 (.04)	1.73 (.24)	2.06 (.25)
Large	.69 (.06)	.70 (.06)	.12 (.02)	.11 (.04)	1.86 (.23)	2.12 (.29)
Exp. 3: Generation						
Read	.66 (.08)	.71 (.06)	.19 (.06)	.15 (.04)	1.51 (.27)	1.78 (.24)
Generate	.87 (.04)	.92 (.02)	.19 (.06)	.15 (.04)	2.27 (.24)	2.70 (.25)
Exp. 4a: Enhanced generation						
Read	.56 (.06)	.61 (.06)	.17 (.04)	.13 (.02)	1.20 (.20)	1.55 (.25)
Generate	.88 (.04)	.94 (.02)	.17 (.04)	.13 (.02)	2.42 (.25)	2.95 (.25)
Exp. 4b: Enhanced generation (replication)						
Read	.60 (.06)	.66 (.06)	.23 (.05)	.15 (.04)	1.15 (.22)	1.63 (.23)
Generate	.86 (.04)	.93 (.03)	.23 (.05)	.15 (.04)	2.07 (.26)	2.77 (.25)

Note. The false-alarm rate for each group is repeated for both levels of the within-subjects factor to indicate that new items cannot be ascribed to these levels given that they were not studied

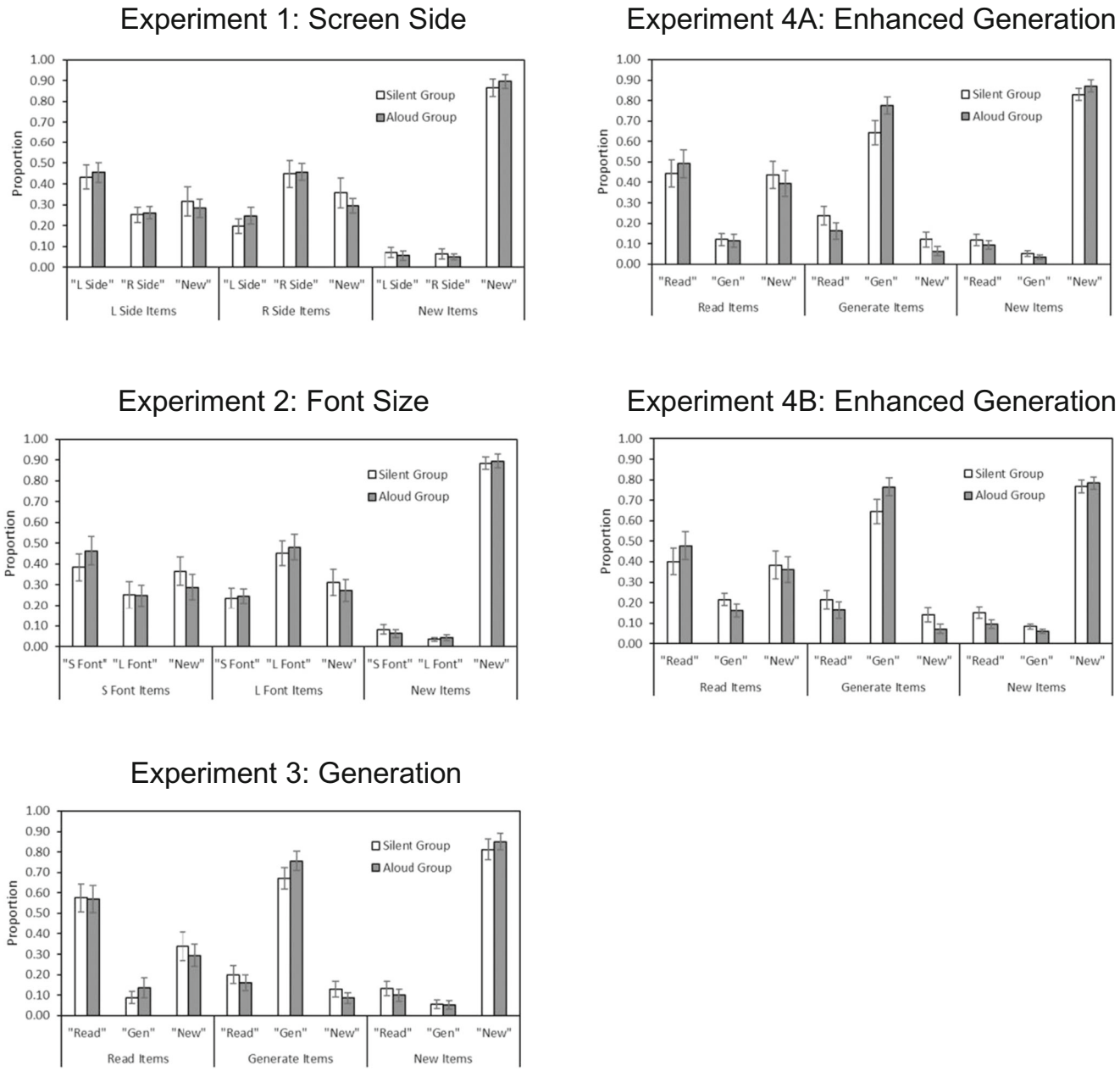

Fig. 1 Mean proportion of source judgments (and 95% CI) by experiment, group, and item type (L/R = left/right; S/L = small/large; Gen = generate)

Table 2. Experiments 1–4b: Mean (± 95% CI) identification-of-origin (IO) proportion correct

	Group	
Experiment	Silent	Aloud
Exp. 1: Screen Side	.66 (.04)	.64 (.04)
Exp. 2: Font Size	.63 (.04)	.65 (.04)
Exp. 3: Generation	.81 (.04)	.82 (.04)
Exp. 4a: Enhanced generation	.75 (.04)	.82 (.04)
Exp. 4b: Enhanced generation (replication)	.72 (.05)	.78 (.04)

for both the aloud group, $t(31) = 7.10$, $SEM = .02$, $p < .001$, $d = 1.26$, and the silent group, $t(31) = 7.69$, $SEM = .02$, $p < .001$, $d = 1.36$, indicating reliable source memory for screen side. However, memory for screen side was equivalent in the aloud and silent groups (.64 vs. .66), thus production did not improve memory for this item detail ($t < 1$, $p_{BIC} = .87$).

Multinomial-model-based analyses of item recognition and source memory Following studies of generation (e.g., Mulligan, 2004; Nieznański, 2011) and survival processing (e.g., Kroneisen & Bell, 2018; Misirlisoy et al., 2019), we also conducted multinomial-model-based analyses (Batchelder & Riefer, 1990; Dodson, Prinzmetal, & Shimamura, 1998). This

approach allowed estimation of separate parameters associated with item recognition and source memory, while isolating potential response biases in recognition and source judgments. This approach overcomes a known issue with the IO and other typical source measures, namely their confounding of item recognition and source memory (see Murnane & Bayan, 1996).

Parameter processing trees used in the multinomial-model analyses followed Dodson et al. (1998) and are shown in Table 3. During the test phase, participants shown a studied word either recognize the word (parameters D_1 and D_2) or not ($1 - D_1$ or D_2). If they recognize it, they then either identify its source (parameters d_1 and d_2) or not ($1 - d_1$ or d_2). If they fail to identify its source, they will have a certain bias to guess Source 1 (parameter a) over Source 2 ($1 - a$). Whenever they fail to recognize the word (or if the word is new), they will have a certain bias to guess that it is old (parameter b) over new ($1 - b$). If they guess that it is old, then will have a certain bias to guess Source 1 (parameter g) over Source 2 ($1 - g$). Otherwise, the item is deemed new. As outlined in Dodson et al. (1998), this model allows source-judgment biases to differ when recognition succeeds (parameter a) versus when recognition fails but the person guesses that the word is old (parameter g). In total, the aloud and silent groups each had seven parameters: two for item recognition for items from each source (D_1 and D_2), two for source memory for items from each source (d_1 and d_2), and three for the response biases described above (a, g, b).

Table 3. Multinomial model of source judgments

Item type	Parameter trees	Source judgment response
Source 1	$D_1 \bullet d_1$	"Source 1"
	$D_1 \bullet (1- d_1) \bullet g$	"Source 1"
	$D_1 \bullet (1 - d_1) \bullet (1 - g)$	"Source 2"
	$(1 - D_1) \bullet b \bullet g$	"Source 1"
	$(1 - D_1) \bullet b \bullet (1 - g)$	"Source 2"
	$(1 - D_1) \bullet (1 - b)$	"New"
Source 2	$D_2 \bullet d_2$	"Source 2"
	$D_2 \bullet (1 - d_2) \bullet g$	"Source 2"
	$D_2 \bullet (1 - d_2) \bullet (1 - g)$	"Source 1"
	$(1 - D_2) \bullet b \bullet g$	"Source 2"
	$(1 - D_2) \bullet b \bullet (1 - g)$	"Source 1"
	$(1 - D_2) \bullet (1 - b)$	"New"
New	$b \bullet g$	"Source 1"
	$b \bullet (1 - g)$	"Source 2"
	$1 - b$	"New"

Note. The model structure was identical for the aloud and silent groups. D_1/D_2 = probability of recognition of items from Source 1 and Source 2; d_1/d_2 = probability of source memory for items from Source 1 and Source 2; a = probability of guessing Source 1 when item recognition occurs; g = probability of guessing Source 1 when item recognition fails; b = probability of responding old when item recognition fails. See text, Dodson et al. (1998), and Batchelder and Riefer (1990) for details

The fit of this model to the response frequency data for the source judgment responses is shown in the right column of Table 3. As detailed in Dodson et al. (1998), fits were determined using the Solver function in Microsoft Excel and assessed using G^2 (the log-likelihood ratio statistic) as a goodness-of-fit measure. Under the assumption that item recognition was similar for the left and right screen side items, we constrained $D_1 = D_2$ in each group (i.e., aloud, silent). We also assumed that source memory would be similar for left and right items, so we constrained $d_1 = d_2$. We also assumed that the bias toward Source 1 (here, left side; see Table 1) would be similar whether the item was recognized or not, so we constrained $a = g$ in each group. This model, with eight free parameters, fit the data, $G^2(8) = 7.79 <$ chi-squared critical value 11.07.

We next tested two restricted versions of this model, which we dub the *base model*, as outlined in Table 4. First, we tested whether the model fit under the additional assumption that there was no production effect on item recognition (NoPE-R model column in Table 4), by constraining item recognition parameters in the aloud group to equal those in the silent group. The NoPE-R model, with seven free parameters, failed to fit the data, $G^2(7) = 19.40 >$ critical value 12.49. Thus, consistent with the d' measure-based analyses, the model-based analyses indicated a production effect on item recognition. Second, we tested whether the model fit under the assumption that there was no production effect on source memory (NoPE-S model column in Table 4), by constraining the source memory parameters in the aloud group to equal those in the silent group. The NoPE-S model, with seven free parameters, fit the data, $G^2(7) = 9.43 <$ critical value 12.59, and fit as well as the base model. Thus, consistent with the IO measure, the model-based approach indicated that there was a pure-list production effect on item recognition, but not on source memory.

Discussion

Both measure-based and model-based analyses indicated that pure-list production enhanced item recognition (e.g., Taikh & Bodner, 2016), but did not enhance source memory for screen side. This pattern was also obtained for location judgments by Mulligan (2004, Experiment 4) when generation was manipulated between groups. Requiring a source judgment reduced recognition discrimination (cf. Mulligan, Besken, & Peterson, 2010), but, importantly, the production effect on item recognition was similar whether or not source judgments were made.

The null effect of pure-list production on source memory fits with Fawcett and Ozubko's (2016) claim that pure-list

Table 4. Experiments 1–4b: Parameter estimates, *df*, and fits for models assuming no production effect no-recognition (NoPE-R) or no-production effect on source memory (NoPE-S)

Group/ parameter/ *df* − free/ Fit	Exp. 1: Screen side		Exp. 2: Font size		Exp. 3: Generation		Exp. 4a: Enhanced gen.		Exp. 4b: Enhanced gen. (replication)	
	NoPE-R model	NoPE-S model	NoPE-R model	NoPE-S model	NoPE-R model	NoPE-S model	NoPE-R model	NoPE-S model	NoPE-R model	NoPe-S model
Silent group										
D_1	.65[a]	.61[a]	.65[a]	.62[a]	.62[a]	.59	.51[a]	.47	.53[a]	.46
D_2	.65[a]	.61[a]	.65[a]	.62[a]	.87[b]	.84	.89[b]	.85	.87[b]	.81
d_1	.35[b]	.32[b]	.28[b]	.31[b]	.67[c]	.66[a]	.54[c]	.61[a]	.45[c]	.53[a]
d_2	.35[b]	.32[b]	.28[b]	.31[b]	.67[c]	.66[a]	.54[c]	.61[a]	.45[c]	.53[a]
a	.48[c]	.48[c]	.44	.44	.65	.65	.57	.60	.43	.44
g	.48[c]	.48[c]	.71	.71	.71	.71	.69	.68	.65	.64
b	.13	.13	.11	.12	.18	.19	.16	.17	.22	.24
Aloud group										
D_1	.65[a]	.68[d]	.65[a]	.69[c]	.62[a]	.65	.51[a]	.55	.53[a]	.59
D_2	.65[a]	.68[d]	.65[a]	.69c	.87[b]	.90	.89[b]	.93	.87[b]	.92
d_1	.30[d]	.32[b]	.33[c]	.31[b]	.66[d]	.66[a]	.66[c]	.61[a]	.58[c]	.53[a]
d_2	.30[d]	.32[b]	.33[c]	.31[b]	.66[d]	.66[a]	.66[c]	.61[a]	.58[c]	.53[a]
a	.50[e]	.50[e]	.48	.49	.48	.48	.48	.46	.40	.39
g	.50[e]	.50[e]	.60	.60	.66	.66	.73	.74	.61	.61
b	.11	.10	.11	.10	.16	.15	.13	.13	.16	.15
df − free	6	6	4	4	3	2	3	2	3	2
G^2	19.40*	9.43	18.75*	6.61	15.89*	0.56	23.98*	7.49*	40.80*	7.83*

Note. Parameters constrained to be equal in a given model share a superscript. D_1/D_2 = item recognition parameters for items from Source 1 and Source 2; d_1/d_2 = source memory parameters for items from Source 1 and Source 2; a = guessing bias toward Source 1 when item recognition occurs; g = guessing bias toward Source 1 when item recognition fails; b = bias toward responding old when item recognition fails; *df* − free = model *df* − number of free parameters in model = *df* for critical value. See text, Dodson et al. (1998), and Batchelder and Riefer (1990) for details. * = G^2 > critical value, indicating that the model did not fit the data

production increases item familiarity, but not recollection. If recollection of encoding episodes had been enhanced in the aloud group, and if those episodes had bound together item and context information, then this group might have better recollected screen side (e.g., Boywitt and Meiser, 2012a, b; Meiser & Bröder, 2002). This was not the case.

Experiment 2: Font size

Experiments 2–4 provided additional tests of the influence of pure-list production on item recognition and source memory. In Experiment 2, at study, half the words were presented in a small font and half in a large font (after Taikh & Bodner, 2016, Experiment 1). Font size can be considered an intrinsic item detail. Mulligan (2004, Experiment 3) found that pure-list generation impaired memory for a similar intrinsic item detail, font color, consistent with generation favoring item encoding over context encoding. If reading aloud also favors item encoding over context encoding, then memory for font size might be worse in the aloud group. On the other hand, if the aloud group adopts a recollection-based distinctiveness heuristic at test, it might show improved memory for font size. Finally, if the pure-list production effect is driven solely by increased item strength, then source accuracy should be equivalent in the two groups.

Method

Participants were randomly assigned to the aloud or silent group (32 per group). The method followed the source group in Experiment 1, except as follows. At study, words appeared in the center of the screen either in small font (25-point Times New Roman) or large font (125-point Times New Roman). At test, words appeared in 75-point Georgia font. For the source judgment task, the cues SMALL and LARGE prompted participants to indicate which font size the word had appeared in at study.

Results

Measure-based analyses of item recognition and source memory Taikh and Bodner (2016, Experiment 1) found significantly higher *d′* in the aloud group than in the silent group (2.20 vs. 1.78); here, the pure-list production effect was of similar magnitude, but was marginal (2.09 vs. 1.79; Table 1), $F(1, 62) = 2.99$, $MSE = .92$, $p = .09$, $\eta_p^2 = .05$, $p_{BIC} = .64$. Discrimination was not significantly greater for large (vs. small) font items (1.99 vs. 1.89), $F(1, 62) = 3.41$, $MSE = .09$, $p = .07$, $\eta_p^2 = .05$, $p_{BIC} = .59$, consistent with other findings (Castel & Rhodes, 2008; McDonough & Gallo, 2012; Mueller, Dunlosky, Tauber, & Rhodes, 2014). The interaction was not significant ($F < 1$, $p_{BIC} = .87$).

For source memory (see Table 2), font size judgments were above chance for both the aloud group, $t(31) = 6.34$, $SEM = .01$, $d = 1.12$, and the silent group, $t(31) = 5.43$, $SEM = .02$, $p < .001$, $d = 0.96$, but was again equivalent in the aloud and silent groups (.65 vs. .63; $t < 1$, $p_{BIC} = .88$).

Multinomial-model-based analyses of item recognition and source memory Following Experiment 1, we first established a suitable base model. Given that item recognition was not significantly different for small and large fonts, we again constrained $D_1 = D_2$ in each group. We also assumed that source memory would be similar for small and large font, so we constrained $d_1 = d_2$. We initially assumed that the bias toward Source 1 (here, small font; see Table 1) would be similar whether the item was recognized or not, by constraining $a = g$ in each group. However, the model did not fit the data with this assumption in place. Instead, as shown in Table 4, the bias to guess that the word was in small font was greater when recognition failed than when recognition occurred (i.e., $g > a$). Therefore, this constraint was removed. The resulting base model, with 10 free parameters, fit the data, $G^2(10) = 5.37 <$ critical value 7.82. We then tested the two nested versions of this base model outlined in Table 4. The NoPE-R model failed to fit the data, $G^2(9) = 18.75 >$ critical value 9.49. Thus, although the production effect on *d′* was only marginal, a model assuming no production effect on item recognition did not fit the data. The NoPE-S model, on the other hand, did fit the data, $G^2(9) = 6.61$, and did so as well as the base model. Thus, consistent with the IO measure, this model indicated that production did not enhance memory for font size.

Discussion

Pure-list production did not enhance source memory for an intrinsic item detail—font size. Using font color as the intrinsic item detail, Mulligan (2004, Experiment 3) found that pure-list generation enhanced item recognition, but impaired source memory. Production did not have the same negative effect on source memory for an intrinsic item detail as generation, likely because items are presented equivalently (and thus receive equivalent perceptual processing) in both silent and aloud conditions. Experiments 3 and 4 provide additional tests of whether production and generation are dissociable in this respect. So far, our experiments have not yielded evidence that reading aloud bolsters recollection of item details—whereas the lack of improvement is consistent with pure-list production selectively strengthening item memory.

Experiment 3: Generation

In Experiment 3, half the items at study were read intact and half were generated via anagrams (after Taikh & Bodner, 2016, Experiment 3). Generation then served as the source dimension at test. Unlike manipulations of intrinsic item details such as screen side, font size, or font color, generation is well known to enhance memory relative to reading (for a review, see Bertsch, Pesta, Wiscott, & McDaniel, 2007). Taikh and Bodner found that pure-list production enhanced item recognition similarly for read and generated items; additivity of production and generation effects has also been shown in mixed-list designs (Forrin, Jonker, & MacLeod, 2014; MacLeod et al., 2010). Relative to screen side and font size, generation was expected to be more recollectable, and thus more likely to sponsor higher source accuracy. To the extent that the aloud group in Experiment 3 experience and/or rely more on recollection, this group might show greater source memory. On the other hand, if pure-list production enhances only item strength, then source memory should continue to be similar across groups. The latter prediction follows given the noted additivity of production and generation. Therefore, production was expected to strengthen generate and read items equally, and was not expected to improve source accuracy.

Method

Participants were randomly assigned to the aloud or silent group (32 per group). The same method was used, except generation was varied at study (as in Taikh & Bodner, 2016, Experiment 3). Words were presented in 75-point Times New Roman font at study and test. One set of words was presented as anagrams in which the first and third letters were always switched (generate items), and another set was presented intact (read items). At study, participants were asked to read each word and anagram solution (after switching the first and the third letters) either aloud or silently. Participants could request hints, but rarely did so. Following others (Forrin et al., 2014; Taikh & Bodner, 2016), we assumed that the silent group performed the generation task and thus that their generation effect would be comparable to that of the aloud group.

To accommodate the anagram task, participants pressed the space bar to advance to the next study item. At test, the cues ANAGRAM and WORD prompted the source judgment.

Results

Measure-based analyses of item recognition and source memory The pure-list production effect on item recognition d' was significant (2.24 vs. 1.89), $F(1, 62) = 4.48$, $MSE = .91$, $p = .04$, $\eta_p^2 = .07$. We also obtained a robust generation effect on d' (2.49 vs. 1.64), $F(1, 62) = 125.46$, $MSE = .18$, $p < .001$, $\eta_p^2 = .67$. The production effect was similar for read and generate items, $F(1, 62) = 1.15$, $MSE = .18$, $p = .29$, $p_{BIC} = .82$, for the interaction (see Table 1).

As expected, source accuracy was higher here than in Experiments 1 and 2 (see Table 2)—and well above chance for both the aloud group, $t(31) = 18.82$, $SEM = .02$, $p < .001$, $d = 3.55$, and the silent group, $t(31) = 20.08$, $SEM = .02$, $p < .001$, $d = 3.33$. But source accuracy was again very similar for the aloud and silent groups (.82 vs. .81; $t < 1$, $p_{BIC} = .89$). Thus, for the third time, pure-list production did not enhance source memory.

Multinomial-model-based analyses of item recognition and source memory The base model was similar to Experiment 2, except that, given the generation effect on item recognition, the item recognition constraint $D_1 = D_2$ was removed in each group to fit the data, $G^2(12) \approx 0.00 <$ critical value 3.84 (see Table 4). To test the NoPE-R model, we modified the base model by setting D_1 in the silent group equal to D_1 in the aloud group, and D_2 in the silent group equal to D_2 in the aloud group. This model, which assumes no pure-list production effect on item recognition, failed to fit the data, $G^2(10) = 15.89 >$ critical value 7.82. To test the fit of the NoPE-S model, the four source parameters were constrained to be equal. This model fit the data, $G^2(11) = 0.56 <$ critical value 5.99, and fit it as well as the base model, indicating no pure-list production effect on source memory.

Discussion

The production effect in item recognition persisted in the presence of a very effective generation manipulation; this additivity of production and generation replicates previous findings (Forrin et al., 2014; Taikh & Bodner, 2016). Yet pure-list production again failed to enhance source memory. If the aloud group had adopted a recollection-based distinctiveness heuristic, it did not render them more likely to recollect whether items that they read aloud had been generated from an anagram. Instead, so far, our results fit better with the claim that production merely increases the familiarity of items, which would not facilitate source judgments.

Experiment 4a: Enhanced generation

Using a within-subject manipulation of imagery, Taikh and Bodner (2016) did not obtain a pure-list production effect on recognition. They suggested that the aloud group based their recognition decisions on whether they could recollect the highly memorable imagery task, thus bypassing use of a production-based distinctiveness heuristic or an evaluation of memory strength. Experiment 4a posed two questions. First, would an enhanced version of the generation task, like an imagery task, also eliminate the pure-list production effect? The generation task was augmented to provide participants with more "recollective fodder" through two means: (1) varying the anagram solution rule at study, and (2) presenting all test items as anagrams to remind participants of their previous encounter with generate items (see MacLeod, Pottruff, Forrin, & Masson, 2012). Second, would pure-list production improve source memory for this "enhanced" version of the generation task?

Method

Participants were randomly assigned to either the aloud or silent group (32 per group). The procedure followed Experiment 3, except as follows. At study, half of the anagrams were presented with the second and fourth letters switched—to make the generate task more varied relative to Experiment 3. Participants could request hints, but rarely did so. Second, all test items were presented as anagrams—to increase the salience of the generation manipulation relative to Experiment 3. At test, generate items were presented with the same letter rearrangement that was used at study. Half of the read items and lure items were presented with each type of letter rearrangement at test, determined randomly. Participants were not required to report the anagram solutions before making their recognition decisions. We assumed that participants generally solved the anagrams at test, but recognition decisions for some generate items may have been based on recognition of the anagrams themselves.

Results

Measure-based analyses of item recognition and source memory Enhanced generation resulted in numerically higher d' scores for generate items, but lower d' scores for read items, relative to Experiment 3. The latter result might reflect the mismatch in presentation of the read items between encoding (items shown as intact words) and retrieval (items shown as anagrams). Alternatively, the focus on item-specific processing for solving the anagrams may have spilled over to the read items, resulting in impaired relational processing (see Huff et al., 2015; Huff, Bodner, & Gretz, 2019). The pure-list production effect on item recognition remained with enhanced

generation (2.25 vs. 1.81), $F(1, 62) = 8.24$, $MSE = .75$, $p < .01$, $\eta_p^2 = .12$. The generation effect was again robust (2.68 vs. 1.38), $F(1, 62) = 298.21$, $MSE = .18$, $p < .001$, $\eta_p^2 = .83$, and the pure-list production effect was again similar for read and generate items, $F(1, 62) = 1.37$, $MSE = .18$, $p = .25$, $p_{BIC} = .80$, for the interaction.

As in Experiment 3, source accuracy was well above chance for both the aloud group, $t(31) = 18.82$, $SEM = .02$, $d = 3.38$, and the silent group $t(31) = 11.96$, $SEM = .02$, $p < .001$, $d = 2.11$. Critically, unlike in Experiments 1–3, source accuracy was reliably higher in the aloud group than in the silent group (.82 vs. .75), $t(61) = 2.35$, $SEM = .03$, $p = .02$, $d = 0.60$, indicating a pure-list production effect on source memory.

Multinomial-model-based analyses of item recognition and source memory The Experiment 3 base model fit the data, $G^2(12) \approx 0.00 <$ critical value 3.84 (see Table 4). The NoPE-R model failed to fit the data, $G^2(10) = 23.98 >$ critical value 7.82, in line with the pure-list production effect on item recognition. Unlike in Experiments 1–3, the NoPE-S model also failed to fit the data, $G^2(11) = 7.49 >$ critical value 5.99, consistent with the pure-list production effect on IO scores. In the base model, the source memory parameters (i.e., for $d_1 = d_2$) were higher in the aloud group than in the silent group (.66 vs. .55), consistent with the pure-list production effect on the *IO* scores; otherwise the base model parameters were very similar to the Experiment 3 models in Table 4.

Discussion

Experiments 1–3 suggested that reading an entire list of words aloud increases their strength, rather than leading to greater use of a recollection-based strategy. Experiment 4a invites more nuance, given that the enhanced generation manipulation yielded a pure-list production effect both on item recognition (cf. the imagery manipulation reported in Taikh & Bodner, 2016) and source memory. When the two sources differ sufficiently, it appears that production increases the ability to recollect distinctive source features, yielding a pure-list production effect on source memory. Before further considering this possibility, we report an exact replication of Experiment 4a.

Experiment 4b: Enhanced generation (replication)

Method

University of Southern Mississippi undergraduates participated for course credit and were randomly assigned to either the aloud or silent group (32 per group). The method was identical to Experiment 4a.

Results

Measure-based analyses of item recognition and source memory Pure-list production improved item recognition (2.20 vs. 1.61), $F(1, 62) = 14.27$, $MSE = .78$, $p < .001$, $\eta_p^2 = .19$. The within-list generation effect was robust (2.42 vs. 1.30), $F(1, 62) = 203.04$, $MSE = .17$, $p < .001$, $\eta_p^2 = .77$, and the pure-list production effect was similar for read and generate items, $F(1, 62) = 2.38$, $MSE = .17$, $p = .13$, $p_{BIC} = .71$, for the interaction. Source IO scores were above chance for both aloud, $t(31) = 14.83$, $SEM = .02$, $p < .001$, $d = 3.71$, and silent groups, $t(31) = 9.01$, $SEM = .02$, $p < .001$, $d = 2.25$, and critically, were greater in the aloud versus silent group (.78 vs. .72), $t(61) = 2.04$, $SEM = .03$, $p < .05$, $d = 0.52$.

Multinomial-model-based analyses of item recognition and source memory The Experiments 3–4a base model fit the data, $G^2(12) \approx 0.00 <$ critical value 3.84 (see Table 4). The NoPE-R model failed to fit the data, $G^2(10) = 40.80 >$ critical value 7.82, in line with the pure-list production effect on item recognition. The NoPE-S model also failed to fit the data, $G^2(11) = 7.83 >$ critical value 5.99, consistent with the pure-list production effect on source memory. Base model source parameters (i.e., for $d_1 = d_2$) were again greater in the aloud group than in the silent group (.57 vs. .46), consistent with the IO scores.

Discussion

Experiment 4b fully replicated Experiment 4a using the same method but a different participant sample. Critically, the aloud group again showed more accurate source memory for the enhanced generation factor. Thus, it appears that production can enhance source memory for item details if these details are sufficiently recollectable at test.

General discussion

Building on recent interest in the memorial consequences of reading aloud (MacLeod & Bodner, 2017), we examined whether reading a list of words aloud enhances source memory for item details. By a strength account of the pure-list production effect, reading aloud should make all test items more familiar, but this should not facilitate making more accurate source judgments. In contrast, if reading an entire list of items aloud leads participants to adopt a recollection-based distinctiveness heuristic at test, then enhanced recollection of the encoding episode may help bring other aspects of the encoding experience to mind—thereby facilitating source memory (e.g., Boywitt and Meiser, 2012a, b; Meiser & Bröder, 2002). Pure-list production typically enhanced item recognition, replicating prior findings (Bodner et al., 2014;

Bodner et al., 2016; Fawcett, 2013; Fawcett & Ozubko, 2016; Forrin et al., 2016; Taikh & Bodner, 2016). Moreover, consistent with the strength account, pure-list production did not improve source memory in Experiments 1–3. However, consistent with the distinctiveness account, the aloud group showed higher source accuracy for our enhanced generation manipulation in Experiment 4a (replicated in 4b).

Experiments 1–3 suggest that reading aloud does not generally enhance the ability to recollect aspects of the encoding event such as the location of the items or their size, or whether they had been read intact or generated from anagrams (under some circumstances; cf. Experiments 4a–4b). Whether this is true of other encoding strategies, such as survival-related processing and drawing, remains an important question—to date, studies have only examined source judgments about the encoding tasks themselves (Kroneisen & Bell, 2018; Misirlisoy et al., 2019; Wammes et al., 2018). The benefits of drawing items, which may be considered a form of production, have been attributed to the "seamless integration of elaborative, motoric, and pictorial components of a memory trace" (Fernandes, Wammes, & Meade, 2018, p. 304). By this integrated-trace hypothesis, the act of drawing binds together the meaning of the item, what it looks like, and how the drawing was rendered by hand. Whether memory for item details is bound with these component processes at encoding remains to be tested.

Generation has variously been found to lead to better or worse memory for item details, or to have no effect, depending on aspects of the task and materials (Mulligan, 2004, 2011; Mulligan et al., 2006; Nieznański, 2011, 2012; Riefer et al., 2007). In contrast, our experiments provided no indication that production disrupts memory for item details. Production does not appear to induce a trade-off between encoding of the items and encoding of their details or contextual elements, perhaps because—unlike for generation—items are presented identically whether they are read aloud or silently. Thus, although production and generation have similarities, our findings echo others in suggesting that the two encoding strategies differ (e.g., Forrin, Jonker, & MacLeod, 2014; MacLeod et al., 2010).

Experiments 4a and 4b add an important caveat. Here, with an "enhanced generation" manipulation, the aloud groups showed improved source memory for whether words had been read or generated at study. This condition involved presenting a greater variety of anagram types at encoding and then representing the anagrams at test. We suggest that these changes facilitated recollection of the encoding episodes in the aloud groups, enabling them to recollect whether they had read the item or solved an anagram of it at study (Boywitt & Meiser, 2012a, 2012b; Meiser & Bröder, 2002; Starns & Hicks, 2005). Recently, Hourihan and Churchill (2020) reported that production of labels for pictures of objects enhanced both item recognition and memory for item details. Specifically, participants could better discriminate which exemplars of the same object type (e.g., a potted cactus) they had studied if they had said the object name aloud rather than silently at study. Whether a pure-list manipulation of production would enhance item and source memory for pictures remains to be tested. Taken together, there are indications that production may help bind together aspects of one's encoding experiences, in line with the integrated-trace account of the drawing effect (Fernandes et al., 2018). Certainly, this possibility is worthy of further attention.

The improved item recognition in the aloud groups in our experiments was always similar for both levels of our within-subjects factors (see also Forrin et al., 2014; MacLeod et al., 2010). This additivity is important. If the benefits of reading aloud on recognition had interacted with the within-subject factors, then participants would potentially have been able to base their source judgments on strength differences between the two levels. For example, had production strengthened memory more for generate items than for read items, then accurate source judgments could have been made by classifying stronger memories as "anagram" and weaker memories as "read" (see Ozubko et al., 2014). The observed additivity removes this potential concern.

Previous studies of the influence of production on source memory have shown that people are more likely to remember the source or "input modality" for words that were read aloud than for words that were read silently (Conway & Gathercole, 1987; Gathercole & Conway, 1988; Ozubko et al., 2012a; Ozubko et al., 2014), suggesting that production increases the distinctiveness of items in memory. However, in these studies, the production effect on item recognition may have boosted the production effect on modality judgments, due to the confounding of item and source memory in basic IO type source measures (Murnane & Bayan, 1996). Using a multinomial-modeling approach to generate independent measures of item and source memory, Pereira, Sampaio, and Pinheiro (2019) failed to find enhanced source memory judgments about production in a mixed-list design. The researchers noted several methodological details that may explain why they did not replicate Ozubko et al.'s (2014) source memory advantage for aloud items. An additional possibility is that Pereira et al.'s (2019) within-subjects manipulation of positive, negative, and neutral valence words may have drawn attention away from the production manipulation at encoding, thereby weakening its effect on source memory. Consistent with this possibility, literature reviewed in Pereira et al. shows that source memory either is not affected or is impaired for items of emotional valence relative to neutral items. However, if the production effect on source memory for input modality proves to be unreliable, then Fawcett and Ozubko's (2016) claim that production boosts recollection in a mixed-list design may need to be qualified. Further tests of whether mixed-list production enhances source memory for encoding task and item details, such as those that we tested here, are warranted.

Another important question for future research to address is whether encoding strategies can selectively enhance source memory without improving item recognition. To the extent that item and source memory are deemed to reflect distinct processes (e.g., Yonelinas, 2002), it may be possible to show an influence of an encoding manipulation on source memory in the absence of an influence of item memory (e.g., Kurilla & Westerman, 2010). In support of that possibility, theta entrainment following a study phase has been found to enhance source memory without improving item recognition (Roberts, Clarke, Addante, & Ranganath, 2018).

Limits on the effectiveness of production for enhancing source memory join other potential boundary conditions on the utility of a production strategy. For one, there is no evidence that production enhances implicit memory (cf. MacLeod et al., 2010; see also Forrin, MacLeod, & Ozubko, 2012), although this should occur if production increases item strength. For another, Hourihan and Smith (2016) did not find that production of names enhanced people's ability to learn face-name associations—perhaps because the face, which serves as the test cue, was not the item produced at study. Most strikingly, production does not seem to improve free recall: The pure-list effect is generally absent, and the mixed-list effect reflects a cost to silent items rather than a benefit to aloud items—consistent with an item-order account (Forrin & MacLeod, 2016; Jones & Pyc, 2014; Jonker, Levene, & MacLeod, 2014; Lambert, Bodner, & Taikh, 2016).

For every limitation, however, it seems that a new example of the applied value of production is discovered. As examples, production can improve memory for text (Ozubko et al., 2012b), vocabulary learning (Rumbaugh & Landau, 2018), and second-language learning (Hopman & MacDonald, 2018). For a simple strategy, there is complexity under the hood. Pinpointing the "engines of production" will enable researchers to identify and dial in other suitable applications for this often-effective and easily implemented encoding strategy.

Open practices statement The data and materials for all experiments are available at https://osf.io/k73r4/

References

Addante, R. J., Ranganath, C., & Yonelinas, A. P. (2012). Examining ERP correlates of recognition memory: Evidence of accurate source recognition without recollection. *NeuroImage, 62*, 439-450. doi:https://doi.org/10.1016/j.neuroimage.2012.04.031

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review, 97*, 548–564. doi:https://doi.org/10.1037/0033-295X.97.4.548

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*, 201–210. doi:https://doi.org/10.3758/BF03193441

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. doi:https://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. doi:https://doi.org/10.3758/s13423-013-0485-1

Bodner, G. E., Jamieson, R. K., Cormack, D. T., McDonald, D.-L., & Bernstein, D. M. (2016). The production effect in recognition memory: Weakening strength can strengthen distinctiveness. *Canadian Journal of Experimental Psychology, 70*, 93–98. doi:https://doi.org/10.1037/cep0000082

Boywitt, C. D., & Meiser, T. (2012a). Bound context features are integrated at encoding. *The Quarterly Journal of Experimental Psychology, 65*, 1484–1501. doi:https://doi.org/10.1080/17470218.2012.656668

Boywitt, C. D., & Meiser, T. (2012b). The role of attention for context-context binding of intrinsic and extrinsic features. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 1099–1107. doi:https://doi.org/10.1037/a0026988.

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

Diana, R. A., Yonelinas, A. P., & Ranganath, C. (2008). The effects of unitization on familiarity-based source memory: Testing a behavioral prediction derived from neuroimaging data. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 1099–1107. doi:https://doi.org/10.1037/0278-7393.34.4.730

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. doi:https://doi.org/10.3758/BF03196152

Dodson, C. S., Prinzmetal, W., & Shimamura, A. P. (1998). Using Excel to estimate parameters from observed data: An example from source memory data. *Behavioral Research Methods, Instruments & Computers, 30*, 517–526. doi:https://doi.org/10.3758/BF03200685

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavioral Research Methods, Instruments & Computers, 28*, 1–11. doi:https://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. doi:https://doi.org/10.1016/j.actpsy.2012.10.001

Fawcett, J. M., & Ozubko, J. D. (2016). Familiarity, but not recollection, supports the between-subject production effect in recognition memory. *Canadian Journal of Experimental Psychology, 70*, 99–115. doi:https://doi.org/10.1037/cep0000089

Fernandes, M. A., Wammes, J. D., & Meade, M. E. (2018). The surprisingly powerful influence of drawing on memory. *Current Directions in Psychological Science, 27*, 302–308. doi:https://doi.org/10.1177/0963721418755385

Forrin, N. D., MacLeod, C. M., & Ozubko, J. D. (2012). Widening the boundaries of the production effect. *Memory & Cognition, 40*, 1046–1055. doi:https://doi.org/10.3758/s13421-012-0210-8

Forrin, N. D., Jonker, T. R., & MacLeod, C. M. (2014). Production improves memory equivalently following elaborative vs. non-elaborative processing. *Memory, 22*, 470–480. doi:https://doi.org/10.1080/09658211.2013.798417

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

Forrin, N. D., Groot, B., & MacLeod, C. M. (2016). The *d*-prime directive: Assessing costs and benefits in recognition by dissociating mixed-list false alarm rates. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*, 1090–1111. doi:https://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. doi:https://doi.org/10.3758/BF03213478

Hopman, E. W. M., & MacDonald, M. C. (2018). Production practice during language learning improves comprehension. *Psychological Science, 29*, 961–971. doi:https://doi.org/10.1177/0956797618754486

Hourihan, K. L., & Churchill, L. A. (2020). Production of picture names improves picture recognition. *Canadian Journal of Experimental Psychology, 74*, 35–43. doi:https://doi.org/10.1037/cep0000185

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. doi:https://doi.org/10.3758/s13423-014-0648-8

Huff, M. J., Bodner, G. E., & Gretz, M. R. (2019). Distinctive encoding of a subset of DRM lists yields benefits, but also costs and spillovers. *Psychological Research*. Advance online publication. doi:https://doi.org/10.1007/s00426-019-01241-

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*, 3–28. doi:https://doi.org/10.1037/0033-2909.114.1.3

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. https://doi.org/10.1037/a0033337

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

Kroneisen, M., & Bell, R. (2018). Remembering the place with the tiger: Survival processing can enhance source memory. *Psychonomic Bulletin & Review, 25*, 667–673. doi:https://doi.org/10.3758/s13423-018-1431-z

Kurilla, B. P., & Westerman, D. L. (2010). Source memory for unidentified stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 398–410. doi:https://doi.org/10.1037/a0018279

Lambert, A. M., Bodner, G. E., & Taikh, A. (2016). The production effect in long-list recall: In no particular order? *Canadian Journal of Experimental Psychology, 70*, 165–176. doi:https://doi.org/10.1037/cep0000086

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

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

MacLeod, C. M., Pottruff, M. M., Forrin, N. D., & Masson, M. E. J. (2012). The next generation: The value of reminding. *Memory & Cognition, 40*, 693–702. doi:https://doi.org/10.3758/s13421-012-0182-8

Macmillan, N. A., & Creelman, C. D. (1991). Detection theory: A user's guide. New York, NY: Cambridge University Press.

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

Mather, M. (2007). Emotional arousal and memory binding: An object-based framework. *Perspectives in Psychological Science, 2*, 33–52. doi:https://doi.org/10.1111/j.1745.6916.2007.00028.x

McCurdy, M. P., Leach, R. C., & Leshikar, E. C. (2017). The generation effect revisited: Fewer generation constraints enhances item and context memory. *Journal of Memory and Language, 92*, 202–216. doi:https://doi.org/10.1016/j.jml.2016.06.0070

McDonough, I. M., & Gallo, D. A. (2012). Illusory expectations can affect retrieval-monitoring accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 391–404. https://doi.org/10.1037/a0025548

Meiser, T., & Bröder, A. (2002). Memory for multidimensional source information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 116–137. doi:https://doi.org/10.1037/0278-7393.28.1.116

Misirlisoy, M., Tanyas, H., & Atalay, N. B. (2019). Does survival context enhance memory for source? A within-subjects comparison. *Memory, 27*, 781–791. doi:https://doi.org/10.1080/09658211.2019.1566928

Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language, 70*, 1–12. https://doi.org/10.1016/j.jml.2013.09.007

Mulligan, N. W. (2004). Generation and memory for contextual detail. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 838–855. doi:https://doi.org/10.1037/0878-7393.30.4.838

Mulligan, N. W. (2011). Generation disrupts memory for intrinsic context but not extrinsic context. *The Quarterly Journal of Experimental Psychology, 64*, 1543–1562. doi:https://doi.org/10.1080/17470218.2011.562980

Mulligan, N. W., Lozito, J. P., & Rosner, Z. A. (2006). Generation and context memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 836–846. doi:https://doi.org/10.1037/0278-7393.32.4.836

Mulligan, N. W., Besken, M., & Peterson, D. (2010). Remember-know and source memory instructions can qualitatively change old-new recognition accuracy: The modality-match effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 558–566. doi:https://doi.org/10.1037/a0018408

Murnane, K., & Bayan, U. J. (1996). An evaluation of empirical measures of source identification. *Memory & Cognition, 24*, 417–428. doi:https://doi.org/10.3758/BF03200931

Nieznański, M. (2011). Generation difficulty and memory for source. *The Quarterly Journal of Experimental Psychology, 64*, 1593–1608. doi:https://doi.org/10.1080/17470218.2011.566620

Nieznański, M. (2012). Effects of generation on source memory: A test of the resource tradeoff versus processing hypothesis. *Journal of Cognitive Psychology, 24*, 765–780. doi:https://doi.org/10.1080/20445911

Nieznański, M. (2014). Context reinstatement and memory for intrinsic versus extrinsic context: The role of item generation at encoding or retrieval. *Scandinavian Journal of Psychology, 55*, 409–419. doi:https://doi.org/10.1111/sjop.12153

Overman, A. A., Richard, A. G., & Stephens, J. D. W. (2017). A positive generation effect on memory for auditory context. *Psychonomic Bulletin & Review, 24*, 944–949. doi:https://doi.org/10.3758/s13423-016-1169.4

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. doi:https://doi.org/10.1037/a0020604

Ozubko, J. D., Gopie, N., & MacLeod, C. M. (2012a). Production benefits both recollection and familiarity. *Memory & Cognition, 40*, 326–338. doi:https://doi.org/10.3758/s13421-011-0165-1

Ozubko, J. D., Hourihan, K. L., & MacLeod, C. M. (2012b). Production benefits learning: The production effect endures and improves memory for text. *Memory, 20*, 717–727. doi:https://doi.org/10.1080/09658211.2012.699070

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. doi:https://doi.org/10.1080/09658211.2013.800554

Pereira, D. R., Sampaio, A., & Pinheiro, A. P. (2019). Differential effects of valence and encoding strategy on internal source memory and judgments of source: Exploring the production and the self-reference effect. *Frontiers in Psychology, 10*, 1326. doi:https://doi.org/10.3389/fpsyg.2019.01326

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137*, 615–625. https://doi.org/10.1037/a0013684

Riefer, D. M., Chien, Y., & Reimer, J. F. (2007). Positive and negative generation effects in source monitoring. *The Quarterly Journal of Experimental Psychology, 60*, 1389–1405. doi:https://doi.org/10.1080/17470210601025646

Roberts, B. M., Clarke, A., Addante, R. J., & Ranganath, C. (2018). Entrainment enhances theta oscillations and improves episodic memory. *Cognitive Neuroscience, 9*, 181–193. doi:https://doi.org/10.1080/17588928.2018.1521386

Rumbaugh, C. M., & Landau, J. D. (2018). Recognition and recall performance both benefit from the production effect with content-specific information. *Reading Psychology, 39*, 29–40. doi:https://doi.org/10.1080/02702711.2017.1361494

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

Starns, J. J., & Hicks, J. L. (2005). Source dimensions are retrieved independently in multidimensional monitoring tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 1213–1220. doi:https://doi.org/10.1037/0278-7393.31.6.1213

Taikh, A., & Bodner, G. E. (2016). Evaluating the basis of the between-group production effect in memory. *Canadian Journal of Experimental Psychology, 70*, 186–194. doi:https://doi.org/10.1037/cep0000083

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

Wammes, J. D., Meade, M. E., & Fernandes, M. A. (2018). Creating recollection-based memory through drawing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44,* 734–751. doi:https://doi.org/10.1037/xlm0000445

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. doi:https://doi.org/10.1006/jmla.2002.2864