Memory & Cognition
https://doi.org/10.3758/s13421-023-01442-2

Reading text aloud benefits memory but not comprehension

Brady R. T. Roberts[1] · Zoey S. Hu[1] · Eloise Curtis[2] · Glen E. Bodner[2] · David McLean[1] · Colin M. MacLeod[1]

Accepted: 19 June 2023

Abstract
The production effect—that reading aloud leads to better memory than does reading silently—has been defined narrowly with reference to memory; it has been explored largely using word lists as the material to be read and remembered. But might the benefit of production extend beyond memory and beyond individual words? In a series of four experiments, passages from reading comprehension tests served as the study material. Participants read some passages aloud and others silently. After each passage, they completed multiple-choice questions about that passage. Separating the multiple-choice questions into memory-focused versus comprehension-focused questions, we observed a consistent production benefit only for the memory-focused questions. Production clearly improves memory for text, not just for individual words, and also extends to multiple-choice testing. The overall pattern of findings fits with the distinctiveness account of production—that information read aloud stands out at study and at test from information read silently. Only when the tested information is a very close match to the studied information, as is the case for memory questions but not for comprehension questions, does production improve accuracy.

Keywords Production effect · Reading · Text · Comprehension · Memory

Testing is commonly used to assess both formal classroom learning and informal everyday learning, and there are many ways to study for such tests. Consequently, numerous study strategies have been researched with the goal of making learning more effective. A few examples include self-quizzing (Roediger & Karpicke, 2006), spaced learning across time (Carpenter et al., 2012), and explaining knowledge to oneself elaboratively (Chi et al., 1994). These strategies have been shown to improve learning but they all share a common drawback: They can be time-consuming when studying a large volume of content, which is typical in many educational settings (e.g., university courses).

A simpler study strategy that might enhance subsequent memory is reading aloud. This idea was first mentioned by Gates (1917), when he noted that participants "reported that practice in accurate pronunciation of the material was an aid in learning" (p. 67). Barlow (1928) first put this introspection to experimental test: He reported better memory for nonsense syllables learned by reading aloud than by reading silently. In a controlled laboratory setting, Hopkins and Edwards (1972) confirmed that this strategy was successful in improving memory for lists of words. Decades later, the phenomenon that reading aloud improves memory (compared with reading silently) was labelled *the production effect* by MacLeod et al. (2010). MacLeod et al. reported a series of experiments demonstrating a robust and consistent benefit of reading aloud (Experiments 1 and 3), mouthing (Experiment 5), and reading aloud in addition to generation (Experiment 7) or in addition to semantic processing (Experiment 8), relative to reading silently.

Initial studies observed a benefit for production in within-subject, mixed-list experiments but not in between-subjects, pure-list experiments (Dodson & Schacter, 2001; Hopkins & Edwards, 1972; MacLeod et al., 2010). A mixed list contains both aloud and silent words whereas a pure list contains only aloud words or only silent words. Thus, the production effect appeared to emerge only when both aloud and silent words were studied within the same list. This apparent limitation of the effect to a within-subject design prompted researchers, beginning with Conway and Gathercole (1987), to speculate that the production

✉ Brady R. T. Roberts
bradyrtroberts@gmail.com

[1] Department of Psychology, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada

[2] College of Education, Psychology and Social Work, Flinders University, Adelaide, Australia

Published online: 13 July 2023

benefit could be due to the distinctiveness of the aloud content compared with the silent content (Gathercole & Conway, 1988; MacLeod et al., 2010; Ozubko & MacLeod, 2010; Ozubko et al., 2014). Specifically, the proposal was that, during study, words that were read aloud resulted in additional encoding that was then useful at the time of test to diagnose whether an item had been studied. This fit with what Dodson and Schacter (2001) labeled a *distinctiveness heuristic*.

Subsequent work has shown that, in fact, a production effect can be detected in pure-list, between-subjects designs as well (e.g., Fawcett, 2013), but is often considerably smaller than its mixed-list, within-subject counterpart (see MacLeod & Bodner, 2017, for a brief review). Based on findings showing that the pure-list result is expressed only in familiarity whereas the mixed-list result is expressed in both familiarity and recollection, Fawcett and Ozubko (2016) argued that there is a small strengthening effect of production in both designs but that only in the mixed-list case is the larger effect of distinctiveness added to this strength boost.

In essence, earlier work contrasting the production effect in within-subject relative to between-subjects designs, and in pure-list relative to mixed-list paradigms, has led researchers to suggest that there are both distinctiveness-based and strength-based contributions to improved memory for content read aloud (see MacLeod & Bodner, 2017, for a review). This is an important point to consider when making predictions about how production may or may not aid memory in other research contexts.

To date, almost all of the research specifically examining the "production effect" has used word lists as the to-be-learned material. Only three published studies have used text passage materials—Ozubko et al. (2012), Kline (2019), and Icht et al. (2022). Using a within-subject design, Ozubko et al. had participants read some paragraphs of a text aloud and other paragraphs silently, resulting in a reliable production advantage on a fill-in-the-blanks test. Icht et al. (2022) implemented a mixed-list design in which half of the sentences for a given text passage were read aloud and half were read silently, once again using a fill-in-the-blanks test. Icht et al. found that this format led to a production benefit for text passage content both in younger and in older adult participants. In contrast, when using a between-subjects design in which participants read passages either all aloud or all silently, Kline reported no production advantage for recalling a cued paragraph or for yes/no recognition of intact versus altered sentences. It is likely, therefore, that the between-subjects production effect will be smaller than the within-subject effect for text just as is the case for single words. Perhaps more importantly, text passages read entirely aloud should elicit less of a distinctiveness advantage for the produced content because those passages would constitute a type of 'pure-list' design.

Thus far, too, studies of the production effect have focused on memory, using primarily recognition tests, occasionally recall tests, and more rarely other forms of testing (e.g., fill-in-the-blank tests). All of these tests are aimed at memory for single words—literal, verbatim memory. But might production also influence other cognitive skills? In particular, given that memory is improved by production, would comprehension, which necessarily relies on memory, also show a benefit? In studying, the goal is for learners not only to remember what they have read but also to comprehend what the material means.

Seminal work from Bransford and Johnson (1972) demonstrated that comprehension and memory are intimately related. More recent evidence also suggests that verbal memory and comprehension share underlying neural substrates (Leff et al., 2009). Perhaps most critically, it has also been shown that—when episodic memory for a passage is high—those with poor comprehension abilities perform just as well as controls on tests of inferential ability (Hua & Keenan, 2014). This highly relevant work from Hua and Keenan (2014) underscores how comprehension is necessarily reliant on memory, and that when episodic memory for a passage is poor, inferencing is difficult.

A simple extrapolation might, then, expect a benefit to comprehension via its reliance on memory. But comprehension goes beyond literal memory to understanding, and consequently tests of comprehension typically do not rely solely on the exact wording of what has been read. In the education literature, comprehension has been divided into three subtypes: literal, inferential, and evaluative (Basaraba et al., 2013). Literal comprehension is essentially rote memorization, inferential comprehension requires drawing inferences not explicitly stated, and evaluative comprehension takes into account broader contextual factors that underlie the passage or the society in which the author was writing (Basaraba et al., 2013). In inferential and evaluative comprehension, the reader should be able to extract gist, themes, and inferences in comprehending—all of which go beyond retrieving the verbatim text from memory.

Under the strengthening account of the production effect, the benefit derives from overall strengthened memory for produced items. In the context of studying text passages, strengthened memories for passages read entirely aloud should enhance overall recall, supporting later inferencing on a comprehension test. That is, if all of the information read aloud is strengthened, then recall of that content should be facilitated and so too should any resulting comprehension of the remembered content. In short, remembering details of a passage should aid logical formulations about broader themes or connections within that passage. This should occur regardless of whether passages are read entirely aloud, constituting—as previously mentioned—a pure-list design of sorts.

On the other hand, under the distinctiveness account, the memory benefit stems from matching verbatim memory on the retrieval test. If distinctiveness is the operative mechanism in production, then comprehension beyond the literal would not be expected to show a benefit. If the words read aloud are distinctive in memory, recognition of those words should be facilitated but overall understanding of the passage should not be. Therefore, performance on questions targeting verbatim facts from a passage should be enhanced by reading aloud, but broader comprehension as supported by connections between various aspects of the passage should not benefit.

We set out to adjudicate between these two accounts. In so doing, we expected to replicate the memory advantage for text passages reported by Ozubko et al. (2012) and Icht et al. (2022), extending their findings from a fill-in-the-blanks test to a multiple-choice test. In the education literature, multiple-choice tests are thought to assess the same knowledge as open-format tests, and multiple-choice tests have been shown to allow for discrimination of reading comprehension levels (Alonzo et al., 2009; Rupp et al., 2006). Critically, here we sought to determine whether the production benefit is limited to verbatim memory, as a distinctiveness account would predict, or whether the advantage for text read aloud extends to non-literal comprehension, as a strength account would predict.

Experiment 1

Each of our experiments used university-level reading comprehension test materials as the information to be studied. These reading materials required participants to comprehend the content and to answer both memory-focused and comprehension-focused questions. By including both types of questions, we could also use the memory-focused questions as a kind of manipulation check (cf. Festinger, 1953) of the production advantage, permitting more straightforward interpretation of the results for the comprehension questions. The test format was also novel in the realm of production studies: We switched to multiple-choice testing, as widely used in universities. Finally, given the influence of experimental design on the production effect, participants in the current set of studies read multiple passages, some entirely aloud and some entirely silently—a within-subject (but not within-passage) procedure.

Method

Participants A target sample size was calculated a priori using G*Power software (Version 3.1.9.3; Faul et al., 2007), estimating a medium effect size of production on comprehension ($d_z = 0.5$). This power analysis suggested a total of 34 participants to achieve 80% power (two-tailed t test, $\alpha = .05$), which was then set as the minimum recruitment goal. On this basis, we collected data from 50 undergraduate students at the University of Waterloo who participated for course credit. Data of two participants were not included due to incomplete files. *R* statistical software was then used to exclude any participant whose performance was ± 3 standard deviations away from the mean on accuracy of responses to test questions, calculated separately for each of the four levels of the experimental design; this resulted in the exclusion of one participant. The final sample of 47 participants used in the statistical analyses was 76% female, with age ranging from 17 to 36 ($M = 21$, $SD = 3.7$).

Because of the relatively demanding reading task, we recruited only participants who self-reported being fluent in speaking, reading, and writing English on a prescreening questionnaire administered at the beginning of the academic term. This study was approved by the Office of Research Ethics at the University of Waterloo (Project #41191).

Materials The reading material consisted of 10 short passages, 5 each from forms G and H of the *Nelson-Denny Reading Test* (Nelson & Denny, 1929; see Brown et al., 1993).[1] The Nelson-Denny is a standardized reading test designed to measure reading abilities of high school and university students. Each form contains one long passage and six shorter passages. For our purpose, we discarded the long passage and one of the shorter passages from each form. The 10 remaining passages included the following topics: (1) successes of George Carver, (2) concept of self-involvement, (3) uses of common acids, (4) work of hydrographers, (5) schooling in simple and complex societies, (6) famous poet Gwendolyn Brooks, (7) concepts of extraversion and introversion, (8) basic facts about chemical compounds, (9) work of soil conservationists, and (10) concepts of referential and expressive types of symbols. Passages were on average 199 words in length (min = 156, max = 226).

On the test, each passage was followed by five multiple-choice questions, each question having five response options and only one correct answer. Selection of passages for a given participant was entirely random (i.e., the passages from Forms G and H were randomly intermingled). The order of questions and answer options after each passage was not randomized: Their order was constant and preserved that of the Nelson-Denny.

[1] We gratefully acknowledge Riverside Publishing, holders of the copyright, for granting permission to use these materials.

Between one and four (of the five total) questions following a passage were judged to be memory-focused in that they asked about facts directly presented in the passage, with the wording of the correct answer in these questions typically very similar to that in the passage. For example, one passage about schools in simple versus complex societies asked this question: "*It was said that in simple societies children learn what is needed to:*" where the correct answer from among the five choices was "*survive*." This information was directly presented in the passage in the sentence "*She points out that children in relatively simple societies learn what everyone agrees they should know because they must survive*."

Between one and four questions following a passage were judged to be comprehension-focused. Matching the criteria for 'inferential' and 'evaluative' comprehension questions as outlined by Basaraba et al. (2013), these questions asked about elements such as the theme or tone of the passage, or they probed inferences from the passage. For example, from the same schools and societies passage just referenced, one of the questions asked: "*The main idea in the second paragraph was developed primarily through*:", where the correct answer from among the five choices was "*contrast*." The answer did not appear in the text, so this question required understanding of the goal of the second paragraph which was to contrast schools in simple versus complex societies.

Our critical research question required that we distinguish the memory-focused questions from the comprehension-focused questions. Two of the authors did this independently for the 50 questions, and there was almost complete agreement in identifying 27 questions as memory-focused and 23 questions as comprehension-focused. There were disagreements for only three questions, which were then discussed in-depth before a decision was made to assign the questions as memory-focused or comprehension-focused questions.

Apparatus The experiment was programmed using E-Prime 3.0 (Psychology Software Tools, 2016) running on Windows software using a 15-inch monitor. Each passage and its five questions were printed in black Calibri size 18 font on a white background, intended to resemble normal text.

Procedure Participants signed up for the study via an online recruitment site and came to the laboratory individually. Before they began the task, they were given (1) an information letter that briefly outlined the procedure and (2) a consent form to sign. At the outset, participants were shown two instruction screens that asked them to read each passage once only, reading at their normal reading speed for good understanding, and then to answer the questions. Participants were told that they would read some of the passages aloud and others silently. No mention was made of the production effect.

The order of the five aloud and five silent passages was randomly generated for each participant. Successive passages read by a participant were not permitted to be in the same condition more than three times in a row. Assignment of aloud and silent conditions to each passage was counterbalanced. Before each passage appeared, the word ALOUD or SILENT was presented in the center of the screen, indicating to the participant how they should read that passage. Then, the participant pressed the space bar to continue to the passage. After reading a passage, the participant pressed the space bar again which removed the passage and replaced it with a blank screen for 500 ms. After the blank screen, the first question, along with its five multiple choice options (in the same order as in the Nelson-Denny test), appeared in the center of the screen. For each passage, all five questions were presented individually with the participant selecting an answer by pressing the corresponding letter on the keyboard. Once the keyboard registered a response, the next question appeared on the screen after a 500-ms blank screen. Participants read all questions and answer options silently. Reading the passages and responding to the questions were both self-paced with no time limit.[2]

After the participant had read all 10 passages and answered all 50 questions, a final screen announced the end of the experiment. Before leaving the laboratory, participants were given a feedback letter that provided details about the study and the researchers' contact information, and any questions that they had were answered. An experimenter was present for the entire session, ensuring that the passages were read as instructed.

All experimental materials, programs, data, and statistical code for this experiment are available on the Open Science Framework (OSF; https://tinyurl.com/PE-and-Comp-Peer-Review).

Results

Data cleaning and statistical analyses throughout all four experiments were performed using *R* (Version 4.1.1; R Core Team, 2020), enlisting the *afex* (Version 1.1-1; Singmann et al., 2022) and *emmeans* (Version 1.8.1-1; Lenth et al., 2022) packages. The mean proportions for Experiment 1 are displayed in the top section of Table 1. Figure 1 displays the proportion correct scores for each condition for every participant.

The critical question was whether production would benefit not only memory but also comprehension of text. A 2 × 2

[2] While prior work (e.g., Salasoo, 1986) has shown that reading aloud is often slower than reading silently—potentially offering more encoding support—this has been shown to have little to no influence on the production effect (MacLeod et al., 2010). The General Discussion provides more detail on this issue.

Table 1 Descriptive statistics for participant samples and accuracy in each condition

Experiment	Question Type	Encoding Condition	*N*	% Female	Age		Test Accuracy	
					M	*SD*	*M*	*SD*
1	Memory	Aloud	47	76.00%	21.0	3.7	0.81	0.14
	Memory	Silent					0.72	0.16
	Comprehension	Aloud					0.80	0.14
	Comprehension	Silent					0.74	0.16
2	Memory	Aloud	64	73.02%	35.7	15.6	0.81	0.15
	Memory	Silent					0.74	0.15
	Comprehension	Aloud					0.82	0.18
	Comprehension	Silent					0.81	0.19
3	Memory	Aloud	151	49.67%	39.2	11.7	0.85	0.11
	Memory	Silent					0.81	0.13
	Comprehension	Aloud					0.87	0.12
	Comprehension	Silent					0.85	0.12
4	Comprehension	Aloud	131	58.78%	34.9	12.3	0.54	0.15
	Comprehension	Silent					0.52	0.15

Demographic figures represent a single group of participants in each experiment

repeated-measures ANOVA compared the mean accuracy of responses across question type (memory vs. comprehension) and encoding condition (aloud vs. silent). The main effect of question type was not reliable, $F(1, 46) = 0.19$, $p = .663$, $\eta_p^2 < .01$, $BF_{01} = 5.95$[3]; overall accuracy was equivalent for memory questions ($M = .76$, $SD = .16$) and comprehension questions ($M = .76$, $SD = .17$). There was, however, the predicted main effect of encoding condition such that questions were answered more accurately for passages read aloud ($M = .80$, $SD = .14$) than for passages read silently ($M = .72$, $SD = .17$)—a production effect, $F(1, 46) = 25.79$, $p < .001$, $\eta_p^2 = .36$, $BF_{10} = 37$. The interaction between encoding condition and question type was not significant, $F(1, 46) = 0.76$, $p = .389$, $\eta_p^2 = .02$, $BF_{01} = 3.47$.

Because of our a priori hypotheses, we conducted planned comparisons to assess the mean accuracy of memory-focused questions and of comprehension-focused questions under each encoding condition. As is evident in Fig. 1, aloud passages produced a significantly higher mean score than did silent passages both for memory-focused questions, $t(46) = 4.14$, $p < .001$, $d = 0.60$, CI_{95} [0.29, 0.92], $BF_{10} = 163$, and for comprehension-focused questions, $t(46) = 2.26$, $p = .029$, $d = 0.32$, CI_{95} [0.03, 0.33], $BF_{10} = 1.58$. Put simply, production apparently benefited both memory and comprehension in this initial experiment.

[3] Throughout this article, Bayes factors were calculated using the *BayesFactor* (Version 0.9.12-4.4; Morey et al., 2011) package for *R*, enlisting a default Jeffreys–Zellner–Siow (JZS) prior with a Cauchy distribution (center = 0, $r = 0.707$). Bayes factors for the alternative (BF_{10}) are in comparison to intercept-only null models, except for interaction terms which are in comparison to models containing all main effects.

Discussion

Experiment 1 successfully demonstrated that reading text aloud as opposed to silently generates a significant improvement in multiple-choice accuracy for memory-based questions. The results also seemed to be promising with respect to comprehension: Production of text passages appeared to lead to higher accuracy for comprehension-based multiple-choice questions. However, the magnitude of the production effect observed for the comprehension test questions ($d = 0.33$) was smaller than that observed for the memory test questions ($d = 0.60$), and the Bayes factor for the effect on comprehension was weak. Both of these factors prompted follow-up experiments.

Experiment 2

Using the same *Nelson-Denny Reading Test* materials and a similar participant sample—but conducted at a different institution—Experiment 2 provided a second test of production's ability to improve both memory and comprehension of text. A new condition, reading text silently in an unusual font (Sans Forgetica; see Earp, 2018), was included in this experiment but is not the focus of the current investigation and is therefore described in the Appendix. Experiment 2 also measured participants' metamemory beliefs about the effectiveness of production. Using word lists, Castel

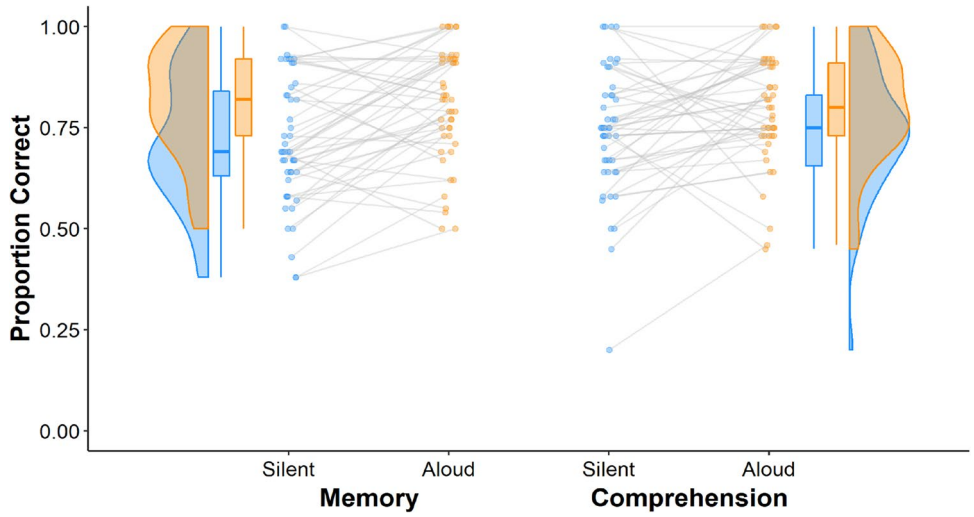

Fig. 1 Experiment 1: Proportion correct in each condition

et al. (2013) found that participants expected production to enhance memory for words, but whether participants would expect production to enhance memory for and/or comprehension of text was not known.

Method

Participants Due to the COVID-19 pandemic, this experiment and all subsequent experiments were conducted online.[4] A target sample size was calculated a priori using G*Power software (Version 3.1.9.3; Faul et al., 2007), estimating the effect size of production on comprehension at $d_z = 0.40$ in a within-subject manipulation. This power analysis suggested a total of 52 participants to achieve 80% power (two-tailed *t* test, $\alpha = .05$), which was then set as the minimum recruitment goal. We collected data from 114 participants using the MTurk (https://www.mturk.com/) online testing platform ($n = 73$; USA sample; HITs approved = 1,000–100,000; HIT approval rate = 98+), and an online research participation system at Flinders University ($n = 41$). Participants on MTurk took part in exchange for $3 USD, whereas undergraduate participants could choose to receive $5AUD or 0.5 course credits. Because the experiment involved a reading task and a memory test, we recruited only participants who self-reported being fluent in English. This study was approved by the Human Research Ethics Committee at Flinders University (Project #8337).

From this initial sample, participants' data were filtered out in sequential steps if they (1) were missing audio recordings (to confirm that they followed instructions) or failed attention checks (see Materials, below; $n = 36$), (2) had incomplete data ($n = 12$), or (3) were ±3 standard deviations away from the mean on test accuracy in any condition ($n = 2$). The final sample of 64 participants used in the statistical analyses was 71.88% female, with ages ranging from 17 to 71 ($M = 35.7$, $SD = 15.6$).[5] From this final sample, 58 reported English as their first language, and the remaining participants had an average self-reported mean fluency in English of 84%.[6]

Materials The materials were the same as in Experiment 1, except that we added an attention check question for which the answer should have been obvious if the text had been read (e.g., "This text was about: reading, writing, dance, travel, or research"). The purpose of this question was to determine whether participants were paying attention to the passage as they were reading it. Text passages and questions were presented in Arial font. To allow for equivalent numbers of trials in each encoding condition, text passages were randomly chosen anew for each participant.

[4] The production effect has been shown to be replicable in online settings, even amidst the COVID-19 global pandemic (Mama & Icht, 2020). Although some prior studies have been conducted live with an experimenter via video conferencing, we had no reason to suspect that the production effect would fail to manifest when participants were instructed to record their experiment sessions alone.

[5] Because age is known to affect memory performance, analyses for Experiments 2–4 were re-run using separate age groups (under or over 45 years old). The pattern of results for both age groups was identical to the overall findings reported here.

[6] In removing nonnative English speakers from the data set, the main effect of Condition became nonsignificant, $F(1, 57) = 3.15$, $p = .081$. All other results matched those reported here, including the significant production effect for memory-focused but not comprehension-focused questions.

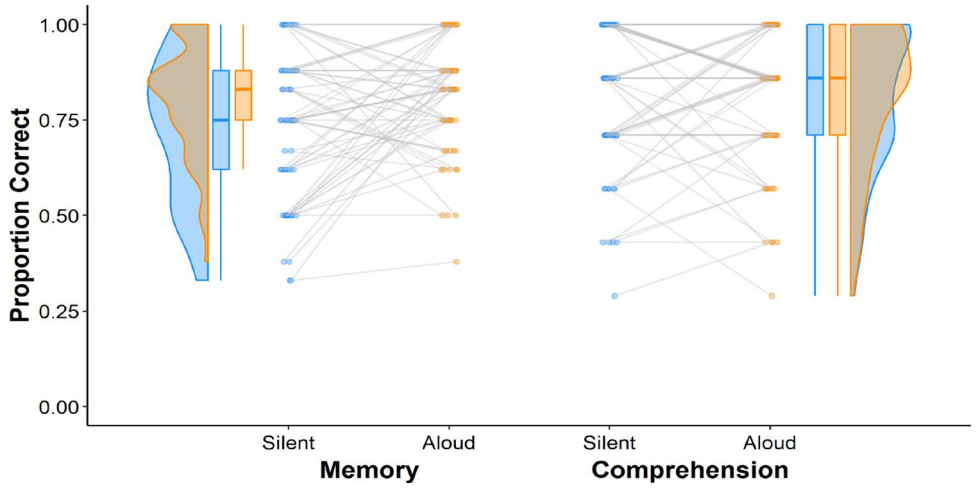

Fig. 2 Experiment 2: Proportion correct in the aloud and silent conditions

Apparatus Starting with this experiment, each experiment in this series was conducted online so participants took part using their personal computers. Participants also used their personal microphones (integrated or discrete) to permit audio recording during the study phase.

Procedure The procedure for this experiment was highly similar to that of Experiment 1. A 3 (encoding condition: silent, aloud, silent with unusual font Sans Forgetica) × 2 (question type: memory, comprehension) fully within-subject design was used. Participants completed the study online through Qualtrics software on a computer (use of a mobile device was blocked). For text passage trials that involved reading aloud, participants were required to record their reading of each text passage using the third-party Addpipe (Addpipe Development Team, 2022) plug-in which was integrated into Qualtrics via JavaScript code to check compliance. On each aloud reading trial, participants were instructed to click the record button before reading the passage. The 'next' button used to continue the experiment was invisible until a recording had been made.

Participants were told that they would read several text passages after each of which they would answer some multiple-choice questions about that text. They were told that they would read three texts silently, three aloud, and three silently but in an unusual font, as cued before each text. Order of the nine texts was randomized for each participant. Test questions appeared one at a time and participants clicked on their answer, guessing if necessary. After answering each question, participants rated their confidence in their response from 0–100% using a slider.

Finally, participants were asked whether they had genuinely done their best to follow the experiment instructions, and whether anything had happened during their participation that would influence their performance. Afterward, participants were provided with a feedback letter that outlined the purpose of the study and included the researchers' contact information. All experiment data, programs, statistical code, and a preregistration for this experiment are available on the Open Science Framework (OSF; https://tinyurl.com/PE-and-Comp-Peer-Review).

Results

The mean proportions of correct responses for Experiment 2 are displayed in the second section of Table 1. Figure 2 displays the proportion correct scores for every participant.

A 2 × 2 repeated-measures ANOVA was conducted to compare the accuracy of responses across question type (memory vs. comprehension) and encoding condition (aloud vs. silent). There was a reliable main effect of question type, $F(1, 63) = 4.33$, $p = .025$, $\eta_p^2 = .06$, $BF_{10} = 0.68$, with superior accuracy for comprehension questions ($M = .81$, $SD = .18$) relative to memory questions ($M = .77$, $SD = .17$; see Fig. 2). There was also the predicted main effect of encoding condition such that questions were answered more accurately for passages read aloud ($M = .81$, $SD = .16$) than for passages read silently ($M = .77$, $SD = .19$), $F(1, 63) = 5.25$, $p = .042$, $\eta_p^2 = .08$, $BF_{10} = 0.67$. The interaction between encoding condition and question type was nonsignificant, $F(1, 63) = 3.49$, $p = .066$, $\eta_p^2 = .05$, $BF_{01} = 1.89$.

Given our critical question, we next carried out planned comparisons to assess the effect of production on memory questions versus comprehension questions. As is evident in Fig. 2, aloud passages produced a significantly higher mean score than did silent passages for memory-focused questions, $t(63) = 2.69$, $p = .009$, $d = 0.34$, CI_{95} [0.08, 0.59], $BF_{10} = 3.71$, but not for comprehension-focused questions, $t(63) = 0.33$, $p = .742$, $d = 0.04$, CI_{95} [−0.20, 0.29], $BF_{01} = 6.93$.

Metacognition (accuracy predictions) Participants' mean predicted accuracy was 64.12% (SD = 24.70) in the silent condition and 59.36% (SD = 25.62) in the aloud condition. In terms of the expected value of production, participants' predictions did not differ significantly for the aloud texts vs. the silent texts, $t(65) = 1.42$, $p = .16$. Thus, their predictions did not align with their performance given that overall test accuracy actually was higher in the aloud condition than in the silent condition—demonstrating a mismatch between their metacognition and performance.

Discussion

As a conceptual replication, Experiment 2 failed to confirm a benefit of production on comprehension of text: Production benefitted only memory. Although nonsignificant, the Bayes factor for the interaction fell short of providing substantial evidence for the null (i.e., $BF_{01} > 3$). An analysis targeting the production effect in comprehension did, however, yield substantial Bayesian evidence for the null effect. Clearly, our manipulation had worked, given the significant production effect found in memory, matching prior literature. It is possible, of course, that Experiment 2 was underpowered to detect the production effect for comprehension because that effect may be smaller than the effect for memory. That there was also substantial Bayesian evidence for the null effect of production on comprehension in this experiment further raises suspicions that the benefit of production on comprehension seen in Experiment 1 may have been a false positive result—a Type I error. In addition, it seems that moving to an online format attenuated even the reliable production effect found for memory questions ($d = 0.60$ in Experiment 1 vs. $d = 0.35$ in Experiment 2). To address these possibilities, we proceeded with another conceptual replication in Experiment 3, this time with a much larger sample size.

We also observed that participants' metacognition in terms of accuracy predictions was not well aligned with their actual accuracy. They did not experience improved accuracy for texts that they read silently rather than aloud. Of course, given the modest differences in accuracy between the conditions, the null difference in their predictions may not be surprising. Moreover, we did not obtain separate predictions for memory versus comprehension questions.

Finally, a limitation of this study was that the typical 50–50% share of aloud and silent trials was violated in favor of a 33–33–33% split (aloud, silent, silent with Sans Forgetica font). Previous work by Icht et al. (2014) showed that statistical distinctiveness may contribute to better memory for rarer items in the production effect paradigm. While there was ostensibly an even 33-33% split of aloud and silent trials in this experiment, it is also possible that the silent and Sans Forgetica trials were seen as a combined set of 'silent' trials, meaning that the share of trials was actually 33% for aloud and a combined 66% for silent and Sans Forgetica trials. As a result, statistical distinctiveness of aloud trials may have biased results toward improved memory performance for passages read aloud. In each subsequent experiment, we returned to the more typical even distribution of aloud and silent trials.

Experiment 3

In Experiment 1, production enhanced accuracy for comprehension questions about texts, but this novel effect did not replicate in Experiment 2. Therefore, the primary goal of Experiment 3 was to revisit the question of whether production improves text comprehension. To collect data from a large and diverse sample online while ensuring high data quality, in Experiment 3 we switched from MTurk to Prolific, an online data collection platform that is well reputed, in part for its superior data quality (Peer et al., 2022).

Method

Participants A target sample size was calculated a priori using G*Power (Version 3.1.9.3; Faul et al., 2007), estimating the effect size of production on comprehension that was observed in Experiment 1 ($d_z = 0.32$). This power analysis suggested a total of 129 participants to achieve 95% power (two-tailed t test, $\alpha = .05$), which was then set as the minimum recruitment goal. On this basis, we collected data from 167 participants using the online recruitment system Prolific (https://www.prolific.co/). Participants took part in exchange for $8.08USD. We only recruited participants ages 18–64 who self-reported being fluent in English and having normal or corrected-to-normal vision on a prescreening questionnaire. This study was approved by the Office of Research Ethics at the University of Waterloo (Project #43987).

From this initial sample, participants' data were filtered out in sequential steps if they (1) took less than 15 minutes to complete the study ($n = 0$), (2) took more than 100 minutes to complete the study ($n = 2$), (3) were ±3 standard deviations away from the mean of the remaining participants for study duration ($n = 4$), (4) responded negatively on our instruction check which asked them whether they read the passages as instructed (i.e., aloud or silently; $n = 0$), (5) had read the passages before the experiment ($n = 2$), or (6) had self-reported nonideal conditions while completing the experiment ($n = 5$). R statistical software was then used to exclude participants whose performance was ± 3 standard deviations away from the mean on accuracy of responses to questions on the test, calculated separately in each of the four levels of the experimental design ($n = 3$). The final sample of 151 participants used in the

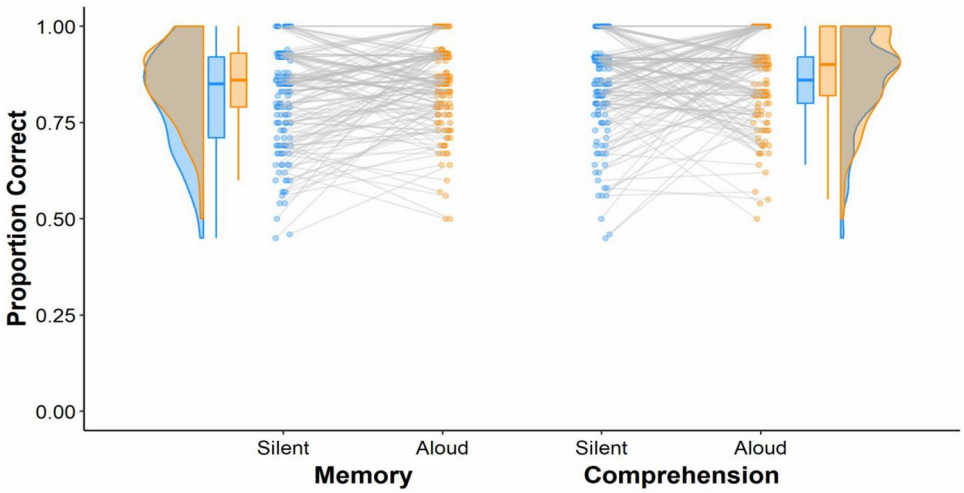

Fig. 3 Experiment 3: Proportion correct in each condition. *Note.* Median values are depicted by the center line on each box-and-whisker plot in this figure. Unique to this experiment, median values for the Aloud and Silent conditions were more disparate for comprehension-focused questions than for memory-focused questions; mean values showed the opposite pattern (as summarized in Table 1)

statistical analyses was 49.67% female, with ages ranging from 19 to 65 ($M = 39.2$, $SD = 11.7$).

Materials and apparatus The materials were those used in the first experiment. Here, though, each passage and its five questions were printed in black Arial size 16 font on a white background, again intended to resemble normal text. As in Experiment 2, participants completed this study online using their personal computers and audio was again recorded during the study phase using participants' personal microphones. New to this experiment, all trials were audio-recorded (including *silent* trials) to ensure compliance with instructions. Note that, because we were unaware of the default in the Addpipe software used to record participants during the study phase, regrettably audio recordings were auto-deleted after 28 days, before we could review them.

Procedure The procedure was similar to Experiment 1 except for a few minor changes: (1) there was no restriction placed on the number of possible consecutive aloud or silent passages, (2) there was no longer a 500-ms blank screen between passage and questions, (3) all questions were presented on the screen at once rather than individually, (4) questions were answered with a mouse click rather than a key press, and (5) to save time, only 6 of the 10 passages were used. The recording process was similar to that used in Experiment 2, making use of the third-party Addpipe service integrated into Qualtrics via JavaScript code. Once again, the assignment of passages to Aloud or Silent conditions was randomized, as was the presentation order of passages; the order of questions and answer options were not randomized.

After the participant had read all 6 passages and answered all 30 questions, a series of questions probed their intuitions and how ideal their testing environment was during the experiment. Participants were asked whether they thought that reading aloud versus silently (1) would affect their ability to remember the passage, and (2) would affect their ability to understand the passage. Participants were then asked whether they had read any of the passages before their involvement in the experiment. Finally, under assurance that their credit was already secured, participants were asked whether their participation had occurred under 'ideal' conditions (e.g., they were paying attention, understood the tasks, and were not distracted).

All experiment materials, programs, data, statistical code, and a preregistration for this experiment are available on the Open Science Framework (OSF; https://tinyurl.com/PE-and-Comp-Peer-Review).

Results

Our use of the Sans Forgetica font in this experiment arose from separate, exploratory research questions and bears no theoretical weight in our current investigation of production and comprehension. As such, we do not consider data from the Sans Forgetica trials in the analyses reported here.

The mean proportions of correct responses for Experiment 3 are displayed in the third section of Table 1. Figure 3 displays the proportion correct scores for every participant.

A 2 × 2 repeated-measures ANOVA compared the mean accuracy of responses across question type (memory vs. comprehension) and encoding condition (aloud vs. silent).

There was a reliable main effect of question type, $F(1, 150) = 13.54$, $p < .001$, $\eta_p^2 = .08$, $BF_{10} = 6.78$, with superior accuracy for comprehension questions ($M = .86$, $SD = .12$) relative to memory questions ($M = .83$, $SD = .12$; see Fig. 3).[7] There also was the predicted main effect of encoding condition: Questions were answered more accurately for passages read aloud ($M = .86$, $SD = .11$) than for passages read silently ($M = .83$, $SD = .13$), $F(1, 150) = 14.37$, $p < .001$, $\eta_p^2 = .09$, $BF_{10} = 9.36$. The interaction was, however, not significant, $F(1, 150) = 0.80$, $p = .372$, $\eta_p^2 = .01$, $BF_{01} = 6.35$.

We then conducted planned comparisons to assess the mean accuracy of memory-focused questions and of comprehension-focused questions under each encoding condition. Reading passages aloud produced a significantly higher mean score for memory-focused questions than did reading passages silently, $t(150) = 3.53$, $p < .001$, $d = 0.29$, CI_{95} [0.12, 0.45], $BF_{10} = 32.05$, but the encoding manipulation again failed to influence performance on comprehension-focused questions, $t(150) = 1.76$, $p = .079$, $d = 0.14$, CI_{95} [−0.02, 0.30], $BF_{01} = 2.43$. Put simply, as in Experiment 2 (but unlike in Experiment 1), production benefited memory but not comprehension.

Discussion

Experiment 3 again showed that the manipulation of reading aloud versus silently resulted in a significant memory benefit for text passages. The size of the production benefit for memory was again more modest than in Experiment 1, perhaps due to the use of online testing. Matching the results of Experiment 2, however, we again did not observe an effect of production on comprehension performance. In the fourth and final experiment, we sought again to achieve high statistical power, and therefore again collected a large sample. In addition, we thought it would be informative to generalize our findings to a new set of materials, albeit similar in being taken from another reading comprehension test.

Experiment 4

Our final experiment aimed to ascertain whether the effects observed thus far generalized to a new set of materials. It was possible that the Nelson-Denny materials lead to a production benefit for comprehension in the laboratory but not online, or that those materials simply do not provide a good test of comprehension, at least in our experimental context. Because production has been found to benefit memory so reliably, both in the broader literature and in the current series of studies, here we focused solely on whether production would influence comprehension—we did so using a new set of comprehension test materials, again making use of similar length text passages followed by multiple-choice testing.

Method

Participants A target sample size of $N = 129$ was calculated using the procedure outlined in Experiment 3. On this basis, we tested 167 participants recruited from the online recruitment system Prolific. Participants took part in exchange for $6.54 USD. Sign-up restrictions were identical to those in Experiment 3. This study was approved by the Office of Research Ethics at the University of Waterloo (Project #43987).

From this initial sample, participants' data were filtered out in sequential steps if they (1) took less than 10 minutes to complete the study ($n = 0$), (2) took more than 60 minutes to complete the study ($n = 5$), (3) were ±3 standard deviations away from the mean of the remaining participants for study duration ($n = 1$), (4) responded negatively on our instruction check which asked if they had read the passages as instructed (i.e., aloud or silently; $n = 1$), (5) had read the passages before the experiment ($n = 4$), (6) did not read a passage as instructed (as determined by audio recordings; $n = 18$), and (7) had self-reported non-ideal conditions while completing the experiment ($n = 7$). *R* statistical software was then used to exclude participants whose performance was ± 3 standard deviations away from the mean on accuracy of responses to questions on the test, calculated separately for each of the two levels of the experimental design ($n = 0$). The final sample of 131 participants used in the statistical analyses was 58.78% female, with ages ranging from 18 to 64 ($M = 34.9$, $SD = 12.3$).

Materials All text passages, questions, and answers were sourced from the educational website Test Prep Review (TPR; https://www.testprepreview.com/modules/reading1.htm), which offers example comprehension exams for free. Six new passages were obtained for use in this experiment, spanning the following topics: (1) the explorations of Magellan, (2) the accomplishments of Marie Curie, (3) the eruption of Mount Vesuvius, (4) the victories of ancient Athens, (5) the fate of Russian princess Anastasia, and (6) the invention of flight. Passages were on average 366 words in length (min = 306, max = 428). For each passage, we retained the five questions that were judged to assess comprehension of the preceding text passage (i.e., the correct answer was not available directly in the relevant passage). All passages and questions were displayed in black Arial size 16 font on a white background,

[7] These analyses were also conducted with the pre-registered target sample size of 129 participants by randomly selecting data from the full data set. The pattern of results was identical.

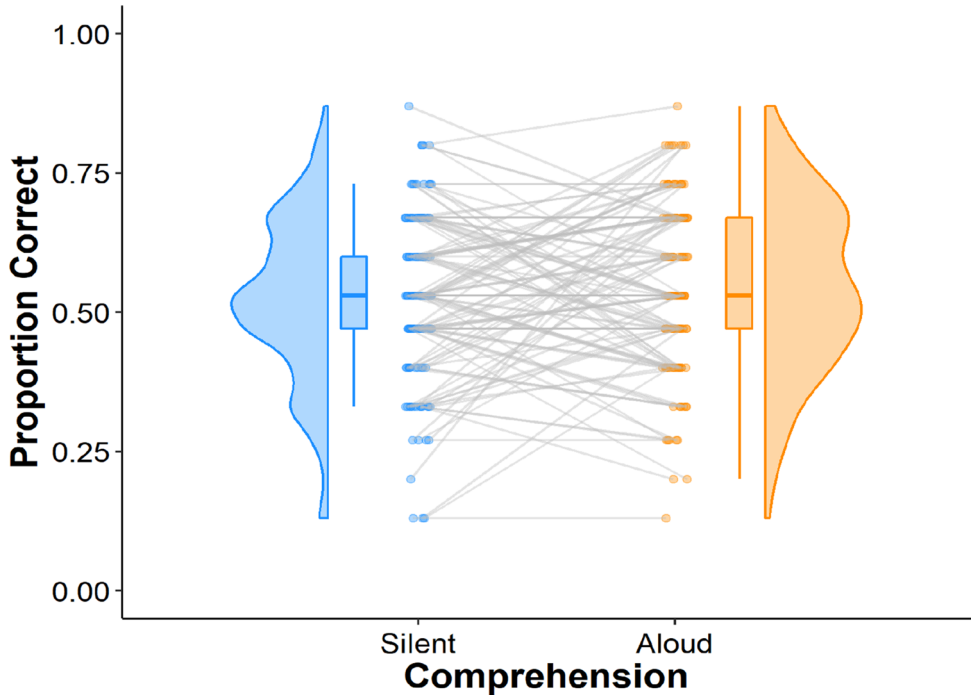

Fig. 4 Experiment 4: Proportion correct in each condition

intended to resemble normal text. Each question offered four multiple-choice answer options.

To ensure that the new TPR materials matched the NDRT materials used in the preceding three experiments, we compared passages in these two sets based on several lexical and affective factors using the custom *R* script 'lex-lookup' (Version 0.1.0; Yeung, 2023).[8] The average word count for each passage and for each sentence within those passages were extracted based on annotation using the *udpipe* package (Wijffels, 2023) for *R*. Concreteness (Brysbaert et al., 2014), word prevalence (Brysbaert et al., 2019), word frequency (Brysbaert & New, 2009), and age of acquisition (Kuperman et al., 2012) values were extracted for each word in each passage as well. Between 70–85% of the words had available values from the lexical databases, which were then averaged for each passage. Sentiment (i.e., affective statements) was calculated per sentence using the *sentimentr* package (Rinker, 2021) for *R*, then also averaged for each passage. Wilcoxon rank sum tests were performed on each variable to examine differences across passage type (NDRT vs. TPR). In the end, average word length (number of letters) was similar between material sets (TPR: $M = 4.94$, $SD = 2.53$, NDRT: $M = 5.07$, $SD = 2.86$), as was the average number of syllables per word (TPR: $M = 1.62$, $SD = 0.91$, NDRT: $M = 1.72$, $SD = 1.00$). Passage word count was significantly higher in the TPR ($M = 358.5$, $SD = 52.2$) relative to the NDRT materials ($M = 199.7$, $SD = 15.3$; $p = .001$). The two material sets did not differ on any other factor ($ps \geq .12$).

Apparatus As in Experiments 2 and 3, participants completed this study online using their personal computers. Similarly, audio was recorded with participants' personal microphones.

Procedure The procedure was identical to that of Experiment 3 with one exception: The orders of questions and of answer options were randomized for each participant. The audio recording process was identical to that used in Experiments 2 and 3.

All experiment materials, programs, data, statistical code, and a pre-registration for this experiment are available on the Open Science Framework (OSF; https://tinyurl.com/PE-and-Comp-Peer-Review).

Results

The mean proportions of correct responses for Experiment 4 are displayed in the bottom section of Table 1. Figure 4 displays the proportion correct scores for every participant.

A paired-samples *t* test was used to compare the mean accuracy of the comprehension-focused questions under each

[8] Thank you to Ryan Yeung for his assistance with this text analysis.

encoding condition. As is evident in Fig. 4, aloud and silent passages produced similar performance, $t(130) = 1.34$, $p = .181$, $d = 0.12$, CI_{95} [−0.05, 0.29], $BF_{01} = 4.28$[9]. Even with an entirely comprehension-focused set of materials, production once again failed to benefit comprehension. This was also evident at the level of individual participants: 43.51% showed numerically better performance for aloud trials, 38.17% showed numerically better performance for silent trials, and 18.32% showed equivalent performance in each condition.

Discussion

In this final experiment, our goal was to examine the effect of production on comprehension using a new set of materials. Our thinking was that perhaps the Nelson-Denny materials used in the preceding three experiments were responsible for the inconsistent findings, or they were not ideally suited for online testing. We also viewed generalization as valuable at this juncture. Having clearly demonstrated the benefit of production on memory for text in Experiments 1–3, we chose here to emphasize the question of principal interest: Would reading aloud (relative to reading silently) benefit comprehension? Again, however, the answer was negative: There was no effect of production on a test that assessed deeper understanding of the studied passages.

General discussion

We know from a now considerable body of research that, relative to reading silently, reading aloud benefits memory (see MacLeod & Bodner, 2017, for a brief review). Although the bulk of this research has used single words or other isolated items as the to-be-remembered stimuli, there is also evidence that production benefits memory for text passages (e.g., Icht et al., 2022; Ozubko et al., 2012; Todorovic, 2020). Consequently, we fully expected memory for text to again benefit from production in the present study, which it did. This observation, consistent over Experiments 1–4, provided both a replication of the memory benefit and a manipulation check for our key new question: Does the benefit of production extend to comprehension?

Examination of almost any standard comprehension test will readily reveal that many of the questions involved are heavily reliant on memory (e.g., 'literal' comprehension questions; Basaraba et al., 2013). Many of the questions asked on such tests—routinely in multiple-choice format—point to directly stated facts in the text, essentially constituting recognition tests for verbatim elements of the text. These memory questions should show a benefit of production—and they do. But what about the other questions, the ones that examine deeper knowledge, including the theme or tone of the text, its gist, or the inferences that can be derived? Memory no doubt still plays an underlying role—it is hard to imagine understanding what cannot be remembered—but these types of questions provide a less verbatim index of comprehension.

In this study, we found that reading aloud yielded significantly better scores than reading silently for memory-focused questions but not for comprehension-focused questions. This pattern is consistent with the distinctiveness account of the production advantage (see Gathercole & Conway, 1988; MacLeod, 2011; MacLeod et al., 2010; see MacLeod & Bodner, 2017, for a brief review). The production benefit based on a distinctiveness account is thought to be literal: It is the studied material itself standing out on the recognition test that is important. In tests of memory, study and test very closely match; they do not go beyond the literal and hence do not extend to similar or semantically related material. Therefore, a verbatim memory test for studied material should show a benefit due to distinctiveness, but a test that attempts to go beyond the verbatim text—to the concepts that it conveys—might not.

This study supports a distinctiveness account of the production effect but does not provide evidence for a strength-based production benefit. If reading aloud led to more robust memories for produced content (as a strength account would predict), presumably then passages read entirely aloud should be easier to freely recall and should therefore facilitate inferential thinking when choosing answers on a comprehension test. Indeed, other researchers investigating comprehension of text have argued that the ability to freely recall a passage is critical for supporting inferential thinking (Hua & Keenan, 2014). Therefore, that production failed to aid comprehension in this study could be taken to indicate that reading passages aloud during the encoding phase did not enhance later *recall* of studied passages relative to reading them silently. Only when the tested content closely matched that of the studied content, as for the memory questions (relying more so on *recognition* memory), could production aid performance—doing so via heightened distinctiveness.

A reader familiar with the educational psychology literature would know that there has been a richer history of research on reading aloud versus silently there. Much of the work in that realm has concerned so-called read-aloud strategies whereby teachers read aloud to children, showing that doing so can aid students' comprehension of text, even when students are learning English as a foreign language (for a recent review, see Senawati et al., 2021). In rarer cases, however, students themselves—rather than the teacher—are asked to read aloud, resulting in studies more analogous to ours. Returning to the cognitive psychology literature, MacLeod (2011) showed that the production benefit for memory can

[9] These analyses were also conducted with the preregistered target sample size of 129 participants by randomly selecting data from the full data set. The pattern of results was identical.

occur when listening to others read aloud, but that the size of the effect was significantly attenuated relative to when participants performed the oral reading task themselves. Therefore, it seems possible that the 'read-aloud' strategy employed in the educational literature could benefit students' memory for the produced content, even if teachers, not students, were to read aloud.

Hale et al. (2007) reported that reading aloud as opposed to silently aided what they referred to as 'comprehension' in school-aged children. But these researchers included half 'factual' (memory) and half 'inferential' (comprehension) questions in their tests and they did not analyze these two types of items separately. In essence, what Hale et al. found was a benefit of production collapsed across the two question types. It is possible, therefore, that the 'factual' questions were the only ones to benefit from reading aloud. If this were the case, the findings of Hale et al. would match our current pattern of results whereby a main effect of encoding condition (aloud vs. silent) is demonstrated, even though the effect is driven entirely by performance differences in memory-focused questions.

Matching our results, Salasoo (1986) failed to find any benefit of production on 'comprehension' test performance, despite also including 'literal' comprehension questions (i.e., memory questions) in the material set. Most revealing, perhaps, is the fact that Elbaum et al. (2004) found no benefit of reading aloud on comprehension performance in over 300 middle school and high school students, both with and without learning disabilities. In sum, related work in the field of education research offers evidence that converges with the findings reported here: Reading text aloud can aid learning, but probably only for literal information and not for transformed information, the hallmark of true comprehension.

A critic might point out that reading aloud ordinarily takes longer than reading silently (Brysbaert, 2019). Based on the reading time difference alone, a straightforward total time hypothesis (see Cooper & Pantle, 1967) would suggest that reading aloud results in better encoding than does reading silently simply because reading aloud takes longer, permitting more extensive encoding. But this seems unlikely to be the explanation. In their study using text, Ozubko et al. (2012) showed that there was no significant correlation between reading time and the size of the production effect. Moreover, MacLeod et al. (2010) showed in their earlier studies that the production effect remained robust even when silent words were purposely studied for longer than aloud words. In addition, the Sans Forgetica font condition included in Experiment 2 would certainly result in slower reading times, yet no improvement in test accuracy was found (see the Appendix).

An additional limitation of the studies presented here was that Experiments 2–4 were conducted remotely and unsupervised. Although prior work has shown that the production effect replicates when testing via video conferencing (Mama & Icht, 2020), few production studies have been conducted unsupervised. It is possible, therefore, that production benefited comprehension in Experiment 1 but not in Experiments 2–4 because in Experiment 1 participants read aloud with an experimenter in the room while in Experiments 2–4 no experimenter was present. Because people may find reading aloud in front of others a stressful task—indeed, the Trier Social Stress Test (Kirschbaum et al., 1993) protocol involves giving a speech in front of others—it is possible that this could have influenced performance across our experiments. Indeed, small stressors ('desirable difficulties'; for a review, see Bjork & Bjork, 2020) can enhance performance on cognitive tasks. Consequently, it is possible that production can benefit comprehension when participants must read aloud with others present, or more generally when they are placed in stressful situations. Nonetheless, that significant production benefits for memory were observed in Experiments 1–3—even with alterations in experiment settings—suggests that reading in front of others is not a requirement to gain the memory benefit.

Further exploration of the efficacy of distinctive encoding strategies for memory and for comprehension would test the generalizability of the claims made here. If future studies can confirm that other distinctiveness-based mnemonic techniques benefit memory but not comprehension of text, this would enhance confidence that (1) production conveys its benefit via distinctiveness, and (2) distinctiveness requires matching studied content to tested content. For example, bizarre imagery is also thought to aid memory via a distinctiveness heuristic (Einstein & McDaniel, 1987; Einstein et al., 1989; Worthen & Marshall, 1996). We would predict that a passage containing bizarre imagery would boost memory relative to a passage that was more mundane or congruent with contextual expectations, but that comprehension of the two passages should not differ. At the same time, mnemonic techniques thought to improve memory through strengthened encoding rather than distinctiveness—such as drawing pictures of to-be-remembered information (Roberts & Wammes, 2020; Wammes et al., 2016, 2018)—should benefit comprehension as well as memory. Indeed, recent studies have already shown this to be true (Schmeck et al., 2014; Wammes et al., 2017).

In conclusion, the distinctiveness account (see MacLeod & Bodner, 2017)—that production results in the encoding of additional features—remains the best general explanation of the production benefit. Indeed, recent formal modeling efforts (MINERVA2: Jamieson et al., 2016; REM: Kelly et al., 2022) have successfully used additional encoding to capture a range of the production effect results in the literature. The current study aimed to explore whether production would benefit reading comprehension for text, extending previous findings that production improves memory both for single words and for longer passages. Here, we repeatedly found that production did not benefit comprehension despite benefiting memory for the same texts. Given that the Nelson-Denny Reading Test and the Test Prep Review website were

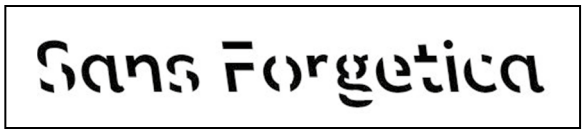

Fig. 5 An example of the Sans Forgetica font

created to measure comprehension skills, and given that the multiple-choice test format used in the current study is a test format widely used in classroom settings, it appears that reading aloud improves memory but not comprehension. It is the actual wording of texts, and not their underlying meaning, that production renders distinctive in memory.

Appendix

Experiment 2 also gauged the value of production relative to another encoding method—something rarely done in production effect studies to date. One encoding method that has received considerable research attention has involved presenting materials in a visually degraded format. Visual degradation has been hypothesized to motivate learners to increase their encoding effort—a "desirable difficulty" (Bjork, 1994) that can result in enhanced memory (e.g., Diemand-Yauman et al., 2011; for reviews, see Weissgerber & Reinhard, 2017; Xie et al., 2018).

In this vein, studies have used a variety of methods of visual degradation, including blurring (Yue et al., 2013), reduced font size (Kornell et al., 2011), and presenting words upside down (Sungkhasettee et al., 2011). A more recent implementation involves presenting materials in Sans Forgetica font, a free-download font designed with the goal of enhancing learning and memory (Fig. 5; see Earp, 2018). Despite this goal, published studies either have supported the effectiveness of this font only under limited circumstances (Eskenazi & Nix, 2021; Geller & Peterson, 2021) or they have not supported it at all (Cui & Liu, 2022; Cushing & Bodner, 2022; Geller et al., 2020; Taylor et al., 2020). Thus, Experiment 2 also provided a novel test of whether studying texts in Sans Forgetica font might improve memory—and possibly comprehension—for text. To this end, in Experiment 2, a (silent) Sans Forgetica encoding condition was included together with the aloud and silent conditions in a within-subject, between-passage design.

Results

We conducted analyses parallel to those on production in the main text to evaluate the effect of reading texts in Sans Forgetica font—a 2 (question type: Sans Forgetica, silent) × 2 (encoding condition: memory, comprehension) repeated-measures ANOVA. There was a significant main effect of question type, $F(1, 63) = 20.04$, $p < .001$, $\eta_p^2 = .24$, $BF_{10} = 52.41$, but not of encoding condition, $F(1, 63) = 0.43$, $p = .514$, $\eta_p^2 = .01$, $BF_{01} = 6.42$. The interaction was also nonsignificant, $F(1, 63) = 0.19$, $p = .664$, $\eta_p^2 < .01$, $BF_{01} = 4.76$. Put simply, performance was similar for texts read silently whether those texts were presented in a normal font ($M = 0.77$, $SD = 0.19$) or in Sans Forgetica ($M = 0.76$, $SD = 0.19$).

Participants' mean predicted accuracy was 64.12% ($SD = 24.70$) in the silent condition and 52.41% ($SD = 26.71$) in the Sans Forgetica condition. Participants thought that their accuracy was lower for the Sans Forgetica texts than for the silent texts in normal font, $t(65) = 5.16$, $SE = 2.27$, $p < .001$, $d = 0.46$, even though they did not show reliably lower accuracy—thus demonstrating a mismatch between their metacognition and their performance.

Discussion

Experiment 2 demonstrated that another simple encoding "hack," reading texts presented in Sans Forgetica font, did not improve test accuracy, in line with other findings in the literature (Cui & Liu, 2022; Cushing & Bodner, 2022; Geller et al., 2020; Taylor et al., 2020).

Acknowledgements Experiments 1, 3, and 4 were supported by a Natural Sciences and Engineering Research Council (NSERC) of Canada Discovery Grant A7459 to CMM and by an NSERC Postgraduate Scholarship to BRTR. An early version of this work was presented at the annual meeting of the Canadian Society for Brain, Behaviour, and Cognitive Science in 2022. We thank Ryan Yeung for assistance in writing the custom script for text analysis of the two material sets used in this study. The data, analysis code, experimental programs, pre-registrations, and other materials are listed on the Open Science Framework (OSF; https://tinyurl.com/PE-and-Comp-Peer-Review).

Declarations

Conflict of Interest The authors express no conflict of interest.

References

Addpipe Development Team. (2022). *Addpipe* (Version July 2022) [Computer Software]. https://addpipe.com/

Alonzo, J., Basaraba, D., Tindal, G., & Carriveau, R. S. (2009). They read, but how well do they understand?: An empirical look at the nuances of measuring reading comprehension. *Assessment for Effective Intervention, 35*(1), 34–44.

Barlow, M. C. (1928). The role of articulation in memorizing. *Journal of Experimental Psychology, 11*, 306–312.

Basaraba, D., Yovanoff, P., Alonzo, J., & Tindal, G. (2013). Examining the structure of reading comprehension: Do literal, inferential, and evaluative comprehension truly exist? *Reading and Writing, 26*(3), 349–379.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition, 9*(4), 475–479. https://doi.org/10.1016/j.jarmac.2020.09.003

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning & Verbal Behavior, 11*, 717–726.

Brown, J. A., Fishco, V. V., & Hanna, G. (1993). *Nelson-Denny reading test: Manual for scoring and interpretation*. Riverside Publishing.

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language, 109*, 104047.

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*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

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

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods, 51*, 467–479. https://doi.org/10.3758/s13428-018-1077-9

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review, 24*, 369–378.

Castel, A. D., Rhodes, M. G., & Friedman, M. C. (2013). Predicting memory benefits in the production effect: The use and misuse of self-generated distinctive cues when making judgments of learning. *Memory & Cognition, 41*, 28–35.

Chi, M. T., De Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439–477.

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

Cooper, E. H., & Pantle, A. J. (1967). The total-time hypothesis in verbal learning. *Psychological Bulletin, 68*, 221–234.

Cui, L., & Liu, J. (2022). Recognition of studied words in perceptual disfluent Sans Forgetica font. *Vision, 6*, 52. https://doi.org/10.3390/vision6030052

Cushing, C., & Bodner, G. E. (2022). Reading aloud improves proofreading (but using Sans Forgetica font does not). *Journal of Applied Research in Memory and Cognition, 11*, 427–436.

Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2011). Fortune favors the bold (and the italicized): Effects of disfluency on educational outcomes. *Cognition, 118*, 111–115.

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.

Earp, J. (2018). Q&A: Designing a font to help students remember key information. *Teacher Magazine*. Australian Council for Educational Research.

Einstein, G. O., & McDaniel, M. A. (1987). Distinctiveness and the mnemonic benefits of bizarre imagery. In M. A. McDaniel & M. Pressley (Eds.), *Imagery and related mnemonic processes*. Springer. https://doi.org/10.1007/978-1-4612-4676-3_4

Einstein, G. O., McDaniel, M. A., & Lackey, S. (1989). Bizarre imagery, interference, and distinctiveness. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(1), 137–146. https://doi.org/10.1037/0278-7393.15.1.137

Elbaum, B., Arguelles, M. E., Campbell, Y., & Saleh, M. B. (2004). Effects of a student-reads-aloud accommodation on the performance of students with and without learning disabilities on a test of reading comprehension. *Exceptionality, 12*(2), 71–87.

Eskenazi, M. A., & Nix, B. (2021). Individual differences in the desirable difficulty effect during lexical acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*, 45–52.

Faul, F., Erdfelder, E., Lang, E.-G., & Buchner, A. (2007). G*Power 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-lists designs: A meta-analysis. *Acta Psychologica, 142*, 1–5.

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.

Festinger, L. (1953). Laboratory experiments. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences* (pp. 136–172). Holt, Rinehart & Winston.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6*, 1–104.

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

Geller, J., & Peterson, D. (2021). Is this going to be on the test? Testing expectancy moderates the Sans Forgetica effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*, 1924–1938.

Geller, J., Davis, S. D., & Peterson, D. (2020). Sans Forgetica is not desirable for learning. *Memory, 28*(8), 957–967.

Hale, A. D., Skinner, C. H., Williams, J., Hawkins, R., Neddenriep, C. E., & Dizer, J. (2007). Comparing comprehension following silent and aloud reading across elementary and secondary students: Implication for curriculum-based measurement. *The Behavior Analyst Today, 8*(1), 9–23.

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

Hua, A. N., & Keenan, J. M. (2014). The role of text memory in inferencing and in comprehension deficits. *Scientific Studies of Reading, 18*(6), 415–431. https://doi.org/10.1080/10888438.2014.926906

Icht, M., Mama, Y., & Algom, D. (2014). The production effect in memory: Multiple species of distinctiveness. *Frontiers in Psychology, 5*, 886. https://doi.org/10.3389/fpsyg.2014.00886

Icht, M., Taitelbaum-Swead, R., & Mama, Y. (2022). Production improves visual and auditory text memory in younger and older adults. *Gerontology, 68*(5), 578–586.

Jamieson, R. K., Mewhort, D. J. K., & Hockley, W. E. (2016). A computational account of the production effect: Still playing twenty questions with nature. *Canadian Journal of Experimental Psychology, 70*, 154–164.

Kelly, M. O., Ensor, T. M., Lu, X., MacLeod, C. M., & Risko, E. F. (2022). Reducing retrieval time modulates the production effect: Empirical evidence and computational accounts. *Journal of Memory and Language, 123*, 104299.

Kirschbaum, C., Pirke, K.-M., & Hellhammer, D. H. (1993). The trier social stress test: A tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology, 28*(1/2), 76–81.

Kline, C. S. (2019). Production effect in complex texts and over time. *Journal of Integrative Behavioral Science, 1*, 1–7.

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease of processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science, 22*, 787–794.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*, 978–990. https://doi.org/10.3758/s13428-012-0210-4

Leff, A. P., Schofield, T. M., Crinion, J. T., Seghier, M. L., Grogan, A., Green, D. W., & Price, C. J. (2009). The left superior temporal gyrus is a shared substrate for auditory short-term memory

and speech comprehension: Evidence from 210 patients with stroke. *Brain, 132*(12), 3401–3410. https://doi.org/10.1093/brain/awp273

Lenth, R. V., Buerkner, P., Herve, M., Jung, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2022). *emmeans: Estimated marginal means, aka least-squares means* (Version 1.8.1-1) [Computer software]. https://github.com/rvlenth/emmeans

MacLeod, C. M. (2011). I said, you said: The production effect gets personal. *Psychonomic Bulletin & Review, 18*(6), 1197–1202. https://doi.org/10.3758/s13423-011-0168-8

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.

Mama, Y., & Icht, M. (2020). Overcoming COVID-19 challenges: A remote adaptation of the production effect task. *Journal of the International Neuropsychological Society, 27*(8), 855–856. https://doi.org/10.1017/s1355617720001277

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2011). *BayesFactor: Computation of Bayes factors for common design* (Version 0.9.12-4.4). [Computer software]. https://richarddmorey.github.io/BayesFactor/

Nelson, M. S., & Denny, E. C. (1929). *The Nelson-Denny reading test*. Houghton Mifflin.

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.

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

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.

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods, 54*(4), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

Psychology Software Tools. E-Prime 3.0 [Computer software]. (2016). Retrieved from https://support.pstnet.com/

R Core Team. (2020). *R: A language and environment for statistical computing* (Version 4.1.1) [Computer software]. R Foundation for Statistical Computing. http://www.r-project.org/

Rinker, T. W. (2021). *sentimentr: Calculate text polarity sentiment* (Version 2.9.1) [Computer software]. https://github.com/trinker/sentimentr

Roberts, B. R. T., & Wammes, J. D. (2020). Drawing and memory: Using visual production to alleviate concreteness effects. *Psychonomic Bulletin & Review, 28*(1), 259–267. https://doi.org/10.3758/s13423-020-01804-w

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23*(4), 441–474.

Salasoo, A. (1986). Cognitive processing in oral and Silent reading comprehension. *Reading Research Quarterly, 21*(1), 59–69.

Schmeck, A., Mayer, R. E., Opfermann, M., Pfeiffer, V., & Leutner, D. (2014). Drawing pictures during learning from scientific text: Testing the generative drawing effect and the prognostic drawing effect. *Contemporary Educational Psychology, 39*(4), 275–286. https://doi.org/10.1016/j.cedpsych.2014.07.003

Senawati, J., Suwastini, N. K., Jayantini, I. G., Adnyani, N. L., & Artini, N. N. (2021). The benefits of reading aloud for children: A review in EFL context. *IJEE (Indonesian Journal of English Education), 1*(1), 73–100.

Singmann, H., Bolker, B., Westfall, J., Aust, F., Ben-Shachar, M. B., Højsgaard, S., Fox, J., Lawrence, M. A., Mertens, U., Love, J., Length, R., & Christensen, R. H. B. (2022). *afex: Analysis of factorial experiments* (Version 1.1-1). [Computer software]. https://cran.r-project.org/web/packages/afex/index.html

Sungkhasettee, V. W., Friedman, M. C., & Castel, A. D. (2011). Memory and metamemory for inverted words: Illusions of competency and desirable difficulties. *Psychonomic Bulletin & Review, 18*, 973–978.

Taylor, A., Sanson, M., Burnell, R., Wade, K. A., & Garry, M. (2020). Disfluent difficulties are not desirable difficulties: The (lack of) effect of Sans Forgetica on memory. *Memory, 28*(7), 850–857.

Todorovic, D. (2020). *Choosing what to read aloud while studying: The role of agency in production*. Unpublished dissertation, University of Waterloo.

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*(9), 1752–1776. https://doi.org/10.1080/17470218.2015.1094494

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. https://doi.org/10.1016/j.actpsy.2017.07.008

Wammes, J. D., Roberts, B. R., & Fernandes, M. A. (2018). Task preparation as a mnemonic: The benefits of drawing (and not drawing). *Psychonomic Bulletin & Review, 25*(6), 2365–2372. https://doi.org/10.3758/s13423-018-1477-y

Weissgerber, S. C., & Reinhard, M.-A. (2017). Is disfluency desirable for learning? *Learning and Instruction, 49*, 199–217.

Wijffels, J. (2023). *udpipe: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the UDPipe NLP toolkit* (Version 0.8.11) [Computer Software]. https://CRAN.R-project.org/package=udpipe

Worthen, J. B., & Marshall, P. H. (1996). Intralist and extralist sources of distinctiveness and the bizarreness effect: The importance of contrast. *The American Journal of Psychology, 109*(2), 239. https://doi.org/10.2307/1423275

Xie, H., Zhou, Z., & Liu, Q. (2018). Null effects of perceptual disfluency on learning outcomes in a text-based educational context: A meta-analysis. *Educational Psychology Review, 30*, 745–771.

Yeung, R. C. (2023). Lex-lookup (Version 0.1.0) [Computer software]. https://doi.org/10.5281/zenodo.7730607

Yue, C., Castel, L., & Bjork, A. (2013). When disfluency is—and is not—a desirable difficulty: The influence of typeface clarity on metacognitive judgments and memory. *Memory & Cognition, 41*, 229–241.

Open practices statement All experimental materials, programs, data, and statistical code for these experiments are available on the Open Science Framework (OSF; https://tinyurl.com/PE-and-Comp-Peer-Review). Experiments 2–4 were preregistered on OSF.