Received: 3 November 2023 | Revised: 14 June 2024 | Accepted: 12 July 2024

DOI: 10.1002/acp.4230

REGISTERED REPORT STAGE 2

WILEY

Does generation benefit learning for narrative and expository texts? A direct replication attempt

Julia Schindler[1] | Tobias Richter[1] | Raymond A. Mar[2]

[1]Department of Psychology IV, University of Würzburg, Wurzburg, Germany

[2]Department of Psychology, York University, Toronto, Canada

Correspondence
Julia Schindler, Department of Education and Psychology, Freie Universität Berlin, Special Educational Needs, Fabeckstraße 35, 14195 Berlin, Germany.
Email: julia.schindler@fu-berlin.de

Present address
Julia Schindler, Department of Education and Psychology, Freie Universität Berlin, Special Needs Education, Berlin, Germany.

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: FOR 5254/1

Abstract

Generated information is better recognized and recalled than information that is read. This generation effect has been replicated several times for different types of material, including texts. Perhaps the most influential demonstration is by McDaniel, Einstein, Dunay, and Cobb (*Journal of Memory and Language*, 1986, 25(6), 645–656; henceforth MEDC). This group tested whether the generation effect occurs only if the generation task stimulates cognitive processes not already stimulated by the text. Numerous studies, however, report difficulties replicating this text by generation-task interaction, which suggests that the effect might only be found under conditions closer to the original method of MEDC. To test this assumption, the present study closely replicated MEDC's Experiment 2 in two separate German and English-speaking samples. The present study provided partial evidence in favor of the expected interaction, which ultimately depended on successful completion of the generation task (with near-to-perfect accuracy). Moreover, it indicates that sentence unscrambling might enhance learning across genres.

KEYWORDS
expository texts, generation effect, learning, narrative texts, replication

1 | INTRODUCTION

Generated information is often better recognized and recalled than information that is passively encoded. This phenomenon is called the *generation effect* and it was first reported for the learning of word pairs (e.g., McDaniel et al., 1988; Slamecka & Graf, 1978). The classical paradigm consists of two experimental conditions in which learners are presented with two words: (1) a context word (e.g., WINTER), and (2) a related target word (e.g., SNOW) which they should memorize for a learning test. In the generation condition, the fragmented target word (e.g., S_ _ W) is presented, and the learner must generate the target word (SNOW). In the reading condition, the same word is presented intact. For both conditions, participants must write down the target words during the learning phase. Typically, learners recognize and recall the generated target words better than the target words they merely read. This effect has been replicated several times for different types of stimulus material, including numbers (e.g., Gardiner & Rowley, 1984), words (e.g., McDaniel et al., 1988; Slamecka & Graf, 1978), sentences (e.g., Graf, 1980, 1981; Lutz et al., 2003), and even rich textual information such as song lyrics (Goldman & Kelley, 2009) and recipes (Goverover et al., 2008, 2010, 2013, 2014). The effect has also been demonstrated using a diverse range of generation tasks, including fill-in-the-blanks (Abel & Hänze, 2019), letter completion (e.g., Einstein et al., 1984; McDaniel & Kerwin, 1987), unscrambling (e.g., Graf, 1982; McDaniel et al., 1986), and mental rotation (e.g., Graf, 1982). Moreover, the memory benefit has also emerged for a wide range of measures, including recognition, cued and free recall (e.g., McDaniel et al., 1988; Slamecka & Graf, 1978),

and for cloze tasks (DeWinstanley & Bjork, 2004). Finally, the generation effect has been demonstrated with different retention intervals and for between- and within-subject designs (see the meta-analysis by Bertsch et al., 2007).

1.1 | The generation effect for narrative and expository texts

One form of the generation effect, which is particularly relevant to educational contexts, is improving memory for complex narrative and expository texts (Einstein et al., 1990; McDaniel et al., 1986; McDaniel et al., 1994; McDaniel et al., 2002). Text generation comprises activities for creating the learning material—or at least parts of it—instead of being presented with an intact text. Two generation paradigms in particular that have been empirically shown to be beneficial for learning with texts are letter completion (filling in letters deleted throughout the text) and sentence unscrambling (arranging randomly-ordered sentences into a meaningful order) (e.g., Einstein et al., 1984, 1990; McDaniel et al., 1986, 2002).

However, in contrast to word learning studies, generation seems to enhance learning for full texts less consistently. Some studies, for example, have found no generation effect for narrative texts (Einstein et al., 1990, Exp. 1; McDaniel et al., 1986, Exp. 1). Others have failed to replicate the effect for learners with good comprehension skills (Abel & Hänze, 2019; McDaniel et al., 2002, Exp. 2A) or for expository texts (Maki et al., 1990, Exp. 1; Schindler et al., 2017; Thomas & McDaniel, 2007, Exp. 1).

These mixed findings suggest that text generation fails to provide a consistent advantage over reading for all learners across different conditions. Generation can be understood as an example of *desirable difficulties*, which are educational measures that make learning intentionally more difficult to improve outcomes (Bjork, 1994; Bjork & Bjork, 2011). One contextual framework advanced by McDaniel and Butler (2010), describes the outcomes of desirable difficulties as a complex interaction of learner characteristics, type of test, learning materials, and tasks (see also Einstein et al., 1990; McDaniel & Einstein, 1989, 2005). According to this framework, learning can be improved only when difficulties stimulate unique cognitive processes that are not already elicited by the learners and when the test requirements match the processes stimulated by the generation task (Schindler et al., 2019).

The material by processing-task interaction proposed by this framework may also explain a phenomenon observed for narrative and expository texts. Learning with narrative texts can be enhanced by letter completion, and learning with expository texts can be enhanced by unscrambling sentences (Einstein et al., 1990; McDaniel et al., 1986, 2002). However, the inverse appears to not to have an effect. That is, unscrambling does not benefit learning from narratives and letter completion does not benefit learning from expository texts (Einstein et al., 1990; McDaniel et al., 1986). A potential explanation for this divergence was first proposed in the *Material Appropriate Processing* (MAP) framework (McDaniel et al., 1986; McDaniel & Einstein, 1989), now one of the components of the contextual framework proposed by McDaniel and Butler (2010). According to McDaniel and colleagues, narrative and expository texts have qualitatively different encoding demands, which interact with the type of generation task. A learning benefit can only be observed when the generation task is appropriate for the learning material such that it stimulates cognitive processing that was not already elicited by the material content.

Narratives typically possess the regular and familiar structure of a story schema (Rumelhart, 1975). This story schema stimulates relational processing of the narrative's propositions, aiding organization and integration (McDaniel et al., 1986). Generating texts by unscrambling sentences also stimulates relational processing because rearranging the sentences into a meaningful order requires organization and integration of the sentence propositions. According to the MAP framework, unscrambling has no effect on learning from a narrative because unscrambling elicits a process already present during narrative comprehension as opposed to eliciting a novel cognitive process. Letter completion, in contrast, stimulates individual-item or proposition-specific processing, that is, processing of lexical concepts or relations between the concepts of a proposition (McDaniel et al., 1986). This process is not present during narrative comprehension. Thus, the letter completion task stimulates unique cognitive processes and thereby enhances learning.

Expository texts, in contrast to narrative texts, stimulate individual-item or proposition-specific processing. This type of text directly focuses on the comprehension of new or unfamiliar concepts instead of stimulating organizational and integrative processing between propositions (Einstein et al., 1990; McDaniel et al., 1986). According to the MAP framework, learning with expository texts is improved by unscrambling because this task stimulates cognitive processes not already elicited by the expository texts. Letter completion, however, has no effect on learning with expository texts because this form of generation task stimulates individual-item or proposition-specific processing, which is already elicited by simply reading the expository text.

1.2 | Inconsistent findings

Despite this appealing explanation, a considerable number of studies report inconsistent findings regarding this genre by generation-type interaction. Letter completion, for example, had no effect on recall for narratives in some studies (Einstein et al., 1990, Exp. 1; McDaniel et al., 1986, Exp. 1; McDaniel et al., 2002; Exp. 2A), and benefited recall for expository texts in other studies (Bjork & Storm, 2011, Exp. 1–4; Burnett & Bodner, 2014; Exp. 1 and 2; DeWinstanley & Bjork, 2004, Exp. 1A-3; Maki et al., 1990, pilot study and Exp. 1), both findings of which were unexpected according to the MAP framework. Similarly, sentence unscrambling had no effect on recall for expository texts in some studies (McDaniel et al., 2002, Exp. 1B; Thomas & McDaniel, 2007, Exp. 1). Also, in some studies, sentence unscrambling unexpectedly benefitted learning from narratives (Einstein

et al., 1990, Exp. 2; McDaniel et al., 1994, Exp. 1–3). Moreover, Schindler and Richter (in preparation) ran a series of seven experiments on this topic under ecologically-valid and methodologically-stringent conditions. In all experiments, expository texts were read (reading control condition) or sentences had to be unscrambled (generation condition). The generation effect was found in only one of the experiments. This particular experiment most closely resembled the original studies by McDaniel and colleagues (e.g., McDaniel et al., 1986, 2002) because learners were not informed of the subsequent test and were allowed to take as much time as needed to read or generate the texts. These findings suggest that the generation effect, though difficult to replicate under ecologically-valid and methodologically-stringent conditions, might be replicable under conditions closer to the setting of the original studies.

These conflicting findings suggest that the text-generation effect is either unreliable (and thus not useful for educational contexts) or it is moderated by contextual factors (McDaniel & Butler, 2010). However, a replication of the original interaction effect is an important and necessary precondition before investigating potential moderating contextual factors. A close replication of the effect under these limited conditions will open the door to methods that control for contextual factors under which the text generation effect might emerge. Thus, this replication is necessary before text generation should be utilized as a learning intervention in pedagogical contexts.

1.3 | The present study

The aim of the present study was to replicate the genre by generation-task interaction found by McDaniel et al. (1986, MEDC). Their study was the first to propose, test, and support the idea that different generation tasks (letter completion vs. sentence unscrambling) work differently in combination with different types of texts (narratives vs. expository). The present study attempted to replicate Experiment 2 of MEDC because the findings provided more convincing evidence for the framework compared with Experiment 1. Our study employed an English-speaking sample to keep the replication as close as possible to the original method, and also investigated a German-speaking sample (using translated materials) to examine whether the generation effect generalizes to another language.

Participants read or generated a short narrative ("The Just Reward" by Guterman, 1945) or an expository text passage ("The Frozen Country," modified from Levy, 1981, by MEDC). Participants in the generation conditions either filled in missing letters or reordered scrambled sentences. Subsequently, in an unannounced learning test, they were asked to write down as much information from the texts as they could remember. We were in close contact with the first author of MEDC to ensure that our method matched as closely as possible to the original study method (i.e., material, study design, instructions, and analyses). We expected to find the generation effect, showing a higher proportion of free recall for narrative texts when missing letters are completed (compared to the reading control condition) and for expository texts when scrambled sentences are reordered (compared to the reading control condition). No generation effect or a substantially smaller effect was predicted for unscrambling narratives and for completing expository texts.

2 | METHOD

2.1 | Ethics statement

The study was conducted in full accordance with the Ethical Guidelines of the German Psychological Society (DGPs), the Canadian Tri-Council Research Ethics guidelines and the American Psychological Association (APA), and it has been approved by the local ethics committees.

2.2 | Participants

As in the original study, participants were enrolled in introductory psychology classes (psychology and teaching students) and received extra course credit for their participation. In the original study, a total of 72 students were randomly assigned to six experimental groups in a 2 (narrative vs. expository text) × 3 (letter completion vs. sentence unscrambling vs. reading control) between-subjects design, resulting in 12 participants per group.

In a meta-analysis of the generation effect, Bertsch et al. (2007) found that the effect was about half the size in between-subjects designs ($d = 0.28$) compared with within-subjects designs ($d = 0.50$). However, this meta-analysis included no studies that used full text generation. Thus, whether these findings can be generalized to text generation is still an open question. To date, the text-generation effect has been demonstrated with both between-subjects designs (e.g., Abel & Hänze, 2019; Einstein et al., 1984; McDaniel & Einstein, 1989) and within-subjects designs (Bjork & Storm, 2011; Goldman & Kelley, 2009; McDaniel & Einstein, 1989).

Using a between-subjects design, MEDC reported main effects and interactions of considerable size. As predicted, the authors reported a large effect showing that participants who completed letters for a narrative recalled more information compared with those who read it, $F(1, 66) = 63.77$ (equivalent to $\eta_p^2 = 0.49$). Participants who unscrambled sentences for an expository text also recalled more information compared with those reading it, $F(1, 66) = 30.57$ ($\eta_p^2 = 0.32$). Lastly, for the overall interaction, letter completion led to the best recall with the narrative text, and unscrambling with the expository text, $F(2, 66) = 33.57$ ($\eta_p^2 = 0.50$).

Despite these large effects in the original study, we entered a medium-sized effect ($\eta^2 = 0.06$; Cohen, 1988) in our power analysis using G*Power (Faul et al., 2007), which revealed a required sample size of $N = 251$, with power (1-β) set to 0.95 and an α-level of 0.05. We increased the target sample size to $N = 300$ (for both the English and German versions, a total of 600) to account for the substitution of participants who would report that they have expected the learning test with participants who would report that they had not, as in the MEDC study.

A total of 300 German students (249 female, 6 missing data; 289 native speakers, 6 missing data) with a mean age of 21.12 years (SD = 3.38, Min = 18, Max = 52) participated at the Universities of Kassel (n = 112) and Würzburg (n = 188). The English-speaking sample comprised 312 participants (219 female, 3 gender diverse, 7 missing data; 227 native speakers, 8 missing data) with a mean age of 19.72 years (SD = 4.96, Min = 17, Max = 59). Participants were recruited from a large urban university in Canada, and tested in small groups of one to four. All provided written consent before testing.

2.3 | Materials and procedure

The materials and procedure in the present study were as described in MEDC. Half of the participants in the English-speaking sample were presented with the English version of the Russian narrative, "The Just Reward" (Guterman, 1945), and the other half with the expository text "The Frozen Continent" (modified from Levy, 1981, by MEDC). Both texts were presented with titles. Each text contained 20 sentences with 83 idea units in the narrative text and 69 idea units in the expository text. For the German-speaking sample, both texts were translated to German. Texts were translated by a speaker with native fluency in both languages, and the quality of translation was assessed with back-translation using a second translator. In the letter completion condition, 18% of the letters were randomly deleted and replaced with blanks of which 40% were vowels. In the sentence unscrambling condition, participants were randomly assigned to one of two conditions, each with text consisting of 20 sentences randomly ordered.

Participants were not informed of the learning test. Instead, they were told that the aim of the study was to investigate their text comprehension. Processing time to read or generate the texts was recorded, and time on task was not limited. Participants were then provided a comprehensibility rating for the text on a 5-point Likert scale (1 = *didn't comprehend the passage at all*, 5 = *comprehended the passage very well*). After a distraction task (i.e., working on math problems for 5 min), the memory test was administered. Participants were asked to write down as much information about the text that they could remember in as much time as they needed.

2.3.1 | Post-experimental questionnaire

In a post-experimental questionnaire, participants were asked to indicate whether they had expected the memory test or not. In addition (and as an extension of the original study), demographics such as gender, age, first language, field of study (psychology or teaching), prior knowledge of the expository text content, and familiarity with the narrative were assessed. These variables were analyzed to check for comparability of the experimental groups. In case of significant differences between groups, these variables were supposed to be statistically controlled in additional analyses.

3 | ANALYSES AND RESULTS

Separate analyses were conducted for the German and the English-speaking samples. As in the original study by MEDC, participants who reported that they had expected the memory test were excluded from analyses. After this exclusion, the analysis samples contained 250 for the English-speaking participants (6 missing data) and 197 for the German-speaking participants (6 missing data). However, because the scores on the memory test did not differ greatly between excluded and included participants, we ran additional analyses with the whole sample, English sample: $t(304) = 0.90$, $p = .37$, $d = 0.13$; German sample: $t(292) = -0.45$, $p = .65$, $d = -0.06$. All statistical significance tests were based on a Type-I error rate of .05. Statistically significant results were subjected to a Bonferroni–Holm correction for multiple testing.

For the sake of comprehensibility and conciseness, we focus on the analyses that correspond to those originally reported by MEDC in the following sections. This entails the English learners who were naïve regarding the learning test. Following the structure of MEDC, we first report findings on generation accuracy (Generation Accuracy Section), followed by processing time (Processing Time Section) and comprehensibility (Comprehensibility Section). The hypothesis tests with free recall as dependent variable, which are central to this replication study, are reported last (Free Recall Section). Results for the whole sample (including those who anticipated the test) are only reported if they deviated from the results of the subset analyses.

The analyses for the German learners are reported in Sections 3.4.2 and 3.4.3, where the central findings are reported, and if they deviated from the English sample. Detailed and complete results for both samples can be found in the supplemental material (Tables S1–S5) available on the Open Science Framework (https://osf.io/wftbr/?view_only=bd602fffe4874d16a64e162f7e81343e).

3.1 | Generation accuracy

3.1.1 | Letter completion

As in the original study, the mean proportions of letters that were generated correctly in the letter completion condition were computed for both texts. Overall, letter completion accuracy was high, with participants correctly inserting letters 93% of the time on average (descriptive statistics appear in Table 1). The difference in letter completion accuracy between genres was not statistically significant, $t(43.40) = 1.90$, $p = .064$, $d = 0.47$.

3.1.2 | Sentence unscrambling

For the sentence unscrambling task, the deviation of sentences from their original position (deviation score) was significantly larger for the expository text than for the narrative, $t(109.68) = 6.19$, $p < .001$, $d = 1.08$. This indicates that unscrambling the expository texts was more difficult than unscrambling the narrative. Participants also made genuine and successful

TABLE 1 Descriptive Statistics for generation accuracy in the letter completion group and in both sentence unscrambling groups (individually and collapsed).

	English sample											
	Test not expected						Total sample					
	Sample size			Generation accuracy			Sample size			Generation accuracy		
	n_{expos}	n_{nar}	n_{total}	Expository *M* (*SD*)	Narrative *M* (*SD*)	Total *M* (*SD*)	n_{expos}	n_{nar}	n_{total}	Expository *M* (*SD*)	Narrative *M* (*SD*)	Total *M* (*SD*)
Letter completion[a]	35	33	68	0.95 (0.05)	0.91 (0.10)	0.93 (0.08)	40	39	79	0.95 (0.04)	0.88 (0.18)	0.92 (0.13)
Sentence unscrambling V1[b]	33	32	65	2.66 (1.48)	0.92 (0.75)	1.80 (1.46)	39	39	78	2.72 (1.43)	0.99 (0.95)	1.86 (1.49)
Sentence unscrambling V2[c]	33	31	64	2.11 (1.44)	1.18 (1.07)	1.66 (1.35)	39	39	78	2.07 (1.39)	1.36 (1.34)	1.71 (1.40)
Sentence unscrambling collapsed[d]	66	63	129	2.38 (1.48)	1.05 (0.92)	1.73 (1.40)	78	78	156	2.40 (1.44)	1.18 (1.17)	1.79 (1.44)
	German sample											
	Test not expected						Total sample					
	Sample size			Generation accuracy			Sample size			Generation accuracy		
	n_{expos}	n_{nar}	n_{total}	Expository *M* (*SD*)	Narrative *M* (*SD*)	Total *M* (*SD*)	n_{expos}	n_{nar}	n_{total}	Expository *M* (*SD*)	Narrative *M* (*SD*)	Total *M* (*SD*)
Letter completion[a]	27	26	53	0.98 (0.02)	0.95 (0.04)	0.97 (0.03)	39	39	78	0.98 (0.02)	0.96 (0.03)	0.97 (0.03)
Sentence unscrambling V1[b]	31	23	54	1.16 (0.89)	0.51 (0.61)	0.89 (0.84)	36	36	72	1.13 (0.87)	0.45 (0.54)	0.79 (0.80)
Sentence unscrambling V2[c]	29	24	53	1.44 (0.83)	0.34 (0.45)	0.94 (0.87)	36	36	72	1.45 (1.01)	0.36 (0.43)	0.90 (0.94)
Sentence unscrambling collapsed[d]	60	47	107	1.30 (0.87)	0.43 (0.53)	0.91 (0.85)	72	72	144	1.29 (0.95)	0.40 (0.49)	0.85 (0.87)

Note: n_{expos} = sample size expository text group, n_{nar} = sample size narrative text group.

[a]Proportions of correctly generated letters.

[b]Mean deviation of the sentences from their original position in the unscrambled text Version 1.

[c]Mean deviation of the sentences from their original position in the unscrambled text Version 2.

[d]Mean deviation of the sentences from their original position in the unscrambled text collapsed across Versions 1 and 2.

TABLE 2 Descriptive statistics for processing time, comprehensibility, and recall as a function of genre and learning condition.

	English sample					
	Test not expected			Test expected		
	Processing Time[a] *M* (SD)	Comprehensibility[b] *M* (SD)	Recall[c] *M* (SD)	Processing time[a] *M* (SD)	Comprehensibility[b] *M* (SD)	Recall[c] *M* (SD)
Expository						
Read	186.44 (137.02)	4.31 (0.84)	0.16 (0.12)	183.41 (121.21)	4.24 (0.86)	0.17 (0.14)
Letter completion	1018.21 (323.73)	3.88 (0.73)	0.12 (0.08)	1030.87 (336.92)	3.90 (0.75)	0.13 (0.09)
Sentence unscrambling	1811.06 (821.89)	3.75 (0.62)	0.23 (0.14)	1781.73 (769.14)	3.75 (0.66)	0.23 (0.14)
Narrative						
Read	155.76 (49.75)	4.12 (0.86)	0.48 (0.15)	166.22 (70.84)	4.05 (0.92)	0.44 (0.19)
Letter completion	1284.94 (543.51)	3.86 (0.80)	0.46 (0.24)	1274.03 (531.81)	3.86 (0.79)	0.45 (0.24)
Sentence unscrambling	1660.66 (599.92)	3.89 (0.68)	0.55 (0.21)	1634.21 (570.56)	3.91 (0.70)	0.52 (0.21)
Total						
Read	171.69 (104.76)	4.21 (0.85)	0.32 (0.21)	174.81 (98.97)	4.14 (0.89)	0.30 (0.21)
Letter completion	1147.53 (460.58)	3.87 (0.76)	0.28 (0.25)	1150.87 (457.62)	3.88 (0.77)	0.29 (0.24)
Sentence unscrambling	1737.64 (723.22)	3.82 (0.65)	0.38 (0.24)	1707.97 (678.99)	3.83 (0.68)	0.38 (0.23)
	German sample					
	Test not expected			Total sample		
	Processing time[a] *M* (SD)	Comprehensibility[b] *M* (SD)	Recall[c] *M* (SD)	Processing time[a] *M* (SD)	Comprehensibility[b] *M* (SD)	Recall[c] *M* (SD)
Expository						
Read	167.06 (64.62)	4.44 (0.70)	0.22 (0.14)	184.00 (76.65)	4.31 (0.73)	0.25 (0.14)
Letter completion	834.26 (216.68)	4.19 (0.75)	0.22 (0.12)	796.18 (215.79)	4.13 (0.78)	0.22 (0.11)
Sentence unscrambling	1254.43 (364.17)	3.97 (0.83)	0.34 (0.13)	1265.04 (354.03)	4.00 (0.81)	0.35 (0.14)
Narrative						
Read	144.42 (56.51)	4.58 (0.61)	0.45 (0.12)	147.44 (49.20)	4.56 (0.60)	0.43 (0.10)
Letter completion	976.92 (218.15)	3.85 (0.92)	0.45 (0.16)	977.31 (242.31)	3.95 (0.87)	0.45 (0.14)
Sentence unscrambling	1135.72 (290.57)	4.00 (0.69)	0.52 (0.18)	1134.32 (293.76)	4.04 (0.70)	0.51 (0.17)
Total						
Read	155.43 (60.83)	4.51 (0.65)	0.34 (0.17)	165.72 (66.58)	4.44 (0.68)	0.34 (0.15)
Letter completion	904.25 (227.02)	4.02 (0.85)	0.33 (0.18)	886.74 (245.49)	4.04 (0.82)	0.34 (0.17)
Sentence unscrambling	1202.29 (337.58)	3.98 (0.77)	0.42 (0.18)	1199.68 (330.72)	4.02 (0.75)	0.43 (0.18)

[a]Processing time in seconds.
[b]Comprehensibility ratings on a 5-point Likert scale (1 = did not comprehend the passage at all, 5 = comprehended the passage very well).
[c]Proportional recall of idea units.

efforts to unscramble the sentences, with their final product deviating less from the original (unscrambled) text than the scrambled versions initially presented to them. This was true for both scrambled versions, and both genres, expository text Version 1: $t(32) = -14.11$, $p < .001$, $d = -2.46$; expository text Version 2: $t(32) = -18.25$, $p < .001$, $d = -3.18$; narrative text Version 1: $t(31) = -40.37$, $p < .001$, $d = -7.14$; narrative texts Version 2: $t(30) = -28.85$, $p < .001$, $d = -5.18$.

3.1.3 | Additional analyses for generation accuracy

In the complete sample, including those who anticipated the learning test, learners filled in more letters correctly in the expository texts compared to the narratives, $t(42.73) = 2.34$, $p = .024$, $d = 0.53$. Otherwise, the results were comparable for the subset and the whole sample.

3.2 | Processing time

Effects of genre (narrative vs. expository) and learning condition (letter completion vs. sentence unscrambling vs. reading control) on processing time were analyzed in a two-factor between-subjects analysis of variance (ANOVA), with processing time measured in seconds as dependent variable. According to MEDC, participants in the letter completion and sentence unscrambling conditions were predicted to have longer processing times for both texts because of increased encoding difficulty compared with those in the reading control condition.

Processing time was missing for five English participants. The relevant descriptive statistics appear in Table 2. Consistent with the findings reported by MEDC, an ANOVA revealed a significant main effect of learning condition, $F(2, 239) = 138.85$, $p < .001$, $\eta_p^2 = 0.54$). Pairwise comparisons indicated that reading took less time than letter completion and sentence unscrambling, letter completion: $t(117) = -9.19$, $p < .001$, $d = -1.70$; sentence unscrambling: $t(178) = -16.61$, $p < .001$, $d = -2.74$. Also, letter completion took less time than sentence unscrambling, $t(192) = -6.79$, $p < .001$, $d = -1.03$.

3.2.1 | Additional analyses for processing time

In addition, we conducted an ANOVA for the whole sample. Processing time was missing for seven participants. Overall, the results for the whole sample were comparable to those obtained for the subset, except that the genre by learning-condition interaction reached significance, $F(2, 299) = 3.44$, $p = .033$, $\eta_p^2 = 0.02$. Just one significant pairwise comparison was responsible for this interaction: Letter completion took less time for expository texts than for narratives, $t(76) = -2.00$, $p = .047$, $d = -0.46$. Together with the higher accuracy for letter completion in the case of expository texts compared to narratives, there is evidence that letter completion was easier for the expository texts than for the narratives.

3.3 | Comprehensibility ratings

Effects of genre and learning condition on text comprehensibility were analyzed in a two-factor between-subjects ANOVA with comprehensibility ratings as the dependent variable. Although it seems plausible to assume that the reported comprehensibility of the two texts differs for the letter completion and sentence unscrambling conditions (compared to the reading control condition), MEDC found no effect of learning condition on the comprehensibility ratings. They found, however, a main effect of genre, indicating better comprehensibility of the narrative compared to the expository text. Thus, in the present study, the narrative was expected to be rated as more comprehensible than the expository text.

In contrast to the findings reported by MEDC, the ANOVA revealed no significant main effect of genre, but a significant effect of learning condition, $F(2, 239) = 5.54$, $p = .004$, $\eta_p^2 = 0.04$. Based on pairwise comparisons, participants who read the texts found them more comprehensible than those who filled in letters or unscrambled the texts, letter completion: $t(118) = 2.51$, $p = .038$, $d = 0.46$; sentence unscrambling: $t(177) = 3.29$, $p = .004$, $d = 0.54$. There were no significant differences in comprehensibility ratings between those who filled-in letters and those who unscrambled the texts ($p = 1.00$). Based on these results, letter completion and sentence unscrambling did indeed decrease text comprehensibility.

3.3.1 | Additional analyses for comprehensibility ratings

When these analyses were repeated for the whole sample (minus 6 missing data), the results did not differ except for the difference between reading and letter completion, which was no longer significant ($p = .09$).

3.4 | Free recall

As in the original study, recall accuracy for each idea unit in the two texts was scored (correctly recalled = 1, not mentioned or incorrectly recalled = 0; no partial scoring). About 9% of the protocols were scored by two different raters. The inter-rater reliability was high (Cohen's $\kappa > 0.76$), hence all of the remaining protocols were rated by one of the two raters, per language. Effects of genre and learning condition on free recall were analyzed in a two-factorial between-subjects ANOVA with proportion of correctly recalled information units as the dependent variable. In line with MEDC, we expected an interaction. Participants should recall more information correctly when they complete letters in the narrative and when they reorder the sentences of the expository text (compared to the reading control condition). However, recall was not expected to improve or was expected to improve less for unscrambling sentences of the narrative or completing letters in the expository text, compared to the reading control condition. This interaction was expected to persist even if processing time was included as a covariate in an analysis of covariance (ANCOVA).

In contrast to the findings of MEDC, the ANOVA for the English sample provided no evidence for the expected interaction between genre and learning condition. Instead, there was a significant main effect of genre, with better recall for the narrative text, $F(1, 244) = 200.14$, $p < .001$, $\eta_p^2 = 0.45$. Moreover, a main effect of learning condition emerged, $F(2, 244) = 7.90$, $p < .001$, $\eta_p^2 = 0.06$. Pairwise comparisons found that more information units were recalled after sentence unscrambling than after letter completion, $t(196) = 3.88$, $p < .001$, $d = 0.58$. Recall was not better, however, for either of the generation conditions when compared to reading. The results of an ANCOVA controlling for processing time did not differ from those of the ANOVA.

3.4.1 | Additional analyses with the whole English sample

When examining the whole sample, the results of the ANOVA and ANCOVA were comparable. The only exception was that pairwise comparisons now revealed a significant recall advantage for sentence unscrambling compared to reading, based on the ANOVA, $t(232) = 2.94$, $p = .011$, $d = 0.41$.

3.4.2 | German sample

The results of the ANOVA and the ANCOVA for the German sample were largely comparable to the results for the English sample. The only difference was that recall was better after sentence unscrambling than after reading, based on the simple effects test conducted to clarify the ANOVA results, $t(143) = 2.93$, $p = .012$, $d = .56$.

3.4.3 | Additional analyses with the whole German sample

No differences were observed when the whole German sample was examined, compared to the subset. The whole sample results indicate a significant main generation effect of sentence unscrambling compared to reading. In other words, more idea units were recalled after sentence unscrambling than after reading. In sum, these results do not support the interaction hypotheses put forward by MEDC.

3.5 | Additional GLMM analysis of free recall

To account for the multilevel structure of the data (idea units and participants), we estimated an additional generalized linear mixed model (GLMM) with a logit link function for recall accuracy of idea units (Dixon, 2008). The model was estimated and tested with the software packages lme4 (Bates et al., 2015) and lmerTest for R (Kuznetsova et al., 2017). Genre (−1 = Narrative, 1 = Expository text) was included as contrast-coded predictor variable in the GLMM. Learning condition was included as two contrast-coded predictor variables, each comparing one of the generation conditions with the reading control condition (−1 = Reading, 1 = Letter completion, 0 = Sentence unscrambling; −1 = Reading, 0 = Letter completion, 1 = Sentence unscrambling). Test expectancy was included as dummy-coded predictor variable with participants who reported that they had not expected a test being the reference group (test not expected = 0, test expected = 1). Also, interaction terms of genre and learning condition were included. Processing time was included as a grand-mean centered control variable. Intercepts for participants and idea units are allowed to vary randomly. Hypotheses were the same as for the ANCOVA reported by MEDC. Prior knowledge did not differ among the six experimental groups (genre by learning condition) in both samples. However, content familiarity was higher for the expository text than for the narrative text in the English-speaking sample, $F(1, 300) = 17.07$, $p < .001$, $\eta^2_p = 0.05$. Familiarity was thus included as a covariate in the GLMM. No significant differences between groups were found for the German-speaking sample.

TABLE 3 Fixed effects and variance components in the GLMM with recall of idea units as dependent variable in the English and German sample.

Parameter	English sample Recall of idea units β (*SE*)	German sample Recall of idea units β (*SE*)
Fixed effects		
Intercept	−1.18 (0.15)*	−0.87 (0.15)*
Genre[a]	−1.07 (0.14)*	−0.58 (0.14)*
Learning condition 1: Read vs. Letter completion[a]	−0.34 (0.12)*	−0.21 (0.09)*
Learning condition 2: Read vs. Sentence[a] Unscrambling	0.33 (0.14)*	0.44 (0.12)*
Test expected[b]	−0.24 (0.20)	0.10 (0.12)
Processing time[c]	0.13 (0.12)	−0.02 (0.11)
Genre* Learning condition 1	−0.14 (0.12)	−0.12 (0.09)
Genre* Learning condition 2	0.07 (0.10)	0.11 (0.07)
Content familiarity	0.09 (0.08)	-
Variance components		
Participants	1.62 (1.27)	0.78 (0.89)
Idea units	1.87 (1.37)	2.49 (1.58)

Note: Genre: −1 = narrative, 1 = expository; Learning condition 1: −1 = read, 1 = letter completion; Learning condition 2: −1 = read, 1 = sentence unscrambling; Test expected: 0 = no, 1 = yes.
[a]Contrast-coded.
[b]Dummy-coded.
[c]Grand mean-centered.
*$p < .05$ (two-tailed).

The results of the GLMM analyses were comparable for the English and German samples (Table 3). Superior recall of idea units from the narrative compared to the expository texts was observed, English: $\beta = -1.07$, $z = -7.69$, $p < .001$; German: $\beta = -0.58$, $z = -4.09$, $p < .001$. Moreover, fewer idea units were recalled after letter completion than after reading, English: $\beta = -0.34$, $z = -2.75$, $p = .006$; German: $\beta = -0.21$, $z = -2.36$, $p = .019$. Recall was better when the participants unscrambled sentences compared to reading, English: $\beta = 0.33$, $z = 2.29$, $p = .022$; German: $\beta = 0.44$, $z = 3.65$, $p < .001$.

In sum, the GLMM results were highly similar to those of the ANOVAs and ANCOVAs. The most important finding was the absence of the expected genre by learning-condition interaction. According to MEDC, letter completion should have been the most beneficial learning condition for narratives, and sentence unscrambling the most beneficial condition for expository texts. Instead, sentence unscrambling was the condition in which participants recalled the most information independent of genre. Moreover, the GLMMs even revealed a learning disadvantage for letter unscrambling compared to reading, independent of genre.

3.6 | Analyses with Bonferroni–Holm correction

A Bonferroni–Holm correction for multiple testing was applied separately for the English and the German samples, and within each sample separately for the subset and the whole sample. The correction was also applied separately for each dependent variable and for each type of analysis. No changes in terms of significance were observed after applying the correction.

3.7 | Exploratory analyses: The role of generation accuracy for the genre by learning-condition interaction

Prior research presumes that successful generation is a necessary precondition for beneficial learning (McDaniel & Einstein, 2005) and extant studies carefully designed their tasks to achieve high generation accuracy (e.g., Bjork & Storm, 2011). The mean for letter completion accuracy was quite high in our study (93% in the English sample, 97% in the German) but still somewhat lower than in the original study (98% for MEDC). To explore whether this difference from the original study was crucial, and to ensure that all necessary preconditions are met, we estimated two additional GLMMs excluding learners who had less than 95% accuracy for letter completion (separate models for the English and German samples). For the remaining learners, mean letter completion accuracy was 98% in both samples (English: $n = 267$, German: $n = 277$). The results are displayed in Table 4.

3.7.1 | English sample

A main effect for genre occurred, with recall better for the narrative text than for the expository one, $\beta = -1.18$, $z = -8.17$, $p < .001$. Moreover, a statistically significant interaction was observed between genre and learning condition (contrast: reading vs. letter completion), $\beta = -0.36$, $z = -2.54$, $p = .011$ (no longer significant after Bonferroni-Holm correction), indicating that recall was better after letter completion than after reading the narrative and better after reading than after letter completion for the expository text. The simple effects, however, were not statistically significant.

3.7.2 | German sample

A main effect for genre occurred, indicating better recall for the narrative than for the expository text, $\beta = -0.61$, $z = -4.28$, $p < .001$. Recall was also better after sentence unscrambling compared to reading, $\beta = 0.38$, $z = 3.10$, $p = .002$. Moreover, the analyses revealed an interaction between genre and learning condition (contrast: reading vs. letter completion), $\beta = -0.19$, $z = -2.02$, $p = .044$ (no longer significant after Bonferroni–Holm correction). Simple effects analyses indicate that recall for reading was better than for letter completion, for the expository text only, $\beta = -0.25$, $z = -2.05$, $p = .041$.

TABLE 4 Fixed effects and variance components in the additional exploratory GLMM with recall of idea units as dependent variable in the English and German subsamples (learners with letter completion accuracy <95% excluded).

Parameter	English sample Recall of idea units β (SE)	German sample Recall of idea units β (SE)
Fixed effects		
Intercept	−0.97 (0.15)*	−0.80 (0.15)*
Genre[a]	−1.18 (0.14)*	−0.61 (0.14)*
Learning condition 1: Read vs. Letter completion[a]	0.06 (0.15)	−0.06 (0.10)
Learning condition 2: Read vs. Sentence[a] Unscrambling	0.10 (0.15)	0.38 (0.12)*
Test expected[b]	−0.25 (0.20)	0.09 (0.12)
Processing time[c]	0.16 (0.12)	−0.03 (0.11)
Genre* Learning condition 1	−0.36 (0.14)*	−0.19 (0.09)*
Genre* Learning condition 2	0.18 (0.11)	0.14 (0.08)
Content familiarity	0.05 (0.08)	-
Variance components		
Participants	1.44 (1.20)	0.75 (0.86)
Idea units	1.88 (1.37)	2.56 (1.60)

Note: Genre: −1 = narrative, 1 = expository; Learning condition 1: −1 = read, 1 = letter completion; Learning condition 2: −1 = read, 1 = sentence unscrambling; Test expected: 0 = no, 1 = yes.
[a]Contrast-coded.
[b]Dummy-coded.
[c]Grand mean-centered.
*$p < .05$ (two-tailed).

4 | DISCUSSION

The aim of the present study was to replicate the generation effect for texts and, more specifically, the genre by generation-task interaction reported by MEDC. To this end, we conducted a close replication study of their Experiment 2, using the original tasks and material in a large sample of 300 English-speaking participants. To test whether this effect generalizes to another language context, we also translated the material and instructions to German and replicated the study with an additional sample of 300 German-speaking participants. According to the findings of MEDC, we expected the generation task of letter completion to benefit learning with narrative texts as compared to reading alone. Sentence unscrambling was expected to benefit learning with expository texts compared to reading. The opposite combination (i.e., completing letters in expository texts, and unscrambling sentences in narratives), however, was expected to elicit a notably smaller generation effect or none at all.

Overall, the results of the preregistered analyses were comparable across samples (English and German) and data sets (the whole dataset or the subset that did not anticipate the test) for the preregistered analyses. In contrast to MEDC and other prominent studies on text generation (e.g., Einstein et al., 1990; McDaniel et al., 2002), no evidence was found for the expected genre by generation-task

interaction. Specifically, letter completion did not benefit the recall of idea units from narratives (compared to reading). Although sentence unscrambling was beneficial for learning with expository texts (as expected), it also unexpectedly led to learning benefits for narratives. These results were largely consistent across the English and German samples, with and without participants who anticipated the learning test, and across the ANOVAs and ANCOVAs that controlled for processing time. (The only exception being for English learners who did not expect the learning test.) The results were also corroborated by the GLMMs, which further revealed that recall was even worse after letter completion than after reading. In other words, the results of the preregistered analyses suggest that completing letters in word fragments hinders learning with texts. These analyses are also based on samples that are much larger than the original study, providing a more accurate estimate of any phenomenon.

Although the expected genre by generation-task interaction was not obtained in the preregistered analyses, a main generation effect was found for sentence unscrambling compared to reading in most hypothesis tests. These findings are in line with the assumption that sentence unscrambling benefits learning from expository texts as postulated by MEDC (see also Einstein et al., 1990; McDaniel et al., 2002; Schindler & Richter, 2023, for example). They further corroborate findings by Einstein et al. (1990, Exp. 2) and McDaniel et al. (1994, Exp. 1–3) who found that sentence unscrambling benefits learning from narratives. However, they do not corroborate the genre by generation-task interaction predicted by the MAP framework and the contextual framework (however, see the end of the Discussion section for an explanation that is compatible with both frameworks). The benefits of sentence unscrambling also cannot be attributed to a time-on-task effect—despite the fact that unscrambling is the most time-intensive condition—because processing time was statistically controlled for in the ANCOVAs and also in the GLMMs.

Extant research (McDaniel & Einstein, 2005) suggests that high generation accuracy is a necessary precondition for the generation effect, especially when the generation task is letter completion. Text comprehension (a necessary precondition for learning, Kintsch, 1994) during the letter completion must naturally be deteriorated to the extent that word forms cannot be recognized correctly without accurate completion. In light of this assumption, it is possible that our study did not meet the critical precondition of nearly perfect letter completion to ensure sufficient text comprehension for the beneficial generation effects to emerge. Therefore, we reran our GLMM analyses including only participants with near-to-perfect letter completion.

The results of these additional analyses provide partial evidence in favor of the genre by generation-task interaction predicted by the MAP framework and the contextual framework. When letter completion accuracy was high, the expected interaction effect for genre by learning condition (contrast: letter completion vs. reading) emerged, indicating that learners recalled more information after letter completion than after reading for the English narrative and less information after letter completion than after reading for the English and the German expository texts. These results have to be interpreted very carefully though because the simple effects were not significant in the English sample. Moreover, the interaction effects in both samples were no longer significant after the Bonferroni–Holm correction. This might possibly be attributed to reduced power after excluding low-accuracy learners in the letter completion condition. Future research is thus required to determine if the letter completion advantage for narratives can be shown more reliably with a larger sample of high generation accuracy learners.

Aside from partially supporting the focal genre by generation-task interaction, these findings serve as initial evidence that the beneficial effects of letter completion in narratives might ultimately depended on near-to-perfect letter completion (at least in the English sample). This would be in line with earlier work emphasizing the role of generation accuracy for beneficial learning effects to occur after generation (McDaniel & Einstein, 2005). It would also be in accordance with findings reported by McDaniel et al. (2002, Exp. 1A and 1B), who demonstrated that the expected learning benefit for completing letters in narratives (as compared to reading) depends on task difficulty. In their study, the expected generation effect was obtained only for readers with high (but not low) word recognition skills. A likely explanation is that word recognition is already challenging in normal reading for the low-skilled readers, and they must invest too many resources for correct identification of words which leave insufficient resources for text comprehension and learning. Thus, learners should be able to complete the letters accurately and without much cognitive effort to benefit from letter completion tasks.

The results, however, are somewhat inconsistent across samples. We found no indication of a letter completion benefit in the German narrative, and sentence unscrambling improved learning in comparison to reading only in the German sample, but for both genres.

The absence of a beneficial effect of letter completion in the German sample cannot be attributed to a speed-accuracy trade-off because letter completion was more time-consuming than reading. One possible explanation is the comparatively small sample size of the letter completion group after excluding participants with generation accuracy less than 95%. Another possible explanation is that the letter completion task did not stimulate sufficient item-specific processing. However, given the strictly parallel task design, it seems difficult to imagine why this should affect recall differently in the English and German samples. Finally, it is possible that either the German narrative stimulated sufficient item-specific processing or the German learners themselves engaged in sufficient item-specific processing, thus rendering letter completion redundant. This possibility is associated with another finding warranting discussion: the overall learning improvement through sentence unscrambling found in all analyses for the German sample. One could argue that the German narrative (and maybe also the English narrative as suggested by some of the preregistered analyses) did not evoke sufficient relational processing, which depends in part on learners' individual story schemas. Fairy tales such as "The Just Reward," with their typical structure might be less common and familiar to the learners of today than to participants of MEDC who took part almost 40 years ago which consequently would result in poorer relational processing of the narrative. According to

McDaniel and Einstein (1989), this would force the readers to rely more on item-specific processing to build understanding which would explain both the absence of a learning benefit after letter completion for the narrative and the unexpected learning benefit after sentence unscrambling. This explanation is further supported by McDaniel et al. (2002, Exp. 2A) who found that poor structure-building comprehenders benefitted from unscrambling sentences in narratives but not from letter completion. Future research is required to further clarify the absence of the letter completion advantage for narratives and the overall learning advantage for sentence unscrambling in the German sample.

Now, why was this main effect of sentence unscrambling compared to reading not obtained in the English sample? A likely explanation is that high generation accuracy is also a necessary precondition for a generation effect to occur after sentence unscrambling. As can be seen from Table 1, sentence unscrambling accuracy is notably higher in the German sample as compared to the English sample which could explain the absence of a beneficial sentence unscrambling effect in the latter one. Differences in generation accuracy (both letter completion and sentence unscrambling) between the current German and English samples and also between the current English sample and the original MEDC sample can possibly be attributed to motivational differences, differences in reading or language abilities, or different reading strategies. At any rate, the obtained sample differences advise caution as to the generalizability of text generation benefits across languages.

4.1 | Practical implications and future research

The exploratory analyses for the English sample suggest that letter completion might be useful for learning with narratives as long as the precondition of almost-perfect letter completion accuracy is met. If this precondition is not met, letter completion can even impair learning in comparison to simple reading (or more elaborate learning strategies). However, if such high levels of accuracy are required for the generation effect to emerge, letter completion may only benefit a small proportion of readers: those capable of such high accuracy. A logical consequence hereof would be to develop letter completion material that is easy enough for all learners to achieve high accuracy, or to adjust generation difficulty to individual abilities. The viability of these strategies in educational settings remains to be seen. Future research needs to corroborate the exploratory findings with a letter completion task that can be mastered by all participants. Furthermore, future research needs to determine the exact letter completion threshold that must be crossed for a generation effect to occur.

The reported text generation effect for sentence unscrambling across genres in the German sample (which was also found in some of the English sample analyses) introduces a possible application for educational practice. An open question arising from the present study is how reliable and generalizable these findings are across languages and material, and if there is a generation accuracy threshold required for sentence unscrambling.

The present study can be understood as a starting point for a systematic investigation of the boundary conditions and potential moderators that limit the efficacy and utility of text generation interventions. One avenue for future research on the genre by generation-task interaction would be to explicitly test further preconditions of this interaction, namely the assumption that narrative and expository texts elicit different types of information processing. Addressing learner characteristics as potential moderators of any text generation effect also seems like a fruitful direction. According to Schindler and Richter (2023), there is only a small body of research focusing on how learner characteristics affect the magnitude and occurrence of the text generation effect. This includes motivation, prior knowledge, and text comprehension skills among possible others.

Finally, there are several open questions that have not yet been touched upon by the present research: Is the generation effect still observable when different texts are used? Or when the learners are informed about the upcoming test (a common practice in schools and universities)? Would it appear for other measures of learning performance prevalent in educational settings (such as multiple-choice questions or transfer tasks), and would sentence unscrambling benefit learning compared to other strategies or interventions? These questions need to be answered first before text generation can be recommended for use in educational contexts. The present study provides the foundation for investigations in this direction.

5 | CONCLUSIONS

The existing research has produced inconsistent findings regarding the reliability and generalizability of the text generation effect. The aim of the present study was to provide a close replication of the most influential demonstration of the genre by generation-task interaction, using both a German- and an English-speaking sample. Replicating this interaction in these two samples would provide valuable evidence that the effect can be reliably reproduced, at least under limited conditions. It would also provide the foundation for directed investigations into the boundary conditions of this effect, with an eye toward its utility in applied contexts.

The present study provided partial evidence in favor of the genre by generation-task interaction for the English sample. There is some indication that letter completion can improve learning with narratives when letters are completed with high accuracy. Sentence unscrambling enhanced learning compared with letter completion and reading, consistently in the German sample (across genres and analyses) and in most additional analyses of the English data. This beneficial effect of text generation cannot be attributed to time on task. In sum, the results stress the importance of successful generation for the text generation effect to occur, and indicate that sentence unscrambling could be useful across genres to enhance learning. They further raise questions concerning an exact generation accuracy threshold, moderating factors, and the generalizability of text generation effects across languages, all of which need to be addressed in future research.

AUTHOR CONTRIBUTIONS

Julia Schindler: Conceptualization (lead); data curation (lead); formal analysis (lead); investigation (lead); methodology (equal); project administration (lead); writing—original draft preparation (lead). **Tobias Richter:** Formal analysis (support); funding acquisition (lead); methodology (equal); resources (equal); supervision (lead); writing—review and editing (equal). **Raymond Mar:** data curation (support); resources (equal); supervision (support); writing—review and editing (equal).

ACKNOWLEDGMENTS

We thank all of our student assistants who have supported data collection and coding. Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

Tobias Richter's work on this article was supported by a grant from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) for the Research Unit "Lasting Learning: Cognitive mechanisms and effective instructional implementation" (Grant FOR 5254/1, project number 450142163).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Open Science Framework at https://osf.io/wftbr/?view_only=bd602fffe4874d16a64e162f7e81343e.

ORCID

Julia Schindler https://orcid.org/0000-0002-5833-1334
Tobias Richter https://orcid.org/0000-0002-0467-9044
Raymond A. Mar https://orcid.org/0000-0002-5307-7031

REFERENCES

Abel, R., & Hänze, M. (2019). Generating causal relations in scientific texts: The long-term advantages of successful generation. *Frontiers in Psychology*, *10*, 199. https://doi.org/10.3389/fpsyg.2019.00199

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4 (Version 1.1–34) [computer software]. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Bjork, E. L., & Storm, B. C. (2011). Retrieval experience as a modifier of future encoding: Another test effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1113–1124. https://doi.org/10.1037/a0023549

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

Burnett, A. N., & Bodner, G. E. (2014). Learnin' 'bout my generation? Evaluating the effects of generation on encoding, recall, and metamemory across study-test experiences. *Journal of Memory and Language*, *75*, 1–13. https://doi.org/10.1016/j.jml.2014.04.005

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

DeWinstanley, P. A., & Bjork, E. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & Cognition*, *32*(6), 945–955. https://doi.org/10.3758/BF03196872

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language*, *59*(4), 447–456. https://doi.org/10.1016/j.jml.2007.11.004

Einstein, G. O., McDaniel, M. A., Bowers, C. A., & Stevens, D. T. (1984). Memory for prose: The influence of relational and proposition-specific processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 133–143. https://doi.org/10.1037/0278-7393.10.1.133

Einstein, G. O., McDaniel, M. A., Owen, P. D., & Coté, N. C. (1990). Encoding and recall of texts: Importance of material appropriate processing. *Journal of Memory and Language*, *29*(5), 566–581. https://doi.org/10.1016/0749-596X(90)90052-2

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Gardiner, J. M., & Rowley, J. M. C. (1984). A generation effect with numbers rather than words. *Memory & Cognition*, *12*(5), 443–445. https://doi.org/10.3758/BF03198305

Goldman, B. A., & Kelley, M. R. (2009). The generation effect in the context of lyrical censorship. *Psi Chi Journal of Undergraduate Research*, *14*(2), 72–78. https://doi.org/10.24839/1089-4136.JN14.2.72

Goverover, Y., Chiaravalloti, N., & DeLuca, J. (2008). Self-generation to improve learning and memory of functional activities in persons with multiple sclerosis: Meal preparation and managing finances. *Archives of Physical Medicine and Rehabilitation*, *89*(8), 1514–1521. https://doi.org/10.1016/j.apmr.2007.11.059

Goverover, Y., Chiaravalloti, N., & DeLuca, J. (2010). Pilot study to examine the use of self-generation to improve learning and memory in people with traumatic brain injury. *American Journal of Occupational Therapy*, *64*(4), 540–546. https://doi.org/10.5014/ajot.2010.09020

Goverover, Y., Chiaravalloti, N., & DeLuca, J. (2013). The influence of executive functions and memory on self-generation benefit in persons with multiple sclerosis. *Journal of Clinical and Experimental Neuropsychology*, *35*(7), 775–783. https://doi.org/10.1080/13803395.2013.824553

Goverover, Y., Chiaravalloti, N., & DeLuca, J. (2014). Task meaningfulness and degree of cognitive impairment: Do they affect self-generated learning in persons with multiple sclerosis? *Neuropsychological Rehabilitation: An International Journal*, *24*(2), 155–171. https://doi.org/10.1080/09602011.2013.868815

Graf, P. (1980). Two consequences of generating: Increased inter- and intraword, two consequences of generating: Increased inter- and intraword organization of sentences organization of sentences. *Journal of Verbal Learning and Verbal Behavior*, *19*(3), 316–327. https://doi.org/10.1016/S0022-5371(80)90248-0

Graf, P. (1981). Reading and generating normal and transformed sentences. *Canadian Journal of Psychology*, *35*(4), 293–308. https://doi.org/10.1037/h0081193

Graf, P. (1982). The memorial consequences of generation and transformation. *Journal of Verbal Learning and Verbal Behavior*, *21*(5), 539–548. https://doi.org/10.1016/S0022-5371(82)90764-2

Guterman, N. (Ed.). (1945). *Russian fairy tales*. Pantheon.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist*, *49*(4), 294–303. https://doi.org/10.1037/0003-066X.49.4.294

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models (Version 3.1-3)

[computer software]. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Levy, B. A. (1981). Interactive processes during reading. In A. M. Lesgold & C. A. Perfetti (Eds.), *Interactive processes in reading* (pp. I–35). Erlbaum.

Lutz, J., Briggs, A., & Cain, K. (2003). An examination of the value of the generation effect of learning new material. *The Journal of General Psychology*, *130*(2), 171–188. https://doi.org/10.1080/00221300309601283

Maki, R. H., Foley, J. M., Kajer, W. K., Thompson, R. C., & Willert, M. G. (1990). Increased processing enhances calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(4), 609–616. https://doi.org/10.1037/0278-7393.16.4.609

McDaniel, M. A., & Butler, A. C. (2010). A contextual framework for understanding when difficulties are desirable. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 175–198). Psychology Press.

McDaniel, M. A., & Einstein, G. O. (1989). Material-appropriate processing: A contextualist approach to reading and studying strategies. *Educational Psychology Review*, *1*(2), 113–145. https://doi.org/10.1007/BF01326639

McDaniel, M. A., & Einstein, G. O. (2005). Material appropriate difficulty: A framework for determining when difficulty is desirable for improving learning. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications* (pp. 73–85). American Psychological Association. https://doi.org/10.1037/10895-006

McDaniel, M. A., Einstein, G. O., Dunay, P. K., & Cobb, R. E. (1986). Encoding difficulty and memory: Toward a unifying theory. *Journal of Memory and Language*, *25*(6), 645–656. https://doi.org/10.1016/0749-596X(86)90041-0

McDaniel, M. A., Hines, R. J., & Guynn, M. J. (2002). When text difficulty benefits less-skilled readers. *Journal of Memory and Language*, *46*(3), 544–561. https://doi.org/10.1006/jmla.2001.2819

McDaniel, M. A., Hines, R. J., Waddil, P. J., & Einstein, G. O. (1994). What makes folk tales unique: Content familiarity, causal structure, scripts, or superstructures? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(1), 169–184. https://doi.org/10.1037/0278-7393.20.1.169

McDaniel, M. A., & Kerwin, M. L. E. (1987). Long-term prose retention: Is an organizational schema sufficient? *Discourse Processes*, *10*(3), 237–252. https://doi.org/10.1080/01638538709544674

McDaniel, M. A., Waddil, P. J., & Einstein, G. O. (1988). A contextual account of the generation effect: A three-factor theory. *Journal of Memory and Language*, *27*(5), 521–536. https://doi.org/10.1016/0749-596X(88)90023-X

Rumelhart, D. E. (1975). Notes on a schema for stories. In D. G. Bobrow & A. Collins (Eds.), *Representation and understanding: Studies in cognitive science* (pp. 211–236). Academic Press. https://doi.org/10.1016/B978-0-12-108550-6.50013-6

Schindler, J., & Richter, T. (2023). Text generation benefits learning: A meta-analytic review. *Educational Psychology Review*, *35*, 44. https://doi.org/10.1007/s10648-023-09758-w

Schindler, J. & Richter, T. (in preparation). *An ecologically valid test of the text generation effect*.

Schindler, J., Richter, T., & Eyßer, C. (2017). Mood moderates the effect of self-generation during learning. *Frontline Learning Research*, *5*(4), 76–88. https://doi.org/10.14786/flr.v5i4.296

Schindler, J., Schindler, S., & Reinhard, M.-A. (2019). Effectiveness of self-generation during learning is dependent on individual differences in need for cognition. *Frontline Learning Research*, *7*(2), 23–39. https://doi.org/10.14786/flr.v7i2.407

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning & Memory*, *4*(6), 592–604. https://doi.org/10.1037/0278-7393.4.6.592

Thomas, A. K., & McDaniel, M. A. (2007). The negative cascade of incongruent generative study–test processing in memory and metacomprehension. *Memory & Cognition*, *35*(4), 668–678. https://doi.org/10.3758/BF03193305

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Schindler, J., Richter, T., & Mar, R. A. (2024). Does generation benefit learning for narrative and expository texts? A direct replication attempt. *Applied Cognitive Psychology*, *38*(4), e4230. https://doi.org/10.1002/acp.4230