Educational Psychology Review (2023) 35:44
https://doi.org/10.1007/s10648-023-09758-w

META-ANALYSIS

Text Generation Benefits Learning: a Meta-Analytic Review

Julia Schindler[1] · **Tobias Richter**[1]

Accepted: 2 March 2023 / Published online: 30 March 2023

Abstract

Learning can be made more efficient when learners generate the to-be-learned text contents instead of passively receiving them. A multi-level meta-analysis was conducted to provide an overall estimate of the text generation effect's magnitude and to identify theoretically and practically relevant moderators. Overall, generation interventions improved learning with texts compared to reading them (Hedges' $g = .41$). This benefit was not attributable to time-on-task and was found across several learning conditions and settings (e.g., narratives and expository texts, multiple generation, and learning assessment tasks). The meta-analysis further suggests that generation benefits learning most strongly if the cognitive processes stimulated by the generation task complement those processes already stimulated by the text. In sum, the findings suggest that text generation can be suitable for educational applications especially if certain conditions are observed.

Keywords Expository texts · Generation effect · Multi-level meta-analysis · Learning with texts · Narrative texts

Learning can be more effective under specific conditions that make learning more difficult and instigate cognitive processes conducive to learning (Bjork & Bjork, 2011). These specific learning conditions are known as *desirable difficulties* (Bjork, 1994). Well-established examples are the distribution of learning sessions compared to blocked learning when learning involves repetition (e.g., Cepeda et al., 2006), interleaving exemplars from different categories in inductive learning (e.g., Brunmair & Richter, 2019), and the testing of already acquired knowledge (e.g., Roediger & Karpicke, 2006). Learning can also be made more effective when the to-be-learned information is generated by the learners themselves instead of being passively received (e.g., McDaniel et al., 1988; Slamecka & Graf, 1978). This *generation effect* has been demonstrated in numerous laboratory experimental studies

✉ Julia Schindler
julia.schindler@uni-wuerzburg.de

[1] Department of Psychology IV, University of Würzburg, Röntgenring 10, 97070 Würzburg, Germany

(for meta-analytic reviews, see Bertsch et al., 2007; McCurdy et al., 2020), often using different versions of a word-generation paradigm. In the classical version of this paradigm, learners are either presented with a context word and an intact target word in the reading control condition (WINTER – SNOW), or the target word is fragmented and needs to be completed by the learners in the generation condition (WINTER – S_ _ _). The finding that memory for the generated target words is better in terms of recognition and recall than for the read target words is quite robust (Bertsch et al., 2007; McCurdy et al., 2020).

Text Generation and Learning

The generation effect has also been demonstrated and replicated for more complex learning materials such as narrative and expository texts (e.g., Einstein et al., 1990; McDaniel et al., 1986, 1994, 2002). Narrative texts tell a story focusing on characters and events. They usually have a familiar story-structure and are meant to entertain the reader (Mar et al., 2021). Expository texts are meant to inform or educate. They usually convey information in the form of concepts, definitions, explanations, and arguments (Mar et al., 2021). Text generation includes all activities that involve the creation of the text material itself (or parts of it) such as letter completion in a fragmented text (e.g., *s_me lett_rs ar_ mis_in_ in t_is se_t_nce*) or reordering scrambled sentences (e.g., Sentence 1 in Position 4, Sentence 6 in Position 1, Sentence 2 in Position 3 etc.) (e.g., Einstein et al., 1984, 1990; McDaniel et al., 1986, 1994, 2002). Learning outcomes of these activities are compared to those that are obtained after reading the intact text. In this regard, text generation is different from commonly used techniques of elaborative or generative learning with texts. The latter usually entail reading the text plus an additional activity, whereas text generation–as understood here–means generating the exact same material which is read in the control condition. Investigating text generation is relevant from both a practical and a theoretical perspective. It addresses the questions how textual information can be efficiently conveyed and which processes or mechanisms that have been found to foster word-pair learning can be transferred to learning with texts.

In contrast to studies using word generation, studies using text generation have failed to produce a learning benefit consistently and reliably (e.g., Einstein et al., 1990; Maki et al., 1990, Exp. 2; McDaniel et al., 1986, Exp.1, 2002, Exp. 2a; Schindler et al., 2017; Thomas & McDaniel, 2007, Exp.1). These inconsistent findings suggest that text generation is not necessarily beneficial for all learners and under all circumstances, and thus raise the question of contextual factors and conditions that possibly moderate the occurrence and magnitude of the text generation effect.

Material Appropriate Processing: Interaction of Text Genre and Generation Task

One framework addressing this question is the *contextual framework* by McDaniel and Butler (2011; see also Einstein et al., 1990; McDaniel & Einstein, 1989, 2005). It comprises the ideas of *material appropriate processing* (MAP,

McDaniel et al., 1986; McDaniel & Einstein, 1989; see also Einstein et al., 1984) and *transfer appropriate processing* (TAP, Morris et al, 1977) and describes desirable difficulties such as text generation as the result of a complex interaction of learning material, difficulty intervention, learning assessment tasks, and learner characteristics. According to the framework, desirable difficulties can be expected to have an additional value for learning only if they stimulate cognitive processes that are relevant for learning (and for the specific learning test) and are not already stimulated by the learning material or initiated by the learners themselves.

The contextual framework and the MAP framework also provide an explanation for the specifically beneficial effects of completing letters in narratives and of unscrambling sentences in expository texts (Einstein et al., 1990; McDaniel et al., 1986, 2002). Learning should be most effective when two types of processing are stimulated during learning: (1) the processing of individual items (idea units) or propositions and (2) relational processing, that is, organizing the individual items and establishing relations between them (integration) (Einstein & Hunt, 1980; Hunt & Einstein, 1981). McDaniel et al. (1986; Einstein et al., 1984; McDaniel & Einstein, 1989) assume that letter completion stimulates individual-item (or proposition-specific) processing especially because it draws the readers' attention to individual words and idea units, whereas sentence unscrambling primarily stimulates relational processing because of the necessity to organize and integrate the contents of the scrambled sentences to unscramble them.

They further assume that also specific types of text stimulate specific types of processing more than others (see also Einstein & Hunt, 1980). Narratives usually have a familiar story schema (Rumelhart, 1975) which specifically fosters the establishment of relational processing between the propositions of a text, but they stimulate propositional processing to a lesser extent. Thus, learning can only be improved beyond reading by a task that stimulates propositional processing such as letter completion. Learners, however, usually have less well-organized schemata for expository texts. Instead, these texts tend to draw the learners' attention to individual propositions such as new words or concepts and to a lesser extent to the relations between them. Hence, learning can only be improved by a task that complementary stimulates relational processing such as sentence unscrambling which requires organizing and integrating the contents of the scrambled sentences (McDaniel & Einstein, 1989; McDaniel et al., 1986).

This genre-by-generation task interaction has been replicated several times (Einstein et al., 1984, Exp.1, 1990, Exp.2; McDaniel, 1984; McDaniel & Kerwin, 1987; McDaniel et al., 1986, Exp. 2), but other studies have produced inconsistent findings (Bjork & Storm, 2011, Exp.1–4; Burnett & Bodner, 2014, Exp. 1 & 2; DeWinstanley & Bjork, 2004, Exp. 1A-3; Einstein et al., 1990, Exp.1 & 2; Maki et al., 1990, pilot study & Exp. 1; McDaniel et al., 1986, Exp.1; McDaniel et al, 1994, Exp. 1–3; McDaniel et al., 2002; Exp. 1B & 2A; Thomas & McDaniel, 2007, Exp.1).

Further Moderators of the Effects of Text Generation on Learning

One possible explanation for these inconsistent findings is that further factors exist that moderate the occurrence or magnitude of the text generation effect.

Learning Assessment Tasks

The meta-analyses by Bertsch et al. (2007) and McCurdy et al. (2020) on the generation effect yielded different mean effect sizes for different learning assessment tasks (note that for all moderator effects reported by McCurdy et al., Cohen's *d* was calculated from the *t*-test statistics in their Tables 1 and 2, based on Lakens, 2013, Formula 2, p. 3). Although overall larger mean effect sizes were obtained by McCurdy et al. for all learning assessment tasks (calculated from *t*-test statistics in Table 1), both meta-analyses found free recall to yield the smallest mean effect size (Bertsch et al.: $d = 0.32$; McCurdy et al.: $d = 0.78$). The largest mean effect size, though, was obtained for cued recall ($d = 0.55$) by Bertsch et al. (with recognition between cued and free recall, $d = 0.46$) and for recognition ($d = 1.09$) by McCurdy et al. (with cued recall between recognition and free recall, $d = 1.07$).

Although these meta-analyses included only studies with simpler study materials such as words, sentences, or numbers, similar findings (i.e., lower effect sizes for free recall compared to recognition and cued recall) might occur for text generation (although it is noteworthy that recognition tasks are rarely used in text generation studies).

Level of Comprehension

Related to the distinction of processing item-specific (or proposition-specific) information and relational information is the level of comprehension that is assessed by a learning outcome measure. Information provided by a learner in a learning assessment task can be coded as explicitly provided in the text (representing the learners' propositional text base) or as elaborated information that goes beyond information explicitly stated in the text (e.g., by combining text information and prior knowledge to establish coherence relations or draw inferences). Such elaborated information represents the quality of the learners' situation model (Kintsch, 1994). Processing item-specific information primarily draws the readers' attention to details in the text, which aids the construction of a rich text base, whereas processing relational information primarily helps to organize individual items and establish (conceptual, chronological, or causal) relations between those items (integration), which in turn fosters an elaborated situation model. The propositional text-base representation is usually assessed in learning tests when only information is scored that has been explicitly mentioned in the text (in contrast to information that goes beyond the text). However, if information is scored that goes beyond the text, it is primarily the quality of the situation model that is reflected by the learners' test score.

Incidental vs. Intentional Learning

Learning in studies on the generation effect can be either intentional, that is, learners are being informed of the upcoming learning test (e.g., Abel & Hänze, 2019; DeWinstanley & Bjork, 2004; Maki et al., 1990), or it can be incidental, that is, learners are not informed of the learning test or even explicitly deceived about the study goal (e.g., Einstein et al., 1984, 1990; Glover et al., 1982; McDaniel et al., 1986, 2002). In their meta-analytic review of the generation effect, Bertsch et al. (2007) found an effect about twice the size for incidental learning ($d = 0.65$) compared to intentional learning ($d = 0.32$). Although the effect sizes calculated from McCurdy et al.'s (2020) *t*-test statistics (Table 2) are noticeably larger, they corroborate the conclusion of Bertsch et al. that generation effects are larger in incidental learning ($d = 1.03$) compared to intentional learning ($d = 0.61$).

It might, however, be that incidental learning does not yield the same advantage in generation with texts. It seems possible that learners read a text more thoroughly when informed of an upcoming learning test in the reading control condition, which would reduce the generation effect. But comprehending a text generally affords a minimum of meaningful and coherent processing independent of learners' intentions. Moreover, depending on the cover task, learners might read a text very closely even if naïve as to the learning test (if, e.g., asked to rate its comprehensibility). It is thus possible that the difference between incidental and intentional learning is less pronounced for texts than for less rich study material.

Learning Time Constraint

Generation tasks that stimulate elaborate processing of texts, such as letter completion and sentence unscrambling, likely are more time-consuming than reading the intact text. To rule out the possibility that potential generation advantages are just a time-on-task effect, several studies limited the learning time in both the generation and reading condition (e.g., Burnett, 2013; DeWinstanley & Bjork, 2004; McDaniel et al., 1989; Weissgerber & Reinhard, 2017). In other studies, participants were allowed as much time as they needed for generation and reading, which usually resulted in learning time differences between conditions (e.g., Abel & Hänze, 2019; Einstein et al., 1990; McDaniel et al., 1986; Thomas & McDaniel, 2007). Meta-analytic reviews investigating the contexts in which the generation effect occurs should thus take time limitation as a possible moderator into account. McCurdy et al. (2020) reported a significant difference between self-paced presentation and timed presentation of stimuli (words, nonwords, and numbers): Although the mean generation effect size (calculated from *t*-test statistics in their Table 2) was larger for self-paced presentation ($d = 0.91$), the effect was still noticeably large for timed presentation ($d = 0.67$), thus rendering it unlikely that differences in time-on-task fully explain the generation effect in less complex learning material. For texts, however, this question is still open.

Retention Interval

One important characteristic of desirable difficulties that is crucial for their utility in educational contexts is their potential to foster long-term learning (Bjork & Bjork, 2011). Bertsch et al. (2007) found the largest effect sizes for retention intervals of more than a day ($d = 0.64$) compared to immediate testing ($d = 0.41$), testing up to 1 min ($d = 0.32$), and testing after 1 min to 1 day ($d = 0.41$). In accordance with these findings, McCurdy et al. (2020) also reported significant differences between retention intervals with the largest mean effect size (calculated from *t*-test statistics in their Table 2) for long retention intervals (> 1 day, $d = 1.34$) followed by short retention intervals (5 min to 1 day, $d = 0.74$), and the smallest though noticeable mean effect size for immediate testing ($d = 0.68$). The magnitude of the text generation effect might depend on specific retention intervals in a way similar to that reported for less complex stimulus material.

Text Length

Enhancing learning difficulty by generation might pose enormous strains on learners' working memory, especially for rich and complex study material, such as texts. The working memory demands might be even more pronounced in longer compared to shorter texts. As a result, the magnitude of the text generation effect might decrease with increasing text length. Consistent with this reasoning, Bertsch et al. (2007) and McCurdy et al. (2020, mean effect sizes calculated from *t*-test statistics in their Table 2) reported decreasing generation effect sizes when the number of stimuli increased (25 or fewer: $d = 0.60$ and $d = 0.88$; 26–50: $d = 0.41$ and $d = 0.70$; > 50: $d = 0.09$ and $d = 0.63$). However, according to established text comprehension models such as the *model of text comprehension and production* by Kintsch and van Dijk (1978) and its successors, text comprehension always poses heavy demands on the reader's working memory when the text is incoherent and inferences have to be drawn in order to comprehend the text. Given that the text is usually made incoherent by generation interventions, establishing a coherent situation model should always be demanding in terms of working memory even for shorter texts. According to this reasoning, effect sizes might not vary as a function of varying text length.

Design

It is debatable whether between-subject designs and within-subject designs have the same underlying true population effect size unless the correlations between the levels of a within-subject factor are truly zero. Furthermore, measurement accuracy can systematically differ between those designs. Even though statistical methods can be used to transform between-subject effects into within-subject effects and vice versa, estimates may be biased by the type of design (Morris & DeShon, 2002). In their meta-analysis, Bertsch et al. (2007) found that the generation effect was about twice the size for within-subject designs ($d = 0.50$) compared to between-subject designs ($d = 0.28$). Similarly, the mean effect size for within-subject designs calculated from

McCurdy et al.'s (Table 1) *t*-test statistics is larger ($d = 1.08$) than for between-subject designs ($d = 0.79$). Although the text generation effect has been demonstrated with between-subject designs (e.g., Abel & Hänze, 2019; Einstein et al., 1984; McDaniel et al., 1989) and within-subject designs (Bjork & Storm, 2011; Goldman & Kelley, 2009; McDaniel et al., 1989), larger effect sizes might be obtained for within-subject designs analogous to Bertsch et al.'s (2007) findings.

Aims of the Present Study

The general purpose of the present study was to provide a systematic quantitative review of the text generation effect. First, we sought to provide an estimate of the magnitude of the text generation effect based on all available studies. Second, we aimed to examine whether occurrence or magnitude of the text generation effect varies as a function of study and sample characteristics, characteristics of the intervention and the texts used for learning, and characteristics of the learning test. Identifying moderators of the text generation effect is important for assessing its generalizability and for practical applications of generation in educational settings. Moreover, some of the moderators included in the meta-analysis, in particular text genre, generation task, and their interaction, are relevant for theoretical accounts of the generation effect such as the contextual framework and the material appropriate processing framework (McDaniel & Einstein, 1989; McDaniel et al., 1986). Therefore, a third aim was to examine the genre-by-generation task interaction which is of particular theoretical relevance for the text generation effect.

Method

Review Criteria

We used combinations of the search string "*generative learning*" OR "*generation effect*" AND "*text comprehension*" to locate scientific publications in the Google Scholar database resulting in 984 hits, and the search string "*generative learning*" OR "*generation effect*" to locate scientific publications in the PsycINFO database resulting in 635 hits. We further used the same search string combined with "Dissertation/Diplomarbeit[dissertation/ thesis]" (WorldCat) and Dissertation/Abschlussarbeit[dissertation/theses]" (ProQuest) to locate pertinent grey literature in the WorldCat and ProQuest databases resulting in 193 and 174 hits. We also checked the reference lists of the meta-analysis conducted by Bertsch et al. (2007) and citations of Slamecka and Graf (1978). The research extended through December 13th, 2021. After an initial screening, we identified 48 potential eligible studies that were more closely examined for inclusion in our meta-analysis. Just one study (deWinstanley & Bjork, 2004) was also included in the meta-analysis by Bertsch et al. (2007) and no study included in our meta-analysis was included in the meta-analysis by McCurdy et al. (2020).

Selection Criteria

Only experimental studies meeting the following criteria were included in the meta-analysis:

(a) The learning materials in the generative study condition and the control condition were coherent texts of at least two sentences.
(b) The study design included a comparison of at least one generative study condition and a reading control condition, either within- or between-subjects.
(c) In experiments based on between-subjects designs, the text materials were the same in the generative study condition and the control condition. In experiments based on within-subject designs, the generative study condition(s) and control condition were not allowed to differ systematically except for the generative manipulation. In other words, assignment of text materials to the generative study condition(s) and the control condition were required to be counterbalanced between participants.
(d) Participants in the generation condition(s) were presented with an explicit generation instruction. The only exception were studies using letter unscrambling. In this case, a specific instruction is not necessary because the reader automatically unscrambles the letters during reading.
(e) Comprehension outcomes in terms of relational or detail-oriented learning or both were assessed.
(f) Studies reported appropriate statistical values to calculate the effect size for the comparison between the generation condition and the control condition.

Dependency between effects (dependent samples) was not a criterion for exclusion, for example, when several learning outcomes were assessed in the same sample of participants. Dependency was addressed by using appropriate statistical techniques (multilevel meta-analysis; see the section Meta-Analytic Strategies). When the data were duplicated (appeared in more than one source), we only used the earlier source in our meta-analysis. For example, results of Burnett and Bodner (2014) were also reported in the master thesis of Burnett (2013) and therefore not included. In total, 20 studies reporting 129 effect sizes based on 74 unique samples fulfilled the inclusion criteria.

Coded Variables

A carefully trained rater coded study and sample characteristics, characteristics of the intervention (generation vs. control condition), the experimental texts used in the study, and the learning outcome for each effect size reported in the studies that met the inclusion criteria. The codings were checked by the authors.

Study and Sample Characteristics

The first groups of variables coded for each study were study and sample characteristics:

(a) **Publication Year**. Effect sizes can depend on the year when the study was published, for example, because of a change in methodological standards or practices, false-positive results of earlier studies, changing research foci, or time-specific confounding variables not otherwise assessed. Occasionally, later studies fail to replicate earlier findings (e.g., Ioannidis & Trikalinos, 2005; Ioannidis et al., 2001). Therefore, we coded the year of publication for every study. The publication year ranged from 1982 to 2019. For the meta-regression analysis, we centered the publication year around 2000. For descriptive statistics and sensitivity analysis, we split the studies into two groups of approximately equal size with the year 2000 used as the cut-off for the grouping. The first group contains studies published (or conducted, in the case of unpublished studies) from 1982 to 2000 and the second group contains studies published (or conducted) from 2001 to 2019.

(b) **Publication Status**. Studies with significant results (and studies with larger effects) tend to be published more often than studies with nonsignificant results (and studies with smaller effects). Therefore, a meta-analysis based only on published studies can overestimate the true population effect. Unpublished studies are assumed to be less prone to a publication bias. For the meta-regression analysis, we dummy-coded journal articles as published studies (coded 0) and theses and dissertations as unpublished studies (coded 1).

(c) **Sample.** According to a second-order meta-analysis by Peterson (2001; see also Hanel & Vione, 2016), university students are usually a more homogenous sample as compared to samples from other populations in social science research. Peterson demonstrated that their responses are significantly more homogenous and that the effect sizes obtained from university-student samples often differ notably from those of nonstudent samples. Directionality of effects even differed in 19% of the meta-analytic findings between both groups. Thus, we also coded the university-student status in our meta-analysis, in order to account for any unexpected differences between university students ($k = 104$) and other populations ($k = 25$) such as high-school students (Abel & Hänze, 2019), multiple sclerosis patients, and mixed samples from the general population (Goverover et al., 2008, 2013, 2014). We used dummy-coding for the meta-regression analysis, with university student samples coded as 0 and other samples coded as 1.

(d) **Design**. We dummy-coded the study design for the meta-regression analysis, with between-subject designs coded as 0 and within-subject designs coded as 1.

Intervention and Text Characteristics

The second group of variables coded for the meta-analysis consisted of characteristics of the intervention and the experimental texts used in the study:

(a) **Generation Task.** We coded the type of generation task, distinguishing between completion (insertion of letters or words) vs. scrambling tasks and the kind of information to be manipulated in the task (letters, words, sentences). These distinctions led to five categories of generation tasks: letter completion (e.g., i_s_r_t

l_t_er_ he_e), word completion (i.e., insert _ _ _ _ here), sentence unscrambling (several mixed sentences have to be reordered with the words within each sentence being in the correct order), word unscrambling (words within a sentence are mixed and need to be reordered), and letter unscrambling (letters within words are scrambled and need to be reordered). However, word unscrambling and letter unscrambling were used in just one study each (Glover et al., 1982; Weissgerber & Reinhard, 2017). Hence, these two types of unscrambling tasks were combined into the category word/letter unscrambling. The task categories were dummy-coded for the meta-regression analysis, with letter completion serving as the reference category (coded with 0) for each of the three dummy-coded variables.

(b) **Text Genre.** All studies included in the meta-analysis used narrative texts (e.g., Grimm and Russian fairy tales, McDaniel et al., 1986) or expository texts on various topics (e.g., solar system; Glover et al., 1982; brain lateralization, Weissgerber & Reinhard, 2017; the Kanchenjunga, the third highest mountain of the world, McDaniel et al, 1986). The inclusion of text genre as moderator is theoretically relevant because of its potential interaction with generation task, as predicted by the material appropriate processing framework. In the meta-regression analysis, we dummy-coded this moderator, with expository texts coded as 0 and narrative texts coded as 1.

(c) **Intentionality.** Learning in studies on the generation effect is either intentional (e.g., DeWinstanley & Bjork, 2004) or incidental (e.g., McDaniel et al., 2002). In some publications, though, no information was provided on whether learning was intentional or incidental (e.g., Glover et al., 1982). In the meta-regression analysis, we used two dummy-coded predictors to capture the three groups of effects, with intentional learning serving as the reference category (coded as 0 in both predictors).

(d) **Text Length.** We coded the number of words in the texts that were used as learning materials. If the specific text length was not explicated in the studies, it was derived from the original texts or estimated based on such information that was provided in the studies. The range was between 182 and 2,298 words. In the meta-regression analysis, we included text length as the number of words, centered around the mean of 533.14 words. For descriptive statistics, we sorted the studies in four categories of text lengths, 0–300, 301–600, 601–900, and 900 + words.

(e) **Learning Time Constraint.** Several studies limited the learning time in the generation and control conditions to make them comparable in terms of time-on-task (e.g., Burnett, 2013; DeWinstanley & Bjork, 2004; McDaniel et al., 1989; Weissgerber & Reinhard, 2017), whereas other studies allowed individual learning time differences between conditions (e.g., Abel & Hänze, 2019; Einstein et al., 1990; McDaniel et al., 1986; Thomas & McDaniel, 2007). In the meta-regression analysis, we dummy-coded this moderator: Studies without learning time constraints were used as the reference category.

Test Characteristics

The third group of variables coded for the meta-analysis consisted of characteristics of the learning test:

(a) **Retention Interval.** We coded time between the end of the study phase and the beginning of the learning assessment test. Not all studies reported exact values, and the retention intervals used in the available studies differed considerably. Therefore, we sorted the retention intervals into four categories: Immediate tests, short retention intervals (2 to 30 min), long retention intervals (1 to 2 weeks), and mixed retention intervals (learners' performance was collapsed across two or more measuring times in the analyses). No study used a retention interval between 30 min and 1 week. We used three dummy-coded predictor variables to include retention interval in the meta-regression model, with immediate tests serving as the reference category (coded as 0).

(b) **Learning Assessment Task.** Different tasks are used for assessing learning outcomes in studies on text generation, most often free-recall (e.g., McDaniel et al., 1986) and cued-recall tasks (e.g., Thomas & McDaniel, 2007) and, in rare instances, other tasks such as verification and matching tasks (e.g., Dee-Lucas & Di Vista, 1980) and single-choice tasks (e.g., Weissgerber & Reinhard, 2017). We created three categories: free recall, cued recall, and others. The category "others" also included recognition tasks which are very rare in text generation studies. In the meta-regression, these three categories were represented by two dummy-coded predictors with free-recall serving as the reference category (coded as 0).

(c) **Level of Comprehension.** We used dummy-coding for the meta-regression analysis, with text-base level scoring (e.g., scoring of information that was explicitly mentioned in the text) as the reference group (coded as 0) and situation-model scoring (e.g., scoring of conceptual or referential information that was not explicitly stated in the text, or global thematic (relational) information presented across several sentences or passages) coded as 1. Both levels were identified by the raters based on the distinction of textbase and situation model by Kintsch (1994).

Effect Size Calculation

We calculated effects for differences between a generation condition and a receptive learning condition (control condition). Whenever available, we used means and standard deviations or standard errors to calculate effect sizes (Cohen's d for between-subject designs and d_{av} for within-subject designs, Formulae 1 and 10 in Lakens, 2013). When this information was missing or incomplete, we used test statistics (e.g., t values or t values computed from F ratios) to calculate effect sizes (d_z for between- or within-subject designs, see Formulae 2 and 7 in Lakens, 2013). Only if none of this information was available, we extracted means and standard deviations from figures displaying the results to calculate effect sizes. We excluded

studies when the available statistical information was not sufficient to compute effect sizes (e.g., Kelley et al., 2009).

When more than one outcome based on the same sample of participants matched our inclusion criteria, we included all of the effects and coded the dependencies. This was the case, for example, in studies reporting generation effects for different types of learning assessments (e.g., Abel & Hänze, 2019, reported four effects for four different learning outcomes) or in studies based on a between-subject design with two or more generation tasks that were compared to the same control group (e.g., McDaniel et al., 1986).

Cohen's d tends to overestimate effect sizes in small samples. Therefore, we applied Hedges' (1981) correction term to convert Cohen's d values into Hedges' g. Hedges' g and Cohen's d can be interpreted the same way.

Meta-analytic Strategies

Several studies included in our meta-analysis provided more than one effect size. Furthermore, multiple effect sizes were dependent on each other because effects for multiple outcome measures were reported for the same sample or two or more different generation conditions were compared to the same control group. We used a three-level meta-analysis to address this hierarchical structure of our data (Assink & Wibbelink, 2016; Cheung, 2014, 2019). The multilevel approach can take into account dependencies and the clustered structure of the effect sizes without losing information from the exclusion of dependent effects or biases through weighting errors (Van den Noortgate et al., 2013). Moreover, the covariances of dependent effects obtained in the same sample need not be known. We used a three-level random effects model to estimate the overall effect across all studies and three-level mixed-effects models for the subsequent analyses of potential moderators, which were included as fixed effects (Borenstein et al., 2010).

We calculated Q statistics to examine the variance of the residuals. Cochran's Q_B statistic (Cochran, 1954) is calculated as the sum of the squared deviations of each study's effect size from the overall effect size, weighted by the inverse of the within-study variance. Q_B follows a χ^2 distribution with k-1 degrees of freedom (k represents the number of effect sizes) and can be used to test whether the variance of effect sizes is significantly different from 0. If Q_B is less than or equal to the degrees of freedom, complete homogeneity is assumed. Likewise, the Q_M statistic can be used to test the variance explained by the model through moderators. A significant Q_M means that the model explains a significant amount of variance between effects.

We also estimated I^2 within clusters of dependent effects (I^2_{within}) and between effects based on independent samples ($I^2_{between}$) (Cheung, 2014, Formula 11), using the formula proposed by Higgins and Thompson (2002) to estimate the typical within-study variance. I^2_{within} may be interpreted as the proportion of the total variability of effects due to heterogeneity within clusters of dependent effects. Likewise, $I^2_{between}$ may be interpreted as the proportion of the total variability of effects due to heterogeneity between effects based on independent samples.

Finally, for the meta-regression models, we estimated R^2_{within} and $R^2_{between}$ to quantify the proportion of variance explained by the predictors within clusters of dependent effects and between independent samples (Cheung, 2014, Formula 17). We used the R-package *metafor* (Viechtbauer, 2010) for all steps of the meta-analysis.

Results

Twenty studies met all of the inclusion criteria. Within these studies, we identified $k = 129$ effect sizes nested in 74 unique samples resulting in a total sample size of $N = 3{,}551$ participants. All studies were published between the years 1982 and 2021. Numbers of effect sizes obtained for each moderator level and for combinations of moderators are displayed in Table 1. The data and analysis scripts underlying the results presented here are available at the repository of the Open Science Framework (OSF) (https://osf.io/y4z7v/).

In the following sections, we first report estimates for the overall text generation effect and the effects in different subsets of studies as defined by the moderators. Afterward, we report the results of a meta-regression analysis that examined the effects of multiple moderators at the same time to estimate their unique effects.

Overall Text Generation Effect

We estimated mean Hedges' g across all studies using a three-level approach. We found a medium-sized positive overall text generation effect, $k = 129$; $g = 0.41$, $SE = 0.05$, $p < 0.001$; 95% CI [0.31, 0.52]. Additionally, we estimated mean Hedges' g for each independent sample using a two-level approach for comparison. The effect was nearly the same, $k = 62$; $g = 0.42$, $SE = 0.05$, $p < 0.001$, 95% CI [0.32, 0.51]. The three-level analysis also revealed significant heterogeneity of effect sizes, $Q(128) = 528.26$, $p < 0.001$, calling for an analysis of moderating variables.

Text Generation Effects in Subsets Defined by Moderators

In a second step, we split the studies into different subsets according to the moderators of interest and estimated the corresponding mean Hedges's g for each of these subsets (see Table 1 for the effect size estimates and test statistics for moderator effects).

Publication Year

We found a significant effect of publication year when studies were split into two groups, until and after the year 2000. The effect was larger in earlier studies ($g = 0.63$, $p < 0.001$) compared to studies that appeared after the year 2000 ($g = 0.28$, $p < 0.001$).

Table 1 Text Generation: Overall Effect and Average Effects for Groups Defined by Moderators

	k	*g*	95% CI		Heterogeneity I^2		Q_M
			LL	*UL*	Within	Between	
Overall	129	0.41***	0.31	0.52	41.54	38.95	
Publication year					48.26	29.16	12.57*** (*df* = 1)
1982–2000	49	0.63***	0.48	0.79			
2001–2019	80	0.28***	0.16	0.40			
Publication status					41.57	38.57	2.42 (*df* = 1)
Published studies	116	0.45***	0.33	0.56			
Unpublished studies	13	0.21	-0.07	0.49			
Sample					41.55	39.13	0.34 (*df* = 1)
University students	104	0.43***	0.31	0.54			
Other samples	25	0.35**	0.10	0.60			
Design					42.34	38.07	0.79 (*df* = 1)
Between-subjects	62	0.36***	0.19	0.52			
Within-subjects	67	0.45***	0.32	0.59			
Generation task					40.13	40.48	5.10 (*df* = 3)
Letter completion	84	0.36***	0.24	0.48			
Word completion	13	0.43**	0.10	0.75			
Sentence unscrambling	24	0.51***	0.28	0.75			
Letter/word unscrambling	8	0.84***	0.38	1.30			
Text genre					42.51	37.46	2.02 (*df* = 1)
Expository texts	78	0.36***	0.23	0.49			
Narratives	51	0.51***	0.34	0.69			
Generation task within studies with expository texts					38.18	42.60	13.12** (*df* = 3)
Letter completion	46	0.23**	0.07	0.39			
Word completion	13	0.43*	0.10	0.76			
Sentence unscrambling	11	0.77***	0.40	1.13			

Table 1 (continued)

	k	*g*	95% CI		Heterogeneity I^2		Q_M
			LL	*UL*	Within	Between	
Letter/word unscrambling	8	0.85***	0.38	1.31			
Generation task within studies with narratives					28.89	49.87	1.58 (*df* = 1)
Letter completion	38	0.57***	0.37	0.76			
Sentence unscrambling	13	0.36*	0.07	0.66			
Intentionality					42.59	36.48	8.27* (*df* = 2)
Intentional	59	0.29***	0.15	0.43			
Incidental	53	0.46***	0.29	0.63			
Not specified	17	0.73***	0.46	1.00			
Text Length					44.96	33.77	10.21* (*df* = 3)
0–300	50	0.36***	0.22	0.51			
301–600	37	0.63***	0.43	0.83			
601–900	30	0.43***	0.20	0.66			
900 +	12	0.05	-0.26	0.37			
Learning time constraint					41.24	39.45	0.65 (*df* = 1)
No time constraint	67	0.37***	0.21	0.52			
Time constraint	62	0.45***	0.31	0.60			
Retention interval					43.45	37.28	1.47 (*df* = 3)
Immediate	36	0.46***	0.27	0.65			
2 -30 min	82	0.39***	0.26	0.52			
1–14 days	9	0.33*	0.01	0.64			
Mixed	2	0.69*	0.05	1.34			
Learning assessment task					45.84	30.07	18.41*** (*df* = 2)
Free recall	61	0.60***	0.47	0.74			
Cued recall	63	0.27***	0.15	0.39			

Table 1 (continued)

	k	*g*	95% CI		Heterogeneity I^2		Q_M
			LL	*UL*	Within	Between	
Other tasks	5	0.03	-0.33	0.40			
Level of comprehension					43.09	37.16	1.36 ($df = 1$)
Text base	123	0.42***	0.32	0.53			
Situation model	6	0.21	-0.14	0.57			

k Number of effect sizes; *g* Mean Hedges' *g*; *CI* Confidence interval for Hedges' *g*; *LL* Lower limit; *UL* Upper limit (of 95%-confidence interval). All effect sizes were estimated using a three-level random-effects/mixed effects model. I^2 (within) = heterogeneity within dependent samples, I^2 (between) = heterogeneity between dependent samples (see Cheung, 2014, Formula 11)

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

Publication Status

We found no significant moderator effect of published vs. unpublished studies. Only in the published studies, the mean generation effect was significant (g = 0.45, $p < 0.001$), whereas the effect was nonsignificant in unpublished studies ($g = 0.21$; $p = 0.136$). However, only 13 effects came from unpublished studies. Therefore, the estimate for the generation effect size in unpublished studies might not be reliable.

Sample

Descriptively, a slightly larger generation effect was found in studies based on student samples ($g = 0.43$, $p < 0.001$) than on non-student samples ($g = 0.35$, $p = 0.007$), but this moderator effect was not significant.

Design

The generation effect was descriptively larger in studies based on a within-subjects design ($g = 0.45$, $p < 0.001$) than in studies based on a between-subjects design ($g = 0.36$, $p < 0.001$), but this moderator effect was not significant.

Generation Task

The overall moderator effect of the type of generation task was not significant. Descriptively, letter completion tasks, on which two-thirds of the effects were based, yielded the smallest generation effects ($g = 0.36$, $p < 0.001$), followed by studies with word completion tasks ($g = 0.43$, $p = 0.010$) and sentence unscrambling tasks ($g = 0.51$, $p < 0.001$). The letter/word unscrambling tasks yielded the largest generation effects ($g = 0.84$, $p < 0.001$), although this latter category included only eight effects. Thus, the estimate might not be reliable.

Text Genre

We found no significant difference in the mean effect sizes from studies based on narrative texts ($g = 0.51$, $p < 0.001$) and from studies based on expository texts ($g = 0.36$, $p < 0.001$).

Generation Task in Studies with Expository vs. Narrative Texts

We explored the interaction of type of generation task and text genre, as hypothesized by the material appropriate processing framework and the contextual framework (Einstein et al., 1990; McDaniel & Butler, 2011; McDaniel et al., 1986, 2000), by examining the moderating effects of the type of generation task separately for studies based on expository and narrative texts. These analyses revealed a significant moderator effect of generation task in studies based on expository texts and no significant moderator effect of generation task in studies based on narrative texts

(see Table 1; the interaction pattern is depicted in Fig. 1; forest plot of effect sizes of narrative and expository text generation are provided in the supplemental material).

However, the pattern of effects for the two types of tasks that have been used in studies based on narrative tasks, letter completion and sentence unscrambling, conforms to the predictions of the material appropriate processing framework and the contextual framework. In studies based on narrative texts, letter completion tasks ($g = 0.57$, $p < 0.001$) yielded descriptively stronger generation effects than sentence unscrambling tasks ($g = 0.36$, $p = 0.017$), although this difference was not significant. In studies based on expository texts, sentence unscrambling tasks yielded stronger generation effects ($g = 0.77$, $p < 0.001$) than letter completion tasks ($g = 0.23$, $p = 0.006$), and this difference was significant, Q_M ($df = 1$) = 6.44, $p = 0.011$, $k = 57$. Moreover, to describe the interaction from a different perspective, letter completion tasks yielded significantly larger generation effects in narrative texts compared to expository texts, Q_M ($df = 1$) = 10.40, $p = 0.001$, $k = 84$, whereas for sentence unscrambling tasks, the difference in mean effect sizes for expository texts compared to narrative texts was not significant, Q_M ($df = 1$) = 1.73, $p = 0.189$, $k = 24$. The latter comparison needs to be interpreted with caution, given the relatively small number of effects based on sentence unscrambling in the two genres (narratives: $k = 13$; expository texts: $k = 11$). In sum, the predictions of the material appropriate processing framework and the contextual framework received partial support.

Intentionality

We found a significant effect for the moderator intentional vs. incidental learning. Generation effects were smallest in studies based on intentional learning ($g = 0.29$, $p < 0.001$), larger in studies based on incidental learning ($g = 0.46$, $p < 0.001$), and largest in studies that did not report whether learning was incidental or intentional ($g = 0.73$, $p < 0.001$). Direct comparisons revealed significantly larger generation effect sizes for studies with unspecified learning compared to studies based on

Summary Forest Plot for the Genre-by-Generation Task Interaction

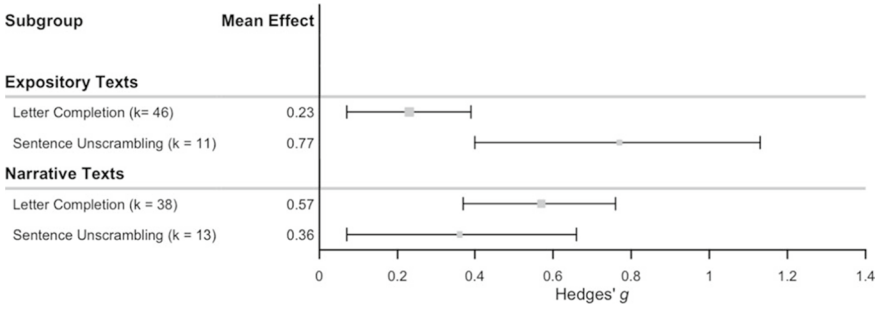

Fig. 1 Summary Forest Plot for the Genre-by-Generation Task Interaction**.** *Note.* Error bars represent the 95% confidence interval

intentional learning, Q_M ($df = 1$) = 9.06, $p = 0.003$, $k = 76$, but nonsignificant differences in generation effect sizes for studies based on intentional learning compared to those based on incidental learning, Q_M ($df = 1$) = 2.67, $p = 0.103$, $k = 112$, and for studies based on incidental learning compared to studies without further specification, Q_M ($df = 1$) = 1.93, $p = 0.165$, $k = 70$.

Text Length

The moderator effect of text length (included as a categorical variable with four groups) was significant. The effect was medium-sized for studies based on short texts (0–300 words, $g = 0.36$, $p < 0.001$), largest for studies with texts ranging from 301–600 words ($g = 0.63$, $p < 0.001$), somewhat smaller in studies with texts ranging from 601–900 words ($g = 0.43$, $p < 0.001$), and small and nonsignificant in texts of more than 900 words ($g = 0.05$, $p = 0.738$). The generation effect sizes differed significantly between studies based on short texts and studies based on texts ranging from 301–600 words, Q_M ($df = 1$) = 4.73, $p = 0.030$, $k = 87$, between studies based on very long texts (900 + words) and studies based on short texts, Q_M ($df = 1$) = 10.21, $p = 0.001$, $k = 62$, and between studies based on very long texts and studies based on texts ranging from 301–600 words, Q_M ($df = 1$) = 10.52, $p = 0.001$, $k = 49$. However, the latter two differences need to be interpreted with caution given the small number of effect sizes based on texts with more than 900 words ($k = 12$). No other comparisons within the four text-length groups reached significance.

Learning Time Constraint

We found no significant moderator effect for whether a time constraint was set for learning. The mean effect for studies with a time constraint was 0.45, $p < 0.001$) and the mean effect for studies without a time constraint was 0.37 ($p < 0.001$).

Retention Interval

No significant moderator effect of retention interval was found. The mean effect sizes for the ordered categories of retention intervals were all very close to each other and to the overall generation effect (immediate: $g = 0.46$, $p < 0.001$; 2–30 min: $g = 0.39$, $p < 0.001$, 1–2 weeks: $g = 0.33$, $p = 0.040$). The only exception was the category with mixed retention intervals ($g = 0.69$, $p = 0.036$) but this category was represented by only two effects.

Learning Assessment Task

We found a significant moderator effect of the learning assessment task. The generation effect was largest in studies based on free-recall tasks ($g = 0.60$, $p < 0.001$), smaller in studies based on cued-recall tasks ($g = 0.27$, $p < 0.001$), and non-existent in studies based on other tasks such as single-choice or verification tasks ($g = 0.03$, $p = 0.851$). The effect sizes in all groups differed significantly from each other (free vs. cued recall: Q_M ($df = 1$) = 13.61, $p < 0.001$, $k = 124$; free recall vs. other tasks:

Q_M ($df = 1$) = 5.14, $p = 0.023$, $k = 66$; cued recall vs. other tasks: Q_M ($df = 1$) = 7.33, $p = 0.007$, $k = 68$). However, the category 'other tasks' includes a variety of heterogenous tasks such as verification or single-choice tasks and is based on only five effect sizes. Comparisons involving this category thus need to be interpreted with great caution.

Level of Comprehension

We found no significant moderator effect for level of comprehension. Descriptively, the mean effect for assessment tasks requiring text comprehension on the text-base level was larger ($g = 0.42$, $p < 0.001$) compared to assessment tasks requiring text comprehension on the situation-model level ($g = 0.21$, $p = 0.242$). However, note that the mean effect for situation-model tasks was based on only six effects.

Multiple Moderator Analyses: Meta-Regression Models

In a third step, we estimated a series of six partly nested meta-regression models to assess the influence of selected moderator variables, while controlling for the effect of other moderators on the magnitude of the generation effect in text learning (Table 2). The main question to be answered by these analyses was whether the effects revealed by the separate analyses of single moderators persisted in the context of multiple and potentially correlated moderators. Some predictors (e.g., text length, publication year) were included in their original metric scale rather than in the form of ordered categories, as in the analysis of separate moderators.

In Model 1, we entered study and sample characteristics. In Model 2a, we added intervention and text characteristics as moderators. In Model 2b, we added the interaction between type of generation task and text genre highlighted by the material appropriate processing framework and the contextual framework (note that because of the dummy coding, the moderator effects for text genre in Model 2b represent effects for studies using letter completion and the moderator effects for generation task represent effects for studies using expository texts). Model 3 included study and sample characteristics as well as test characteristics but not intervention and text characteristics as predictors. Model 4a included all three groups of moderators. Model 4b additionally included the interaction between type of generation task and text genre (again, note that because of the dummy coding, the moderator effects for text genre in Model 4b represent effects for studies using letter completion and the moderator effects for generation task represent effects for studies using expository texts).

The parameter estimates in Table 2 show that the full Model 4b explained a remarkable amount (88%) of the systematic variance in effect sizes between independent samples and 17% of the variance of effect sizes within dependent samples.

Among the study and sample characteristics, publication year had a stable negative effect in all six meta-regression models, indicating that older studies yielded stronger effects than newer studies. Moreover, the text generation effect seems to be stronger in student samples compared to non-student samples when sample

Table 2 Multilevel Mixed-Effect Meta-Regression Models Estimating the Effects of Study and Sample Characteristics, Intervention and Text Characteristics, and Test Characteristics

	Model 1	Model 2a	Model 2b	Model 3	Model 4a	Model 4b
Intercept	0.330***	0.238	0.260	0.518***	0.286	0.290
Publication year	-0.018***	-0.026***	-0.024***	-0.013**	-0.021***	-0.020**
Publication status	0.004	0.022	0.037	0.080	-0.037	-0.029
Sample	0.015	-0.402*	-0.517*	-0.011	-0.261^{+}	-0.366*
Design	0.141	0.009	0.012	0.216*	-0.025	-0.021
Text Length		-0.000	-0.000		0.000	0.000
Generation Task:						
Word completion vs. letter completion		0.853**	1.003**		0.509^{+}	0.626*
Sentence unscrambling vs. letter completion		0.280*	0.542**		0.315*	0.643***
Letter/word unscrambling vs. letter completion		0.462	0.634		0.107	0.259
Intentionality						
Incidental vs. intentional		-0.093	-0.185		-0.361	-0.554*
Not specified vs. intentional		0.071	-0.023		0.317	0.247
Text genre		-0.043	0.116		0.097	0.311^{+}
Learning time constraint		0.232	0.165		0.375	0.285
Generation task (Sentence Unscrambling vs. Letter Completion) X text genre			-0.503*			-0.633**
Retention Interval:						
Short vs. immediate				-0.033	0.216	0.307^{+}
Long vs. immediate				-0.163	-0.032	0.037
Mixed vs. immediate				0.111	0.120	0.156
Learning assessment task:						
Cued vs. free recall				-0.357***	-0.315*	-0.346**
Other tasks vs. free recall				-0.422*	-0.420*	-0.446*
Level of comprehension				-0.157	-0.265	-0.335^{+}

Table 2 (continued)

	Model 1	Model 2a	Model 2b	Model 3	Model 4a	Model 4b
Q_B	429.15*** (df = 124)	356.96*** (df = 116)	356.08*** (df = 115)	349.90*** (df = 118)	307.20*** (df = 110)	300.62*** (df = 109)
Q_M	20.77*** (df = 4)	56.17*** (df = 12)	54.02*** (df = 13)	47.16*** (df = 10)	87.61*** (df = 18)	86.75*** (df = 19)
I^2_{within}	52.56%	60.58%	53.07%	59.88%	66.04%	58.62%
$I^2_{between}$	23.45%	10.35%	19.14%	11.25%	0.31%	8.14%
R^2_{within}	.00	.00	.10	.03	.08	.17
$R^2_{between}$	.51	.51	.65	.80	1.00	0.88

Publication Year: centered around the year 2000; Publication Status: dummy-coded (0 = published studies, 1 = unpublished studies); Sample: dummy-coded (0 = student sample, 1 = other sample); Design: dummy-coded (0 = between-subject design, 1 = within-subject design);

Text Length: Words per text, centered; Generation Task: dummy-coded (Letter Completion = 0 in all three predictors); Text Genre: dummy-coded (0 = Expository Texts, 1 = Narratives); Intentionality: dummy-coded (Intentional Learning = 0 for both predictors); Time Constraint During Learning: dummy-coded (0 = No Time Constraint, 1 = Time Constraint);

Retention interval: dummy-coded (Immediate Test = 0 in all three predictors); Learning Assessment Task: dummy-coded (Free Recall = 0 in both predictors); Level of Comprehension: dummy-coded (text base = 0, situation model = 1)

$^{+}$ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$ (two-tailed)

characteristics and intervention and text characteristics were included, as indicated by Models 2a, 2b, and 4b. Finally, the generation effect was larger for within-study designs compared to between-study designs when sample and test characteristics were included, as indicated by Model 3.

Among the intervention and text characteristics, the generation task exerted relatively consistent effects across all models. Sentence unscrambling yielded larger effect sizes than letter completion across all four models, but no differences were found in effect sizes for letter/word unscrambling compared to letter completion in any of the models. Word completion yielded larger generation effects than letter completion in all models except for Model 4a (note that all generation task moderator effects were estimated for expository texts only in Models 2b and 4b).

Another consistent pattern across models was nonsignificant differences in generation effect sizes between expository and narrative texts (note that in Models 2b and 4b, these moderator effects were estimated only for letter completion tasks). However, a noteworthy result was that the interaction of generation task (sentence unscrambling vs. letter completion) and text genre was significant in both models (Models 2b and 4b), indicating that the positive effect of sentence unscrambling vs. letter completion that was found in expository texts disappeared in narrative texts (see also Fig. 1). No significant moderator effects were found for text length and learning time constraint across all four models. Finally, incidental learning was associated with smaller generation effect sizes than intentional learning in Model 4b.

Among the test characteristics, free recall tasks were associated with significantly larger generation effects than cued recall or other tasks. Finally, no significant effect was found for retention interval and level of comprehension in any of the three models.

Publication Bias

We first used funnel plots (Sterne & Egger, 2001) and the trim-and-fill method recommended by Duval and Tweedie (2000) to analyze small-study effects. We used the mean effect sizes to aggregate multiple effect sizes based on dependent samples because no established multi-level methods exist to analyze publication bias. The trim-and-fill analysis suggested that a small-study effect is present ($g_{\text{Trim \& Fill}} = 0.28$, $p < 0.001$, 95% CI [0.23, 0.32]) with twelve studies missing on the left side (Fig. 2). The Egger regression test for plot asymmetry (Egger et al., 1997) was significant ($z = 4.26$, $p < 0.001$). We also rerun the Egger test as a multi-level meta-regression with the three-level data set and the modified predictor $2/\sqrt{N}$ proposed by Rodgers and Pustejovsky (2021). The multi-level Egger test with the modified predictor was also significant ($t(127) = 3.043$, $p = 0.003$). These results suggest that the research field suffers from a small-study effect, i.e. studies with smaller sample sizes yield larger positive effects than studies with larger sample sizes. This asymmetry may indicate a publication bias.

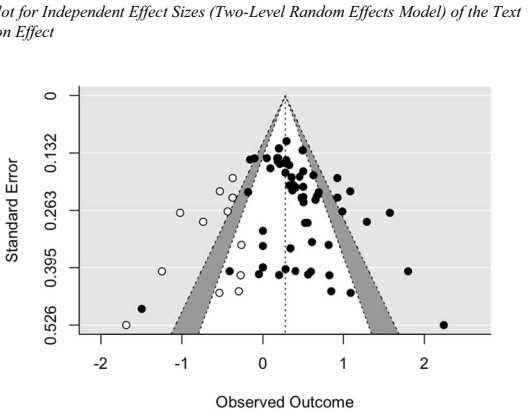

Fig. 2 Funnel Plot for Independent Effect Sizes (Two-Level Random Effects Model) of the Text Generation Effect. *Note*. Black dots represent observed effect sizes and white dots represent effect sizes imputed by the trim-and-fill procedure. The probability that effect sizes fall by chance in the particular area are $p < .01$ in light gray area, $.01 < p < .05$ in the dark-grey area, and $.05 < p < 1.00$ in the white area

As a second method to detect publication bias, we used *p*-curve analysis that has been proposed as a method to evaluate whether selective reporting of significant results in a field mistakenly suggests that an effect exists in the population even though the null hypothesis holds (Simonsohn et al., 2014). The analysis was performed with the online-app Version 4.0 provided by Simonsohn et al. (http://www.p-curve.com/app4/). The basic idea is that the distribution of *p* values for significant effects should be right-skewed, with more effects associated with smaller *p* values ($p < 0.025$) than larger *p* values. Given that a *p*-curve analysis requires independent effects, we selected the first significant effect in each cluster of dependent effects in the data file for inclusion in the analysis. When the test statistics needed for the *p*-curve analysis (*F* or *t* values) were not reported, we computed them from the available information. The analysis produced a clearly right-skewed *p* curve; 35 of the 40 effects were associated with *p* values smaller than 0.025 (Fig. 3). Thus, in contrast to the trim-and-fill analysis, the *p*-curve analysis revealed no evidence for a publication bias.

Discussion

The purpose of the present study was to provide a systematic quantitative review of the text generation effect. Apart from providing an estimate of the overall text generation effect with texts, the study's aim was to identify moderators that affect its occurrence and magnitude, which is important for assessing the generalizability and utility of text generation in educational settings. Finally, we examined the genre-by-generation task interaction which is of theoretical relevance for the contextual framework (McDaniel & Butler, 2011) and the material appropriate processing framework (McDaniel & Einstein, 1989; McDaniel et al., 1986).

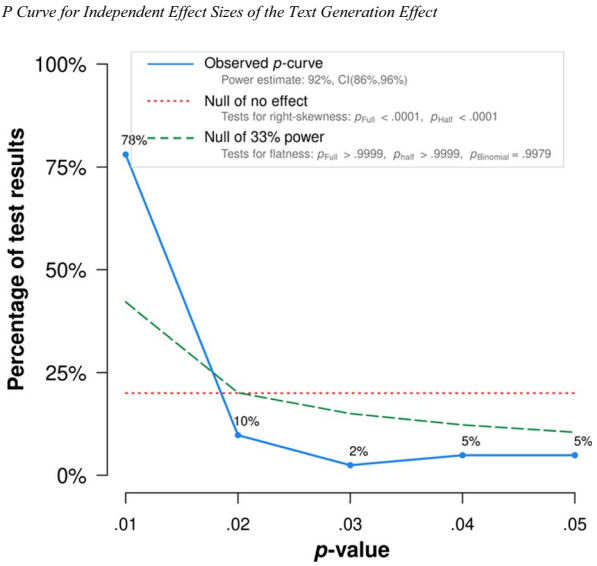

Fig. 3 P Curve for Independent Effect Sizes of the Text Generation Effect. *Note*. Of the 41 significant effects included, the *p* values of 36 effects were smaller than .025. Tests for right-skewness for the full curve ($p < .05$) and the half curve ($p < .025$) indicate that the *p* curve is right-skewed (for details, see Simonsohn et al., 2014)

Overall Text Generation Effect

We found a significant overall effect of text generation ($g = 0.41$), consistent with the findings reported for words, nonwords, and numbers by Bertsch et al. (2007) and McCurdy et al. (2020). Memory for (partially) generated texts was overall better than for texts that had been read. This finding is consistent with classifying text generation as a desirable difficulty that enhances learning by making it intentionally more difficult (Bjork, 1994; Bjork & Bjork, 2011).

Significant Moderator Effects

Generation Task, Text Genre, and Genre-by-Generation Task Interaction

Our findings suggest that generation is beneficial for all types of text generation that have been included in this meta-analytic study and for both narratives and expository texts. However, learning seems to be especially beneficial when sentences in expository texts are unscrambled. Despite the comparatively small number of effects ($k = 24$) for sentence unscrambling, our findings are of practical relevance given the predominant use of expository texts in educational contexts such as school, university, or in any setting of self-directed learning. Although generation effect sizes did not differ significantly between letter completion and sentence unscrambling for narratives, letter completion yielded descriptively larger generation effects. These results provided partial evidence in favor of the material appropriate processing framework and the contextual framework and are broadly consistent with extant studies that have reported the proposed genre-by-generation task interaction in the past (Einstein et al., 1984, Exp.1, 1990, Exp.2; McDaniel, 1984; McDaniel &

Kerwin, 1987; McDaniel et al., 1986, Exp. 2). As to learning with narratives, specific benefits of letter completion beyond those of sentence unscrambling are less clear.

Sample

A significant effect for sample, was found only in three models of the meta-regression analyses including the comprehensive Model 4b which indicates larger effect sizes for student samples as compared to non-student samples. These results suggest that it might be a fruitful future prospect to test the generalizability of the reported moderator effects especially for school-aged children.

Intentionality

No evidence was found for the assumption that learners who are aware of a later learning test would process the texts in the reading condition more thoroughly than learners who are unaware of the test. It rather seems that learners who are aware of the learning test benefit more from generation (see meta-regression Model 4b). One might speculate that learners take the generation task more seriously when expecting a test. The finding that studies without a specification as to intentionality yielded the largest effect sizes is difficult to interpret given that this category probably includes effects that are based on incidental as well as intentional learning and consists of only 17 effects.

Text Length

The largest effect sizes were found for texts of 301 to 600 words in the moderator subset analysis, whereas no generation effect was found for texts with more than 900 words. In contrast to these findings, text length was not a significant moderator effect in the meta-regression analyses. This finding, however, might be due to non-monotonic changes in effect sizes for this moderator. The increase in effect size from very short texts to texts of 301–600 words is probably covered up by the decrease in effect size for very long texts. These results are in contrast with the findings reported by Bertsch et al. (2007) and McCurdy et al. (2020) for the number of stimuli as a moderator variable and also with the alternative prediction that the effect size should not differ with text length. A likely explanation might be that other text related factors (such as readability or generation success) have a much larger effect on generation effect sizes than text length. However, due to a lack of sufficient data, those moderators could not be included in the current meta-analysis.

Learning Assessment Task

The moderator and meta-regression analyses both indicate that the generation effect is largest for free recall and significantly less pronounced for cued recall and other tasks (for the latter no generation effect was found in the subset analyses). These results differ from meta-analytic results reported by Bertsch et al. (2007) and

McCurdy et al. (2020) who reported the smallest effect sizes for free recall and the largest effect sizes for cued recall (Bertsch et al.) and recognition (McCurdy et al.). This divergence is most likely due to the less rich study material (words, non-words, numbers) in their meta-analyses. Likewise, the included generation tasks varied considerably from the tasks addressed in the current meta-analysis. According to the contextual framework and the transfer appropriate processing framework, benefits of generation can be expected to show only (or to a larger extent) if the requirements of the learning assessment task match the requirements of the generation task.

Of all 61 effects based on free recall, 29 effects were based on letter completion, 10 on word completion, and 4 on letter/word unscrambling (a Table with the number of effects for selected combinations of moderator values is provided as supplemental material). These generation tasks can be assumed to stimulate mostly proposition or item-specific processing, which matches the requirements of free recall tasks because in such tasks, learners are usually asked to provide as much explicitly stated information from the text as possible (and only such are coded consequently).

Tasks which have been categorized as cued recall, though, varied considerably including fill-in-the-blanks tasks, answering detail and conceptual questions, and remembering context words or context sentences. Also, most of these effects (54 out of 63) were based on letter completion. It seems likely that item-specific processing matches some of these tasks' demands less well than free recall, which might reduce the generation effect.

Another explanation for the larger generation effect in free as compared to cued recall could be that less retrieval cues (environmental support, Craik, 1990) are available in free recall. Consequently, free recall requires more self-initiated retrieval activity by the learner (Craik, 1990). Generation is supposed to stimulate cognitive processes which are relevant for a rich mental text representation, which provides retrieval structures that might aid free, uncued recall.

Publication Year, Published vs. Unpublished Studies, and Publication Bias

Our analyses indicated that the text generation effect decreased with increasing publication year. One possible explanation for this effect could be an increase in sample size over time. In this case, the effect of publication year might be attributed to small sample effects (only large effects can be detected with small samples) in combination with a publication bias (significant results are more likely being published than non-significant ones) in earlier studies. This explanation is supported by a negative correlation of effect-size variance as an indicator of sample size with publication year ($r = -0.46$, $p < 0.001$). and the funnel plot and Egger regression test which indicate a small-study effect.

However, the *p*-curve analysis did not reveal any evidence for publication bias. Given that the trim-and-fill analysis and the *p*-curve analysis use different criteria and are based on different assumptions to assess the presence of publication bias, it is possible that the two procedures suggest different conclusions. For example, the trim-and-fill analysis is known to lead to false-positive diagnoses of publication bias if the underlying studies represent different effects (heterogeneity), which is the case if moderators exist (Terrin et al., 2003), as in our meta-analysis. Thus, it is unclear

whether the overrepresentation of small studies with significant results indicates a publication bias.

Another possible explanation for the effect of publication year might be that newer studies often neglected the interaction between text genre and generation task. In the earlier studies (i.e., until the year 2000), 33 effects were based on optimal pairings of genre and task, whereas only 16 effects were based on optimal pairings in the studies after 2000. This imbalance could explain the publication-year effect found in the moderator subset analyses. However, in the meta-regression models, with the effects of study material, generation task, and their interaction term partialled out, the publication-year effect was still significant.

Non-significant Moderator Effects

The text generation effect occurred across all levels of several moderators, namely study design, retention interval, comprehension level, and for studies with and without a learning time constraint, which might make text generation attractive as a learning intervention in pedagogical contexts. That said, these findings need to be interpreted with caution because some of the moderators are confounded (e.g., design with learning time control, intentionality, generation task, and learning assessment task). Moreover, for some moderator levels only few effects could be obtained (e.g., long retention intervals: $k = 9$; situation model comprehension: $k = 6$), limiting the interpretability of the findings. The non-significant moderator effect for learning time constraint, though, seems to rule out a time-on-task effect as explanation for the reported generation advantage.

Limitations and Open Questions

The data included in the current meta-analytic study were insufficient to investigate potentially interesting moderators and interactions such as an interaction between generation task and learning assessment task or level of comprehension, and of their more complex three-way interactions with text genre. Generation success and various learner characteristics, such as individual differences in information processing, reading ability, topic interest and prior knowledge could not be included because of insufficient data. And even the investigated two-way interaction of text genre and generation task is difficult to interpret because only a few effects were based on sentence unscrambling.

Related to the foregoing point, study characteristics were often varied in clusters, which further complicates the interpretation of some results. Although some clusters might be useful from a theoretical perspective (e.g., using letter completion especially in combination with narrative texts), others (e.g., intentional learning in combination with time constraints and cued recall: $k = 46$, incidental learning in combination with free recall but without learning time constraints: $k = 38$) point to research gaps that need to be further investigated.

Finally, for 24 of the 67 effects based on within-subjects designs, the statistical information provided was insufficient for computing d_{av}, which was our preferred

measure of effect sizes for within-subjects designs because it is comparable to d_s for between-subjects designs. In cases where d_{av} could not be computed, we had to rely on d_z (computed from the t values) for within-subject designs (Formula 7 in Lakens, 2013). The effect size measure d_z is smaller than d_{av} if the correlation between repeated measures is lower than 0.5 but it is larger than d_{av} when the correlation between repeated measures exceeds 0.5 (Lakens, 2013). To obtain a rough estimate whether and to what extent d_z is over- or underestimated in text generation research, we computed d_{av} and d_z for all 35 effects from within-subjects designs for which the information for computing both measures was available. The (unweighted) mean difference between d_{av} and d_z was 0.07, suggesting that d_z is largely equivalent do d_{av} and, if anything, represents a slightly conservative measure for text generation effects.

Conclusion and Educational Implications

Despite these limitations, the findings of the current meta-analytic review have relevant practical implications. We found a medium-sized positive overall generation effect in learning with text materials and this effect does not appear to be attributable to a simple time-on-task effect. Moreover, this text generation effect was found across several learning conditions. Compared to reading, generation interventions seem to benefit learning for both narratives and expository texts, for all types of generation tasks, for incidental and intentional learning, for cued recall and even stronger for free recall. Our analyses further suggest that generation should especially benefit learning if the text is not too long (i.e., < 900 words) and when the cognitive processes stimulated by the generation task complement those processes already stimulated by the text genre. This last point is probably the most important precondition and advantage of text generation. Text generation improves learning beyond reading to the extent that it stimulates cognitive component processes of reading comprehension that readers do not automatically engage in (McDaniel & Einstein, 1989; McDaniel et al., 1986). Thereby, it improves text comprehension which is a necessary precondition for successful learning and long-term retention (Kintsch, 1994). This line of argumentation suggests that text generation, when administered properly (i.e., by considering possible interactions between generation task and genre), could be quite suitable to improve learning in various educational contexts such as school or university where texts are a primary source of learning. Moreover, generation tasks such as letter completion, word completion or sentence unscrambling are easy to implement in worksheets, textbooks, or digital learning environments.

Perhaps most importantly, this meta-analysis highlighted open research questions that need to be addressed in future research, and it stressed the importance of more high-powered and pre-registered studies on the text generation effect that take theoretically relevant interactions of moderators carefully into account.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10648-023-09758-w.

Acknowledgements We thank Matthias Brunmair for his contributions in an early stage of this research and Marina Klimovich for her valuable help in coding the effects and study characteristics.

Funding Open Access funding enabled and organized by Projekt DEAL. 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).

Declarations

Conflict of Interest We have no known conflict of interest to disclose.

References

References marked with an asterisk indicate studies included in the meta-analysis.

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

Assink, M., & Wibbelink, C. J. M. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. *The Quantitative Methods in Psychology, 12*(3), 154–174. https://doi.org/10.20982/tqmp.12.3.p154

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*(2), 201–210. https://doi.org/10.3758/BF03193441

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, E. L., & Bjork, R. (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.) & FABBS Foundation, *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

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*(2), 97–111. https://doi.org/10.1002/jrsm.12

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin, 145*(11), 1029–1052. https://doi.org/10.1037/bul0000209

*Burnett, A. (2013). *Learnin' 'bout my generation: The effects of generation on encoding, recall, and metamemory.* (Doctoral Dissertation, University of California). Retrieved from https://prism.ucalgary.ca/handle/11023/887. Accessed 13 Apr 2015

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

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

Cheung, M.W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods, 19*(2), 211–229. https://doi.org/10.1037/a0032968

Cheung, M. W. (2019). A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychology Review, 29*, 387–396. https://doi.org/10.1007/s11065-019-09415-6

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*, 101–129. https://doi.org/10.2307/3001666

Craik, F. I. M. (1990). Changes in memory with normal aging: A functional view. *Advances in Neurology, 51*, 201–205.

Dee-Lucas, D., & Di Vesta, F. J. (1980). Learner-generated organizational aids: Effects on learning from text. *Journal of Educational Psychology, 72*(3), 304–311. https://doi.org/10.1037/0022-0663.72.3.304

*DeWinstanley, P. A., & Bjork, E. L. (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

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Einstein, G. Q., & Hunt, R. R. (1980). Levels of processing and organization: Additive effects of individual-item and relational processing. *Journal of Experimental Psychology: Human Learning and Memory, 6*(5), 588–598.

*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: The importance of material appropriate processing. *Journal of Memory and Language, 29*(5), 566–581. https://doi.org/10.1016/0749-596X(90)90052-2

*Glover, J. A., Bruning, R. H., & Plake, B. S. (1982). Distinctiveness of encoding and recall of text materials. *Journal of Educational Psychology, 74*(4), 522–534. https://doi.org/10.1037/0022-0663.74.4.522

Goldman, B. A., & Kelley, M. R. (2009). The generation effect in the context of lyrical censorship. *Psi Chi Journal of Undergraduate Research, 14*(2), 72–77. 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. (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. D., & DeLuca, J. (2014). Task meaningfulness and degree of cognitive impairment: Do they affect self-generated learning in persons with multiple sclerosis? *Neuropsychological Rehabilitation, 24*(2), 155-171. https://doi.org/10.1080/09602011.2013.868815

Hanel, P. H. P., & Vione, K. C. (2016). Do student samples provide an accurate estimate of the general public? *PLoS ONE, 11*(12), e0168354. https://doi.org/10.1371/journal.pone.0168354

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Hunt, R. R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning and Verbal Behavior, 20*(5), 497–514.

Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A., & Contopoulos-Ioannidis, D. G. (2001). Replication validity of genetic association studies. *Nature Genetics, 29*(3), 306–309. https://doi.org/10.1038/ng749

Ioannidis, J. P., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology, 58*(6), 543–549. https://doi.org/10.1016/j.jclinepi.2004.10.019

Kelley, M. R., Goldman, B. A., Briggs, J. E., & Chambers, J. (2009). Ironic effects of censorship. In M. R. Kelley (Ed.), *Applied memory* (pp. 1–18). Nova Science.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist, 49*(4), 294–303.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

*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

Mar, R. A., Li, J., Nguyen, A. T. P., & Ta, C. P. (2021). Memory and comprehension of narrative versus expository texts: A meta-analysis. *Psychonomic Bulletin & Review, 28*, 732–749. https://doi.org/10.3758/s13423-020-01853-1

McCurdy, M. P., Viechtbauer, W., Sklenar, A. M., Frankenstein, A. N., & Leshikar, E. D. (2020). Theories of the generation effect and the impact of generation constraint: A meta-analytic review. *Psychonomic Bulletin & Review, 27*, 1139–1165. https://doi.org/10.3758/s13423-020-01762-3

*McDaniel, M. A. (1984). The role of elaborative and schema processes in story memory. *Memory & Cognition*, *12*(1), 46-51. https://doi.org/10.3758/BF03196996

McDaniel, M. A., & Butler, A. C. (2011). A contextual framework for understanding when difficulties are desirable. 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), 545–656. 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., Waddill, 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., Ryan, E. B., & Cunningham, C. J. (1989). Encoding difficulty and memory enhancement for young and older readers. *Psychology and Aging*, *4*(3), 333-338. https://doi.org/10.1037/0882-7974.4.3.333

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

*McDaniel, M. A., Waddill, P. J., Finstad, K., & Bourg, T. (2000). The effects of text-based interest on attention and recall. *Journal of Educational Psychology, 92*(3), 492–502. https://doi.org/10.1037/0022-0663.92.3.492

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*(1), 105–125. https://doi.org/10.1037/1082-989X.7.1.105

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research, 28*(3), 450–461. https://doi.org/10.1086/32373

Rodgers, M. A., & Pustejovsky, J. E. (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods, 26*(2), 141–160. https://doi.org/10.1037/met000300

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rumelhart, D. E. (1975). Notes on a schema for stories. In D. G. Brown & A. Collins (Eds.), *Representation and understanding: Studies in cognitive science*. Academic Press.

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

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P*-Curve analysis: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534–547. https://doi.org/10.1037/a0033242

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

Sterne, J. A., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology, 54*(10), 1046–1055. https://doi.org/10.1016/S0895-4356(01)00377-8

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine, 22*(13), 2113–2126. https://doi.org/10.1002/sim.1461

*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

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sànchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods, 45*(2), 576–594. https://doi.org/10.3758/s13428-012-0261-6

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

*Weissgerber, S. C., & Reinhard, M. A. (2017). Is disfluency desirable for learning? *Learning and Instruction*, *49*, 199-217. https://doi.org/10.1016/j.learninstruc.2017.02.004