



# Theories of the generation effect and the impact of generation constraint: A meta-analytic review

Matthew P. McCurdy<sup>1</sup> · Wolfgang Viechtbauer<sup>2</sup> · Allison M. Sklenar<sup>1</sup> · Andrea N. Frankenstein<sup>1</sup> · Eric D. Leshikar<sup>1</sup>

Published online: 15 July 2020  
© The Psychonomic Society, Inc. 2020

## Abstract

The generation effect is the memory benefit for self-generated compared with read or experimenter-provided information. In recent decades, numerous theories have been proposed to explain the memory mechanism(s) and boundary conditions of the generation effect. In this meta-analysis and theoretical review, we analyzed 126 articles (310 experiments, 1,653 estimates) to assess 7 prominent theories to determine which theories are supported by the existing literature. Because some theories focus on item memory (memory for the generated target) and others focus on context memory (memory for details associated with the generated target), we examined memory effects for both types of details (item, context) in this meta-analysis. Further, we assessed the influence of generation constraint (how constrained participants are to generate a certain response), which recent work has shown affects the magnitude of the generation effect. Overall, the results of this meta-analysis support some theoretical accounts, but not others, as explanatory mechanisms of the generation effect. Results further showed that generation constraint significantly moderates the magnitude of the generation effect, suggesting that this factor should be rigorously investigated in future work. Overall, this meta-analysis provides a review and examination of generation effect theories, and reveals important areas of future research.

**Keywords** Generation effect · Meta-analysis · Generation constraint · Item memory · Context memory

Techniques and devices that are designed to enhance the storage and retrieval of information from memory are generally referred to as memory strategies, or mnemonics. Memory strategies are useful in everyday life, whether it be to remember someone's name, remember to do a task later, or learn vital information. Thus, when used explicitly, these strategies have wide-ranging implications, from performing everyday tasks to improving educational outcomes, and more. Learning and memory researchers continue to investigate effective strategies and why such strategies promote memory (i.e., the underlying memory mechanisms). A better understanding of the

mechanisms supporting effective memory strategies, and the boundary conditions under which a strategy is optimal, is important in advancing our understanding of memory.

One type of strategy that has garnered substantial interest is the *generation effect* (Slamecka & Graf, 1978). The generation effect is the memory benefit for self-generated information over read, or experimenter-provided, information (Jacoby, 1978; Slamecka & Graf, 1978). Researchers have spent decades investigating how self-generating promotes memory to better understand the underlying mechanisms supporting the generation effect. In this study, we evaluate these existing theories about how the generation effect occurs by meta-analyzing data from numerous studies over the past four decades. In doing so, we aim to provide some evidence of which theories best account for mechanisms underlying the generation effect, while also highlighting some of the shortcomings of these existing theories to be addressed in future research. Overall, this meta-analysis has three goals: review and assess existing theories, investigate the influence of generation constraint as a critical moderator of the generation effect, and examine other experimental design moderators that affect the size of the generation effect.

✉ Matthew P. McCurdy  
matthew.mccurdy3@gmail.com

✉ Eric D. Leshikar  
leshikar@uic.edu

<sup>1</sup> Department of Psychology, University of Illinois at Chicago, 1007 W Harrison St (M/C 285), Chicago, IL 60607, USA

<sup>2</sup> Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Maastricht University, PO Box 616, 6200, MD Maastricht, The Netherlands

Prior research on the generation effect often uses word lists to study different aspects of the memory benefits from self-generation. In a typical study, participants are presented with a list of stimuli (usually words or word pairs). For half of the stimuli, participants self-generate a target word (e.g., open-cl\_\_\_\_), while for the other half, participants simply read an intact target word (e.g., above–below). On a later memory test for the target words, the common finding is that self-generated words are better remembered than read words (i.e., the generation effect). Decades of research has shown that the generation effect is robust across various procedures (Bertsch, Pesta, Wiscott, & McDaniel, 2007). Although several theories have been proposed to explain the generation effect, there is still some debate on the underlying memory mechanism(s) contributing to this phenomenon. Thus, the first goal of this study is to describe the prominent theories of the generation effect (see below) and then perform a meta-analysis of the extant generation effect literature to examine the extent to which these theories can explain findings on the generation effect. Some of these theories focus on item memory effects (i.e., memory for the generated item), whereas others attempt to explain context memory effects (i.e., memory for details associated with that item). We consider both types of accounts (item and context) in this meta-analysis.

Our second goal of this meta-analysis is to investigate the influence of a factor known as generation constraint on the size of the generation effect. Throughout the history of research on the generation effect, various generation tasks have been used to study this effect, but the extent that different generation tasks influence the size of the generation effect has received limited attention. Many of the existing generation effect theories do not differentiate across generation tasks (all generation tasks are typically viewed as equivalent), but recent work strongly challenges this idea. Research shows that the magnitude of the generation effect is influenced by certain features of the task used, such as *generation constraint* (McCurdy, Leach, & Leshikar, 2017; McCurdy, Sklenar, Frankenstein, & Leshikar, 2020), which refers to the amount of information given to the participant that limits, or constrains, what can be produced in a generation task. For example, in a standard generation procedure, participants are often given one or multiple *constraints*, such as a generation rule (e.g., “generate a synonym”), a cue word from which to generate a target, and letter(s) of the target word (e.g., “reply–an\_\_\_\_\_”). These constraints serve to ensure that the participant produces the expected target response to reduce item-selection confounds (i.e., idiosyncratic or unique generated responses). Some work, however, shows that tasks placing fewer constraints on what participants can generate increases the memory benefits for self-generated materials compared with tasks with more constraints (Fiedler, Lachnit, Fay, & Krug, 1992; Gardiner, Smith, Richardson, Burrows, & Williams, 1985; McCurdy et al., 2017; McCurdy et al.,

2020). Given that generation effect studies use tasks that involve varying amounts of constraints (some higher, some lower), it is possible to evaluate the history of work on the generation effect to probe this factor. Thus, we test the extent to which generation constraint has influenced memory for both item and context memory in past work.

A final goal of this meta-analysis is to examine methodological factors that influence the magnitude of the generation effect (e.g., retention interval [immediate, delayed], type of generation task [word stem, anagram, etc.], age group [younger, older]). A previous review provided some insight on important factors that influence the generation effect for *item memory* (Bertsch et al., 2007); however, given that recent work has measured the generation effect for *context memory* (Geghman & Multhaup, 2004; Marsh, 2006; Marsh, Edelman, & Bower, 2001; Mulligan, 2004, 2011; Mulligan, Lozito, & Rosner, 2006), we examine factors that influence the generation effect for both item and context memory in this meta-analysis. Knowledge about such factors can spur future research and offer a richer understanding of this memory phenomenon.

In the next section, we give a summary of the prominent theories of the generation effect that we test in this meta-analysis. Given that the primary goal of this meta-analysis is to examine empirical support for several different theories, we provide a brief overview of the theories examined in this meta-analysis. For a more elaborate review of generation-effect theories, we refer the reader to an earlier systematic review (Mulligan & Lozito, 2004). First, we describe theories that focus on item memory, followed by theories about context memory.

## Item memory theories

**Mental effort** One of the oldest and most intuitive theories of the generation effect is that of mental effort. This theory suggests that self-generation requires more mental effort (i.e., the use of more cognitive operations) at encoding relative to reading, which in turn leads to enhanced memory performance (McFarland, Frey, & Rhodes, 1980; Tyler, Hertel, McCallum, & Ellis, 1979). For instance, Tyler et al. (1979) showed that high-effort generation (difficult sentence completion and anagram solving) led to better recall than did low-effort generation (easy sentence completion and anagram solving). In this meta-analysis, we test the mental effort theory by examining the memory outcomes of studies that included a task difficulty manipulation (low-difficulty generation, high-difficulty generation) as a manipulation of effort. One issue that has plagued the mental effort theory is the lack of a universal measure and manipulation of mental effort across studies. Given the lack of a universal measure, we relied on the author’s subjective definition of effort and only included

studies that attempted to manipulate task effort to examine this theory. In line with Tyler et al. (1979), if high-difficulty generation tasks lead to a larger generation effect than low-difficulty tasks do, then this would lend support to the mental effort theory. On the other hand, finding a similar magnitude generation effect for low-difficulty and high-difficulty generation conditions would argue against a mental effort theory.

**Selective displaced rehearsal** Some research suggests that the generation effect is an artifact of experimental design (Slamecka & Katsaiti, 1987). Specifically, the selective displaced rehearsal theory claims that in within-subjects manipulations involving both generate and read conditions, participants may notice the manipulation (generation versus read) and selectively rehearse (or allocate more attention to) generated items at the expense of rehearsing read items. This account therefore posits that there is nothing special about the act of generation, but instead that the generation effect emerges simply because participants choose to focus more cognitive resources on generated items compared with read items, leading to better memory (Begg & Snider, 1987; Schmidt & Cherry, 1989; Slamecka & Katsaiti, 1987). This theory is supported by the finding that the generation effect is often eliminated in between-subjects designs, as well as pure-list presentations (i.e., blocked presentation of generate and read items; Begg & Snider, 1987; Schmidt & Cherry, 1989; Slamecka & Katsaiti, 1987). We assess the selective displaced rehearsal theory by comparing the generation effect for within-subjects versus between-subjects designs, as well as mixed-list (generate and read conditions presented in the same study list) versus pure-list (generation and read conditions presented in blocked or separate study lists) presentations of generate and read items. Finding a generation effect in within-subjects and mixed-list presentations, but no generation effect in between-subjects and pure-list designs, would support the selective displaced rehearsal concept. In contrast, finding a generation effect in between-subjects and pure-list designs would argue against it.

**Semantic (lexical) activation** The semantic activation theory suggests that the generation effect improves memory for meaningful stimuli (i.e., stimuli that are already represented in one's semantic network), but not for meaningless stimuli. For example, McElroy and Slamecka (1982) had participants generate “word–word” pairs (e.g., sand–band) and “word–nonword” pairs (e.g., sand–dand) using a rhyming generation task (e.g., generate a word that rhymes with the cue: sand–ba\_\_\_; sand–da\_\_\_). In this study, they found that generation improved memory over reading for the word–word pairs, but not for the word–nonword pairs. This work led to the semantic activation theory, which claims that the act of generation promotes access to semantic information, which in turn strengthens that memory trace, leading to better memory (McElroy & Slamecka, 1982). Originally, it was thought that

generation would only improve memory for words, but later research found that in some instances self-generation can improve memory for nonwords, but only if the nonword is meaningful (e.g., such as solutions to mathematics problems; Gardiner & Hampton, 1985; Johns & Swanson, 1988; McNamara & Healy, 2000). We examine this framework by comparing the generation effect for studies using meaningful (i.e., words and numbers) versus nonmeaningful (i.e., nonwords and nonnumbers) stimuli. Finding a generation effect for meaningful materials but not for nonmeaningful materials would support the concept of semantic activation, whereas finding a generation effect for both meaningful and nonmeaningful stimuli would argue against this account.

**Two-factor/multifactor** Perhaps the most prominent item memory theory claims that self-generation improves memory through a conjunction of two memory mechanisms at once, known as the two-factor or multifactor theory (Hirshman & Bjork, 1988; McDaniel, Wadill, & Einstein, 1988). This theory hypothesizes that self-generation improves memory relative to read controls through (1) enhanced item-specific processing and (2) enhanced relational processing. According to this theory, generation enhances the processing of item-specific information, or the characteristics that are unique to the target item. In addition to item-specific processing, this theory also proposes that generation improves memory through enhanced relational processing of details associated with, or occurring simultaneously with, the item. For example, in cue–target word pairs, generating (relative to reading) the target word strengthens the association between the cue and target (Hirshman & Bjork, 1988). Enhanced relational processing can also extend to whole-list (i.e., intertarget) relational information, or the relationship between items across different trials (Hunt & McDaniel, 1993; McDaniel, Riegler, & Wadill, 1990; McDaniel, Wadill, & Einstein, 1988). To understand the predictions that derive from the multifactor theory, it is first important to understand that the multifactor theory relies on transfer-appropriate processing (TAP) principles (Morris, Bransford, & Franks, 1977) to explain how generation tasks improve memory over reading. Transfer-appropriate processing suggests that memory performance will be improved when there is greater overlap in processing between encoding task and memory test. Research shows that different memory tests are sensitive to different types of encoding processes, like those defined by the multifactor theory (item-specific, relational processing). For example, enhanced item-specific processing is thought to enhance discriminability of learned items compared with distractor/similar items, often manifesting as better memory for generated versus read items on *recognition* memory tests, which are especially sensitive to item-specific processing (Begg, Snider, Foley, Goddard, 1989; Hunt & McDaniel, 1993). Enhanced relational processing of the cue–target relationship strengthens

the memory representation of the cue word and target, often leading to better performance on *cued recall* tests (Hirshman & Bjork, 1988). Likewise, enhanced relational processing of the intertarget relationships (i.e., processing of the relationship between different target items) leads to processing of similarities across all target items, often resulting in better performance on *free recall* memory tests (Hunt & McDaniel, 1993; McDaniel et al., 1990; McDaniel et al., 1988). Thus, the multifactor theory posits that processing enhanced by generation tasks (item-specific and relational) often provides a greater overlap (compared with reading) with the type of memory representations needed to perform well on most common memory tests, thereby leading to the generation effect.

To test this multifactor account, we compare the magnitude of the generation effect as measured by different memory tests (recognition, cued recall, free recall). The multifactor theory predicts a significant generation effect for each type of memory test relative to read controls. Given that each memory test is presumed to tap different memory representations, a significant generation effect across each test type (recognition, cued recall, free recall) would support the idea that generation improves memory via both enhanced item-specific processing and enhanced relational processing as predicted by the multifactor theory (Hirshman & Bjork, 1988; McDaniel et al., 1988). Further, it is possible that the magnitude of the generation effect differs between test types. This finding might suggest that one type of processing (i.e., item-specific as measured by recognition tests; relational processing as measured by cued/free recall tests) has a larger influence on the generation effect than the other. Finally, finding no generation advantage over read controls for any of the three types of memory tests would argue against the multifactor theory, suggesting that both types of processing (item-specific and relational) are not necessary to explain the generation effect.

## Context memory theories

More recently, researchers have investigated the impact of self-generation on memory for contextual details (e.g., source, font color). These results are more mixed than findings for item memory. Some work shows that self-generation improves memory for contextual details (Geghman & Multhaup, 2004; Marsh, 2006; Marsh et al., 2001), whereas other work shows generation does not boost context memory over reading (Jurica & Shimamura, 1999; Mulligan, 2004, 2011; Mulligan et al., 2006). These mixed findings have led to different theories to explain how self-generation influences memory for contextual details, which we describe next.

**Associative strengthening** The associative strengthening theory suggests that the act of self-generating materials induces enhanced relational processing which in turn leads to better

memory for contextual details of an episode by binding these details together into a memory representation (Greenwald & Johnson, 1989; Marsh, 2006; Marsh et al., 2001). For example, Marsh (2006; Marsh et al., 2001) found evidence that self-generation improves context memory (location and font color) relative to reading, leading to the conclusion that enhanced relational processing from generation extends to multiple details associated with an item (e.g., the cue word, location where it was learned). Geghman and Multhaup (2004) similarly argue that both item and context memory improve to the same extent under self-generation because they share a similar mechanism (what they called enhanced associative encoding). One way to test this account is by examining whether a generation effect is evident across all types of context memory. Finding a generation effect across studies that included a context memory measure would support the associative strengthening theory, suggesting that generation can improve memory for various associated details (e.g., source, font color). Finding no generation effect across various context memory details would argue against the associative strengthening theory.

**Item–context trade-off** In contrast to the associative strengthening account, the item–context trade-off theory suggests that limited cognitive resources restrict the memory benefits of self-generation for the item, but not the context. This theory suggests that because generating words requires more cognitive resources to process the target item compared with reading, less resources are available to encode extraneous details (context memory), leading to better item memory for generated compared with read items, but worse context memory for generated items compared with read (i.e., an item–context trade-off; Jurica & Shimamura, 1999; Nieznański, 2012). For instance, Nieznański (2012) found that when a high-difficulty generation task was used (which is assumed to require more cognitive resources; Tyler et al., 1979), a generation effect occurred for item memory, but context (source) memory showed a negative generation effect (read > generate), supporting the item–context trade-off account. We aim to assess this theory by comparing the generation effect for item versus context memory measures. Finding an interaction between item memory and context memory, where a generation effect is present for item memory, but a negative effect for context memory (read > generate), would support the item–context trade-off theory. Finding a significant generation effect for both item and context memory would argue against this account.

**Processing account** The associative strengthening and item–context trade-off theories make opposing predictions for context memory (improved versus diminished memory for context), but a more nuanced theory argues against a simple all-or-nothing account. Mulligan (2004, 2011) applied Jacoby's (1983) processing account to explain why only certain types of context details experience a generation effect, whereas other

details do not. The processing account relies on transfer-appropriate processing principles (Morris et al., 1977) to predict which context details are improved by self-generation. Specifically, generation tasks (e.g., generate an antonym) typically require participants to think conceptually (i.e., processing of semantic features, such as the relationship between the cue and target) to generate the target item. Conversely, the more automatic process of reading requires participants to think more perceptually (i.e., processing of the visual [or data-driven] details, such as the letters that make up the word; Jacoby, 1983; Mulligan, 2011). Similarly, different types of context details may be classified as conceptual context details (e.g., source, temporal-spatial location), or perceptual details (e.g., font color, background color, font type). Conceptual context details are thought of as nonvisual details related to the target item (like remembering whether the item was generated or read, or which location it was encoded in), whereas perceptual details are visual details associated with the target item, like its font color or font type (Mulligan, 2004, 2011).<sup>1</sup> The processing account, in line with transfer-appropriate processing principles, predicts that the type of processing required by the generation or read task should lead to memory improvements for context details that tap into the same type of processing as required by the encoding task. Thus, generation should improve memory for conceptually related context details (e.g., source, location), whereas reading should lead to improved memory for perceptually related context details (e.g., font color, font type), an idea supported by empirical findings (Mulligan, 2004, 2011; Mulligan et al., 2006). We aim to test this theory by comparing the generation effect for conceptual context details (e.g., source, location) and perceptual context details (e.g., font color, font type). Finding an interaction in which generation leads to better memory than reading for conceptual context details, but reading provides better memory over generating (i.e., a negative generation effect) for perceptual context details would provide support for the processing account. Finding a similar effect (either a generation effect, no effect, or a negative generation effect) for both types of details would argue against the processing account.

## Present study

In the following analysis, we have four primary goals: First, we examine the overall generation effect across all studies included in this analysis. Second, we test which generation-

<sup>1</sup> Although we rely on the distinction between conceptual and perceptual context details, it should be noted that it is likely the case that *both* conceptual and perceptual processing are involved in the encoding of different types of context details. Thus, the distinction between conceptual/perceptual context merely represents the primary type of processing required to encode the detail, rather than implying that solely one type of processing is engaged and not the other.

effect theories (item, context) are supported by existing empirical work. Third, we examine the extent that generation constraint influences the magnitude of the generation effect. Finding an effect of generation constraint, which has not been well studied to date, would suggest that this is an important factor to consider in future work. Fourth, we test potential moderators of the generation effect to investigate methodological factors that influence the magnitude of the generation effect. Using mixed-effects models, which allow for the inclusion of studies with multiple repeated measures (e.g., item and context memory measures from the same participants), we summarize 126 generation-effect articles with the goal of advancing our understanding of the generation effect and identifying areas that should be examined in future research.

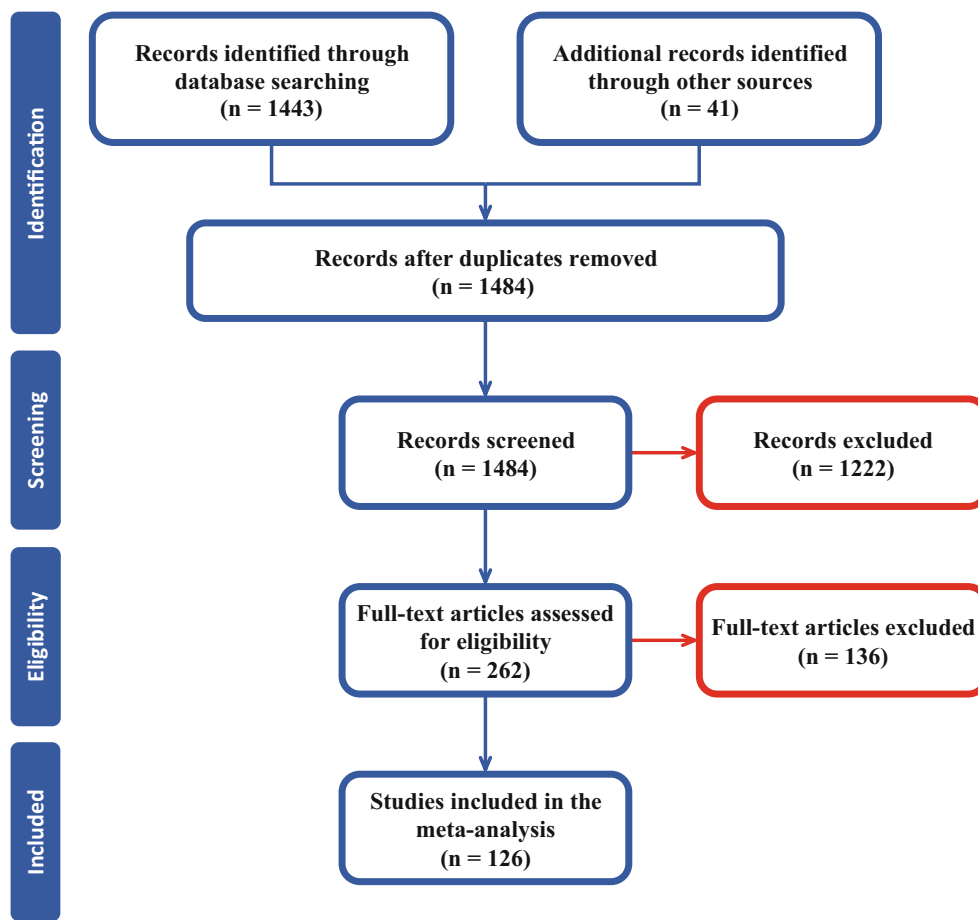
## Method

In this section, we describe our literature search procedure and how we identified articles to include in this meta-analysis. Then we describe how we coded for the various moderators included in our analyses. Finally, we describe the statistical methods used for analyzing these data.

### Literature search and coding procedure

Figure 1 depicts a PRISMA flow chart of the overall study selection process (Moher, Liberati, Tetzlaff, & Altman, 2009). Studies included in the analysis were obtained through three methods. First, we searched two online scientific publication databases (PsycINFO, PsycArticles) using the following search terms: “Generation effect OR Reality monitoring<sup>2</sup> AND source memory OR context memory OR item memory” ( $k = 1,443$  articles). The initial search was conducted on February 19, 2017. Next, we conducted a forward and backward citation search of an existing review article (Bertsch et al., 2007), to find articles not obtained by the initial database search ( $k = 35$  additional articles). Finally, we did a targeted Google Scholar search of authors with more than two publications on the generation effect based on our initial search ( $k = 6$  additional articles). This search method therefore identified a total of 1,484 articles. The titles and abstracts of all 1,484 articles identified were examined for inclusion by the first and last author, separately. To be included, at least one experiment within the article was required to (1) have one or more generation condition(s) and a read (comparison) condition; (2) report numerical or graphical data for a recognition, cued recall,

<sup>2</sup> Reality monitoring refers to distinguishing the source of internally generated and externally presented information (Johnson & Raye, 1981), and has been used to study the source memory benefits for generated versus read (or given) materials. We included reality monitoring studies where participants self-generate materials (compared with materials given to them from another source).



**Fig. 1** PRISMA-style flowchart showing the study selection for meta-analysis on the generation effect

or free recall memory measure of item memory, context memory, or both; (3) include at least one nonclinical sample (healthy younger or older adults). The two raters had acceptable agreement in their inclusion/exclusion ratings Cohen's kappa ( $k$ ) = .67 (Cohen, 1968; Viera & Garrett, 2005), and disagreements were resolved through discussion. One-thousand two hundred and twenty-two (1,222) of the articles were excluded for the following reasons: (a) no generation manipulation or no read control (comparison) condition ( $k$  = 685 articles excluded); (b) memory measure other than recognition, cued recall, or free recall ( $k$  = 143 articles excluded); (c) lack of necessary statistical information in text or graphs ( $k$  = 1 article excluded); (d) use of a clinical population or children ( $k$  = 371 articles excluded); (e) atypical methodology or stimuli (e.g., physically generating actions, drawing images;  $k$  = 22 articles excluded). This examination of the titles and abstracts led to 262 articles that were deemed potentially relevant. As a further vetting, the first author scrutinized the full text of this subset of 262 articles using the same exclusion criteria: (a) no generation manipulation or no read control (comparison) group ( $k$  = 19 additional articles excluded); (b) memory measure other than recognition, cued recall, or free recall ( $k$  = 13 additional articles excluded); (c) lack of necessary statistical

information in text or graphs ( $k$  = 5 additional articles excluded); (d) use of a clinical population ( $k$  = 19 additional articles excluded); (e) atypical methodology or stimuli ( $k$  = 80 additional articles excluded). This process led to a final total of 126 included articles. There was no formal date restriction on our initial search, but the publication dates of the final list of articles included ranged from 1978 to 2019.<sup>3</sup>

### Moderator variables

Each study included was coded for several moderator variables. The moderator variables to be coded and the levels of these variables were determined a priori by the first and last author based on theoretical interest. All studies were coded by the first author, with the exception of the generation constraint variable (which was coded by two independent raters). To assess reliability of the moderator coding, the third author independently coded a random sample of 20% of the articles included (25 articles). There was high agreement among the

<sup>3</sup> An originally unpublished article was published during the time of the writing of the manuscript; thus, the date of that article (2019) exceeds the date we last searched for articles (February, 2017).

coders across the moderators (Cohen's kappa ( $k$ )  $M = .87$ ,  $SD = .08$ ; range: .70–1.00; Cohen, 1968; Viera & Garrett, 2005),<sup>4</sup> and all disagreements were resolved through discussion. For the generation constraint variable, a set of rules to code the amount of constraint provided by the generation task used in each study was developed by the first author. The *generation constraint* variable had three levels (low, medium, high constraint) and was coded based on the number of “cues” given to the participants for generating the target item, with “cues” being defined as “any information given by the experimenter to instruct participants on what to generate.” Using a word-pair stimulus as an example: low-constraint generation would provide a cue word, but no explicit rule on how the participant should generate the target word (e.g., open-\_\_\_\_\_); medium-constraint generation would provide a generation rule and a cue word (e.g., “generate an antonym”; open-\_\_\_\_\_); high-constraint would provide a rule, cue word, and part of the target word (e.g., “generate an antonym”; open-cl\_\_\_\_\_). Two independent coders (the third and fourth authors) were trained on these rules and independently coded this variable for all studies included in the meta-analysis. There was moderate agreement between the coders' independent ratings for this variable, Cohen's kappa ( $k$ ) = 0.59, weighted kappa ( $k_w$ ) = 0.51 (Cohen, 1968; Viera & Garrett, 2005), and all disagreements were resolved through discussion between the coders and the first author to create a consensus code of generation constraint that was used in the analyses.

Each study included in this meta-analysis was coded for a total of 18 different experimental moderators. For ease of comprehension, we first list moderators used to evaluate the different generation effect theories, followed by the moderators used to test various experimental procedures that influence the generation effect. The following variables were used to assess the seven different theories tested: *condition type* (generate, read); *generation difficulty* (low difficulty, high difficulty); *manipulation type* of generate versus read condition (within subjects, between subjects); *presentation style* (mixed list, pure/blocked list); *word status* (words, non-words, numbers), *memory test* (recognition, cued recall, free recall); *memory type* (item, context); *context memory type* (conceptual context, perceptual context); *generation constraint* (low, medium, high constraint).

The remaining variables were tested as experimental moderators to assess their influence on the generation effect: *learning type* (incidental, intentional); *stimuli relation*<sup>5</sup> (semantic, category exemplars, antonym, synonym, rhyme, compound words,

definitions, unrelated); *generation mode* (verbal/speaking, covert/thinking, writing/typing); *generation task* (anagram, letter transposition, word fragment, sentence completion, word stem, calculation); *divided attention* (divided attention, full attention); *presentation pacing* (self-paced, timed); *filler task* (filler, no filler); *subject population* (younger adults, older adults); *retention delay* (immediate, short [5 minutes–1 day], long [>1 day]); *number of stimuli* (25 or fewer, 26–50, more than 50).

## Statistical analysis

The data set was structured for analysis using an “arm-based” model (Salanti, Higgins, Ades, & Ioannidis, 2008; Stram, 1996). For each comparison of memory performance between read (control) versus generate (experimental) condition, the data set includes two rows, corresponding to the two conditions (read, generate). We then quantified the mean proportion of correctly remembered items in the two conditions, using a dummy variable to distinguish the condition type (0 = read; 1 = generate). In some cases, the same read condition was compared with multiple types of generate conditions (e.g., one with low and one with high constraints), in which case additional rows were added to reflect these extra conditions. Moreover, some comparisons (or “pairings” of one or multiple generate conditions with a single read condition) were examined within the same experiment, and the same article sometimes included multiple experiments. In addition, some experiments included a single sample of participants (i.e., within-subject design), whereas other studies included more than one sample (i.e., between-subjects design). We therefore coded this hierarchical structure based on an article identifier, an experiment identifier (i.e., a single experiment within a multiexperiment article), a sample identifier (within experiment), and an observation (i.e., row) within sample identifier. Finally, we created a “pairing” identifier that was nested within experiment, but not strictly within sample or vice versa (i.e., the same pairing might involve two different samples, as in a between-subjects design, or the same sample might be used in two different pairings, as in a within-subjects design). We illustrate in more detail how the structure of the data set was coded using these identifiers in the [Appendix](#).

We then used a mixed-effects multilevel model with random effects for articles, experiments, samples, and observations (an extension of the model described by Konstantopoulos, 2011), plus a crossed random effect for “pairings” within experiments (Fernández-Castilla et al., 2019). The random effects for articles, experiments, samples, and observations account for the multilevel structure of the data set and dependencies that can arise thereof (e.g., from multiple mean proportions coming from the same sample of participants), while the random effect for the “pairings” within experiments ensures that we are comparing “alike with alike” (except for the condition type). That

<sup>4</sup> A table reporting the interrater reliability statistics for each moderator is available at: <https://osf.io/9pv7a/>

<sup>5</sup> *Stimuli relation* refers to the type of relationship between the cue and target when word pairs were used as stimuli (i.e., the cue–target relation). For example, a “semantic” relation would represent a word pair where the cue and target were semantically related (e.g., bank–money), whereas an “antonym” relation represents a cue and target that are opposites (e.g., open–close).

is, rows for the same pairing represent a particular comparison between generate versus read conditions with other factors held constant. Finally, to distinguish the latter, the model includes a fixed effect for the condition type dummy variable, whose coefficient reflects the estimated (average) difference in the mean proportion of items recalled in generate versus read conditions (i.e., the effect size) across all studies included in the model.<sup>6</sup>

In some cases (12.3% of all mean proportions included), numerical data for a mean proportion were not reported, in which case graphical data were converted to numerical data using the WebPlotDigitizer program (Rohatgi, 2015). Additionally, we recorded the variance of the proportions reported for a particular condition when available. For conditions where the variance (or some other derivative variance statistic) was not reported (74% of cases), the variance was imputed based on the expected quadratic relationship between the mean proportion and the variance of the proportions as observed in the studies that report both pieces of information (i.e., if the mean proportion is 0 or 1, the variance must be 0; for mean proportions around 0.5, the variance will tend to be highest). A mixed-effects meta-regression model was used for this purpose, using the log-transformed standard deviation as the outcome (Raudenbush & Bryk, 1987), with the mean proportion, mean proportion squared,  $1/(\text{number of items})$ , and  $1/(\text{sample size})$  as predictors. The model accounted for a substantial amount of heterogeneity ( $R^2 = .61$ ). Predicted values based on this model were exponentiated and squared and were used to impute missing variances. The sampling variance of each mean proportion,  $\bar{p}$ , was then computed with  $\text{Var}[\bar{p}] = (\text{reported/imputed variance of the proportions}) / (\text{sample size})$ , and these variances were then included in the model in the usual meta-analytic manner (as the variances of the sampling errors of the observed outcomes).

Because of the inclusion of within-subjects designs, sampling errors for multiple mean proportions obtained from the same sample are not independent. Computation of the covariances of the sampling errors is not possible in the present case, as this would require access to the raw data from such studies. Instead, after fitting the mixed-effects model, we used cluster-robust inference methods (with clustering at the article level) to account for any dependencies that are not already captured by the random effects included in the model (Moeyaert et al., 2017). All inferences reported (i.e., tests and confidence intervals) are based on this approach.

In meta-analyses, it is widely recognized that data may be influenced by publication bias, which reflects the possibility that studies included in the meta-analysis (which tend to be published studies in the majority of cases) may not be

representative of all studies actually conducted. This phenomenon is especially problematic when the missing studies are more likely to be those that did not find a statistically significant effect (among other factors; Rothstein, Sutton, & Borenstein, 2006), in which case the included studies are biased toward showing a greater effect than actually exists. One common way to assess whether publication bias may be influencing the results of a meta-analysis is to include some measure of precision of the estimates as a predictor in a meta-regression model (also known as Egger's test; Egger et al., 1997). In the present case, where the sampling variances are known to be directly related to the size of the outcomes, the usual procedure of using the standard errors (i.e., square-root of the sampling variances) as predictor is not advisable, as this will suggest the presence of publication bias even in its absence. Instead, we used the inverse sample size as the measure of precision, which is known as Peters' test (Peters, Sutton, Jones, Abrams, & Rushton, 2006). Moreover, for an arm-based analysis, publication bias is not expected to manifest itself through a main effect of the precision predictor, but through an interaction of the predictor with the condition dummy variable (under publication bias, we would expect studies failing to find a significant *difference* between read and generate conditions to be missing). Therefore, main effects for condition, precision, and, most importantly, their interaction were included in the model. Based on this model, we also compute estimated condition effects (i.e., the size of the generate versus read difference) as a function of precision, extrapolating up to the case of infinite precision (i.e., where  $1/(\text{sample size}) = 0$ ), which can cautiously be interpreted as an estimate of the effect "corrected" for potential publication bias, analogous to the now commonly used PET/PEESE methods (Stanley & Doucouliagos, 2014).

To test each theory or experimental moderator, we interacted the condition type (generate, read) factor with the moderator of interest. These interaction models provide coefficients describing the change in the estimated mean proportion for generate compared with read items for each level of the moderator included (e.g., within-subjects design versus between-subjects design), allowing for the comparison of the generation effect (memory benefit for generate compared with read) between the different levels of the moderator. In the results, we report comparison statistics on the interaction coefficient to examine each generation effect theory, and to assess the influence of the other experimental moderators on the size of the generation effect. All analyses were conducted in R (R Core Team, 2018) using the *metafor* package (Viechtbauer, 2010).

## Results

The final data set included 126 articles, 310 experiments, and 579 samples, yielding a total of 1,653 mean proportions (i.e.,

<sup>6</sup> Although the model does not include a corresponding random effect for the condition dummy variable (i.e., a random slope), the observation-level ("pairing") random effect essentially fulfills the same purpose, allowing for heterogeneity in the size of the condition effect across pairings.



rows of data), 804 of which corresponded to a read condition and 849 to a generate condition. Hence, within the 310 experiments, 804 different comparisons (between a read and one or multiple generate conditions) were examined. The estimated mean proportions, 95% confidence intervals, and number of comparisons for each of the models tested are reported in Tables 1 and 2. We report results in the following order: First, we report the results from the model examining the presence of publication bias. Second, we report the overall generation effect across all studies (generate compared with read). Third, we report the models showing whether the different item memory theories (mental effort, selective displaced rehearsal, etc.) and context memory theories (associative strengthening/item–context trade-off, processing account) are supported by existing empirical work. Fourth, we report two models testing the influence of generation constraint as a significant moderator of the generation effect across all the included studies in this meta-analysis. Fifth, we assess other factors known to influence the generation effect (e.g., type of generation task, retention delay).

Next, we report the results from each of our models. The models are numbered by the order in which they are reported here and in Table 1.

## Publication bias

In the publication bias model, the interaction between the inverse sample size and the condition type factor just failed to be significant,  $t(122) = 1.81$ ,  $p = .07$ . Still, larger samples were associated with smaller generation effects. For example, for sample sizes of 8, 24, and 120 participants per condition (corresponding to the minimum, median, and maximum sample size), estimated effects were .142 (95% CI [.100, .184]), .101 (95% CI [.086, .115]), and .084 (95% CI [.056, .111]), respectively. When extrapolating to an infinite sample size, the generation effect was reduced to .080 (95% CI [.048, .111]), which is still significant,  $t(122) = 5.00$ ,  $p < .001$ . Therefore, although we are not willing (or able) to rule out the presence of publication bias altogether, we would cautiously suggest that its possible presence is unlikely to be a major factor in the present data.

## Generation effect

**1. Generation effect** Our model testing the standard generation effect compared the mean proportions of memory for generate compared with read materials across the full sample of studies including both item and context memory results. We expected that generate materials would be better remembered than read materials would, consistent with the standard generation effect. This model showed that memory performance was indeed better for generate compared with read items, with an estimated mean difference ( $M_{\text{diff}}$ ) of .102 (95% CI [.087,

.117]),  $t(124) = 13.85$ ,  $p < .001$ , suggesting that the generation effect is generally robust across various experimental parameters.

## Item memory theories

The following set of analyses pertaining to item memory theories were conducted based on the subset of rows reporting item memory mean proportions ( $k = 1,284$ ). Note that analyses might be based on fewer estimates if the values for particular moderator variables were missing for some cases.

**2. Mental effort** To test the mental effort theory, we ran a model testing the interaction between *condition type* (read, generate) and *generation difficulty* (low difficulty, high difficulty), only for studies that included a difficulty manipulation ( $k = 27$ ). The mental effort theory predicts a larger generation effect for high-difficulty compared with low-difficulty generation. Our model showed that generate items were better remembered than read items for both low-difficulty ( $M_{\text{diff}} = .177$ , 95% CI [.133, .221]),  $t(23) = 8.36$ ,  $p < .001$ , and high-difficulty ( $M_{\text{diff}} = .171$ , 95% CI [.111, .221]),  $t(23) = 10.12$ ,  $p < .001$ , generation tasks. Critically, there was no interaction,  $F(1, 23) = 0.04$ ,  $p = .839$ , suggesting that the magnitude of the generation effect did not significantly differ between low-difficulty and high-difficulty generation tasks. These findings do not support the predictions of the mental effort theory.

**3. Selective displaced rehearsal** We ran two models to assess the selective displaced rehearsal theory,<sup>7</sup> which predicts a generation effect for within-subjects and mixed-list designs, but no generation effect for between-subject sand pure-list designs.

**3a. Condition type by manipulation type** The first model (3a) tested the interaction between *condition type* (read, generate) and *manipulation type* (within subjects, between subjects). This model revealed a significant generation effect for both within-subjects manipulations ( $M_{\text{diff}} = .144$ , 95% CI [.126, .161]),  $t(119) = 16.24$ ,  $p < .001$ , and between-subjects manipulations ( $M_{\text{diff}} = .073$ , 95% CI [.031, .114]),  $t(119) = 7.58$ ,  $p < .001$ . The Condition Type  $\times$  Manipulation Type interaction coefficient was also significant,  $F(1, 119) = 34.27$ ,  $p < .001$ , revealing that between-subjects manipulations led to a significantly smaller generation effect compared with within-subjects manipulations. Although we found a reduced generation effect for between-subjects compared with within-

<sup>7</sup> Two models were necessary to test this theory because the theory makes predictions about the size of the generation effect for within-subjects versus between-subject designs (Model 3a), and for mixed-list versus pure-list presentations (Model 3b). One model tests within-subjects versus between-subjects manipulation types, while the second model tests mixed-list versus pure-list presentation types.

**Table 1.** Estimated mean proportions ( $M_{est}$ ), 95% confidence intervals (CI), and number of comparisons ( $k$ ), by model

<b>Model</b>					
Moderator level		95% CI			
Interaction moderator level	$M_{est}$	LL	UL	Test	$k$
<b>1. Generation effect (condition type)</b>					
Read	0.478	0.446	0.511	—	804
Generate	0.581	0.549	0.612	$t(124) = 13.85, p < .001$	849
				$F(1, 23) = 0.04, p = .839$	
<b>2. Mental effort (Condition Type × Generation Difficulty)</b>					
Low difficulty					
Read	0.473	0.393	0.553	—	71
Generate	0.650	0.583	0.716	$t(24) = 8.36, p < .001$	73
High difficulty					
Read	0.464	0.395	0.533	—	60
Generate	0.635	0.560	0.711	$t(24) = 10.12, p < .001$	73
				$F(1, 119) = 34.27, p < .001$	
<b>3a. Selective displaced rehearsal (Condition Type × Manipulation Design)</b>					
Within subjects					
Read	0.456	0.419	0.494	—	444
Generate	0.600	0.562	0.638	$t(119) = 16.24, p < .001$	468
Between subjects					
Read	0.500	0.451	0.549	—	182
Generate	0.573	0.522	0.623	$t(119) = 7.58, p < .001$	190
				$F(1, 118) = 8.51, p = .004$	
<b>3b. Selective displaced rehearsal (Condition Type × List Presentation)</b>					
Mixed-list presentation					
Read	0.448	0.406	0.489	—	328
Generate	0.591	0.550	0.632	$t(118) = 15.48, p < .001$	337
Pure (blocked)-list presentation					
Read	0.494	0.456	0.533	—	294
Generate	0.597	0.555	0.638	$t(121) = 8.93, p < .001$	317
				$F(2, 117) = 57.86, p < .001$	
<b>4. Semantic activation (Condition Type × Stimuli Type)</b>					
Words					
Read	0.471	0.435	0.506	—	553
Generate	0.603	0.567	0.640	$t(117) = 16.86, p < .001$	585
Nonwords					
Read	0.439	0.377	0.500	—	41
Generate	0.443	0.368	0.518	$t(117) = 0.29, p = .770$	41
Numbers					
Read	0.392	0.296	0.489	—	32
Generate	0.600	0.518	0.682	$t(117) = 16.24, p < .001$	32
				$F(2, 116) = 5.34, p = .006$	
<b>5. Multifactor theory (Condition Type × Memory Test)</b>					
Recognition					
Read	0.622	0.581	0.663	—	251
Generate	0.767	0.733	0.800	$t(116) = 12.35, p < .001$	261
Cued recall					

**Table 1.** (continued)

<b>Model</b>		95% CI			Test	<i>k</i>	
Moderator level	<i>M</i> <sub>est</sub>	LL	UL				
Interaction moderator level							
Read	0.527	0.463	0.590	—		135	
Generate	0.659	0.593	0.725	<i>t</i> (116) = 8.98, <i>p</i> < .001		147	
Free recall							
Read	0.268	0.234	0.303	—		236	
Generate	0.363	0.326	0.401	<i>t</i> (116) = 8.54, <i>p</i> < .001		246	
<b>6. Associative strengthening</b>							
<b>(Condition type – Context memory data only)</b>							
Read	0.606	0.557	0.656	—		169	
Generate	0.638	0.588	0.689	<i>t</i> (41) = 2.10, <i>p</i> = .042		180	
<b>7. Item–context trade-off</b>							
<b>(Condition Type × Memory Type)</b>							
Item memory						<i>F</i> (1, 122) = 32.45, <i>p</i> < .001	
Read	0.468	0.434	0.502	—		626	
Generate	0.591	0.556	0.626	<i>t</i> (122) = 15.87, <i>p</i> < .001		658	
Context memory							
Read	0.525	0.472	0.578	—		169	
Generate	0.557	0.496	0.618	<i>t</i> (122) = 2.11, <i>p</i> = .037		180	
<b>8. Processing account</b>							
<b>(Condition Type × Context Memory Type)</b>							
Conceptual context memory						<i>F</i> (1, 34) = 49.71, <i>p</i> < .001	
Read	0.616	0.564	0.668	—		111	
Generate	0.696	0.653	0.739	<i>t</i> (34) = 4.76, <i>p</i> < .001		119	
Perceptual context memory							
Read	0.594	0.533	0.655	—		40	
Generate	0.541	0.485	0.597	<i>t</i> (34) = -5.75, <i>p</i> < .001		43	
<b>9a. Generation constraint</b>							
						<i>F</i> (2, 121) = 4.89, <i>p</i> = .009	
Lower const. generate	0.642	0.595	0.690	—		207	
Medium const. generate	0.586	0.549	0.622	<i>t</i> (121) = -2.36, <i>p</i> = .020		401	
Higher const. generate	0.528	0.471	0.585	<i>t</i> (121) = -3.04, <i>p</i> = .003		226	
<b>9b. Generation Constraint × Memory Test</b>							
Recognition						<i>F</i> (2, 113) = 0.42, <i>p</i> = .659	
Lower const. generate	0.746	0.686	0.805	—		107	
Medium const. generate	0.734	0.695	0.773	<i>t</i> (113) = -0.32, <i>p</i> = .750		180	
Higher const. generate	0.706	0.645	0.767	<i>t</i> (113) = -0.83, <i>p</i> = .407		119	
Cued recall						<i>F</i> (2, 113) = 3.78, <i>p</i> = .026	
Lower const. generate	0.717	0.638	0.795	—		21	
Medium const. generate	0.642	0.578	0.705	<i>t</i> (113) = -2.49, <i>p</i> = .014		97	
Higher const. generate	0.567	0.435	0.699	<i>t</i> (113) = -2.04, <i>p</i> = .044		32	
Free recall						<i>F</i> (2, 113) = 9.40, <i>p</i> < .001	
Lower const. generate	0.484	0.407	0.561	—		63	
Medium const. generate	0.354	0.300	0.409	<i>t</i> (113) = -2.61, <i>p</i> = .010		118	
Higher const. generate	0.295	0.253	0.337	<i>t</i> (113) = -4.25, <i>p</i> < .001		75	

*Note.* *F* tests reflect either interaction tests or omnibus tests of multiple coefficients, whereas *t* tests reflect contrasts of two conditions with the reference level indicated by a dash (—). Each model is numbered and listed in the same order as presented in the results. LL = lower limit; UL = upper limit

**Table 2.** Estimated mean proportions ( $M_{est}$ ), 95% confidence intervals (CI), and number of comparisons ( $k$ ), of condition type by moderator

<b>Moderator</b>		95% CI			Test	$k$
Moderator level(s)	$M_{est}$	LL	UL			
<b>Learning type</b>					$F(1, 121) = 5.63, p = .019$	
Incidental						
Read	0.458	0.414	0.501	—		240
Generate	0.583	0.541	0.626	$t(121) = 11.45, p < .001$		252
Intentional						
Read	0.489	0.453	0.526	—		553
Generate	0.582	0.547	0.617	$t(121) = 10.32, p < .001$		587
<b>Stimuli relation (for word pair stimuli)</b>					$F(7, 69) = 8.24, p < .001$	
Semantic						
Read	0.493	0.441	0.546	—		172
Generate	0.617	0.564	0.671	$t(69) = 7.92, p < .001$		205
Category exemplars						
Read	0.527	0.476	0.578	—		96
Generate	0.619	0.554	0.684	$t(69) = 4.68, p < .001$		101
Antonym						
Read	0.430	0.333	0.528	—		72
Generate	0.535	0.446	0.624	$t(69) = 5.23, p < .001$		77
Synonym						
Read	0.425	0.314	0.536	—		21
Generate	0.594	0.460	0.727	$t(69) = 7.33, p < .001$		21
Rhyme						
Read	0.423	0.355	0.491	—		109
Generate	0.495	0.439	0.552	$t(69) = 4.81, p < .001$		109
Compound words						
Read	0.497	0.376	0.617	—		5
Generate	0.559	0.439	0.678	$t(69) = 121.30, p < .001$		5
Definitions						
Read	0.482	0.285	0.68	—		16
Generate	0.678	0.523	0.833	$t(69) = 3.84, p < .001$		16
Unrelated						
Read	0.441	0.322	0.561	—		13
Generate	0.504	0.395	0.612	$t(69) = 2.85, p = .006$		14
<b>Mode of Generation</b>					$F(2, 120) = 2.70, p = .072$	
Verbal/Speaking						
Read	0.463	0.417	0.510	—		349
Generate	0.586	0.541	0.631	$t(120) = 9.71, p < .001$		390
Covert/Thinking						
Read	0.517	0.461	0.571	—		85
Generate	0.589	0.530	0.647	$t(120) = 2.74, p = .007$		87
Writing/Typing						
Read	0.481	0.428	0.533	—		365
Generate	0.572	0.520	0.624	$t(120) = 11.30, p < .001$		369
<b>Generation task</b>					$F(6, 109) = 12.28, p < .001$	
Word stem						

**Table 2.** (continued)

<b>Moderator</b>					
Moderator level(s)		95% CI			
Condition type	$M_{est}$	LL	UL	Test	$k$
Read	0.473	0.424	0.523	—	255
Generate	0.574	0.527	0.622	$t(109) = 9.93, p < .001$	266
<b>Anagram</b>					
Read	0.471	0.413	0.529	—	117
Generate	0.584	0.530	0.639	$t(109) = 8.83, p < .001$	117
<b>Letter transposition</b>					
Read	0.507	0.432	0.581	—	67
Generate	0.540	0.454	0.627	$t(109) = 1.84, p = .072$	67
<b>Word fragment</b>					
Read	0.469	0.419	0.519	—	251
Generate	0.570	0.523	0.616	$t(109) = 5.77, p < .001$	260
<b>Sentence Completion</b>					
Read	0.539	0.442	0.636	—	40
Generate	0.606	0.483	0.729	$t(109) = 2.76, p = .007$	40
<b>Calculation</b>					
Read	0.388	0.286	0.491	—	34
Generate	0.600	0.514	0.687	$t(109) = 14.68, p < .001$	34
<b>Cue only</b>					
Read	0.509	0.403	0.614	—	20
Generate	0.657	0.593	0.721	$t(109) = 3.47, p < .001$	45
<b>Divided attention</b>				$F(1, 122) = 0.83, p = .364$	
<b>Divided attention</b>					
Read	0.336	0.282	0.390	—	18
Generate	0.454	0.371	0.537	$t(122) = 7.43, p < .001$	18
<b>Full attention</b>					
Read	0.481	0.449	0.514	—	786
Generate	0.583	0.552	0.614	$t(122) = 13.43, p < .001$	831
<b>Presentation pacing</b>				$F(1, 122) = 10.91, p = .001$	
<b>Self-paced</b>					
Read	0.481	0.411	0.551	—	141
Generate	0.640	0.583	0.696	$t(122) = 7.96, p < .001$	171
<b>Timed</b>					
Read	0.476	0.439	0.514	—	658
Generate	0.565	0.530	0.601	$t(122) = 12.13, p < .001$	673
<b>Filler task</b>				$F(1, 122) = 0.22, p = .642$	
<b>Filler task (&gt;1 min)</b>					
Read	0.478	0.441	0.514	—	463
Generate	0.577	0.539	0.614	$t(122) = 11.46, p < .001$	472
<b>No filler task</b>					
Read	0.480	0.432	0.528	—	341
Generate	0.586	0.541	0.630	$t(122) = 8.76, p < .001$	377
<b>Subject population</b>				$F(1, 122) = 1.80, p = .182$	
<b>Younger adults</b>					

**Table 2.** (continued)

<b>Moderator</b>		95% CI			Test	<i>k</i>
Moderator level(s)	<i>M</i> <sub>est</sub>	LL	UL			
Condition type						
Read	0.485	0.452	0.518	—		763
Generate	0.586	0.554	0.617	<i>t</i> (122) = 13.19, <i>p</i> < .001		805
<b>Older adults</b>						
Read	0.344	0.307	0.381	—		41
Generate	0.473	0.421	0.525	<i>t</i> (122) = 6.51, <i>p</i> < .001		44
<b>Retention delay</b>				<i>F</i> (2, 120) = 3.19, <i>p</i> = .045		
<b>Immediate</b>						
Read	0.479	0.443	0.516	—		677
Generate	0.579	0.544	0.614	<i>t</i> (120) = 12.63, <i>p</i> < .001		719
<b>Short (5 minutes–1 day)</b>						
Read	0.487	0.418	0.556	—		87
Generate	0.580	0.495	0.665	<i>t</i> (120) = 4.87, <i>p</i> < .001		87
<b>Long (&gt;1 day)</b>						
Read	0.444	0.351	0.536	—		40
Generate	0.608	0.517	0.700	<i>t</i> (120) = 6.09, <i>p</i> < .001		43
<b>Number of Stimuli</b>				<i>F</i> (2, 120) = 9.17, <i>p</i> = .001		
<b>25 or fewer</b>						
Read	0.466	0.426	0.506	—		421
Generate	0.588	0.552	0.624	<i>t</i> (120) = 12.95, <i>p</i> < .001		441
<b>26–50</b>						
Read	0.481	0.445	0.517	—		291
Generate	0.570	0.532	0.609	<i>t</i> (120) = 8.63, <i>p</i> < .001		317
<b>More than 50</b>						
Read	0.523	0.467	0.580	—		92
Generate	0.579	0.518	0.641	<i>t</i> (120) = 4.26, <i>p</i> < .001		91

*Note.* *F* tests reflect interaction tests, whereas *t* tests reflect contrasts of two conditions with the reference level indicated by a dash (—). Each model is numbered and listed in the same order as presented in the results. LL = lower limit; UL = upper limit

subjects designs, finding a significant generation effect for between-subjects designs argues against the selective displaced rehearsal theory, which predicts no generation effect in between-subjects designs.

**3b. Condition type by presentation type** The second model (3b) tested the interaction of *condition type* (read, generate) and *presentation type* (mixed list, pure list). Selective displaced rehearsal theory predicts a generation effect for mixed-list presentation, but no generation effect for pure-list presentations. This model showed a significant generation effect for both mixed-list presentations ( $M_{\text{diff}} = .144$ , 95% CI [.125, .162]),  $t(118) = 15.48$ ,  $p < .001$ , and pure-list presentations ( $M_{\text{diff}} = .103$ , 95% CI [.071, .149]),  $t(118) = 8.93$ ,  $p < .001$ . The Condition Type  $\times$  Presentation Type interaction coefficient was also significant,  $F(1, 118) = 8.51$ ,  $p = .004$ ,

revealing that the generation effect was significantly smaller for pure-list presentations compared with mixed-list presentations. Similar to our first model (within subjects vs. between subjects), however, finding a significant generation effect for pure-list designs argues against the selective displaced rehearsal theory. Together, these two models do not support the selective displaced rehearsal theory as a viable account of the generation effect.

**4. Semantic (lexical) activation** To assess the semantic activation theory, we ran a model testing the interaction of *condition type* (read, generate) and *word status* (words, nonwords, numbers). This theory predicts a generation effect for meaningful stimuli (words and numbers), but no generation effect for nonmeaningful stimuli (nonwords). This model showed a significant generation effect when words were used as stimuli

( $M_{diff} = .133$ , 95% CI: .117 to .149),  $t(117) = 16.86$ ,  $p < .001$ , and when numbers were used ( $M_{diff} = .208$ , 95% CI [.162, .253]),  $t(117) = 16.24$ ,  $p < .001$ . When nonwords were used as stimuli, however, we found no differences between the generate and read conditions (i.e., no generation effect;  $M_{diff} = .004$ , 95% CI [-.045, .053]),  $t(117) = 0.29$ ,  $p = .770$ . The Condition Type  $\times$  Word Status interaction was significant,  $F(2, 117) = 57.86$ ,  $p < .001$ . Follow-up analyses revealed that studies using numbers as stimuli showed a significantly larger generation effect compared with studies using words,  $t(117) = 4.96$ ,  $p < .001$ , whereas studies using nonword stimuli showed a significantly reduced generation effect compared with studies using words,  $t(117) = -7.79$ ,  $p < .001$ . This model provides support for the semantic activation theory, suggesting meaningful information (words and numbers) may be a necessary condition to obtain a generation effect.

**5. Two-factor/multifactor** The model assessing the multifactor theory tested the interaction between *condition type* (read, generate) and *memory test* (recognition, cued recall, free recall). The multifactor theory predicts that self-generating improves memory by enhancing both item-specific and relational processing, relative to reading. Finding a generation effect using different memory tests, which are assumed to assess different types of memory representations (item, relational), would support this theory. This model revealed a significant Condition Type  $\times$  Memory Test interaction coefficient,  $F(2, 116) = 5.34$ ,  $p = .006$ , which was driven by a larger generation effect for recognition ( $M_{diff} = .144$ , 95% CI [.121, .167]) and cued recall memory tests ( $M_{diff} = .132$ , 95% CI [.076, .189]) compared with free recall tests ( $M_{diff} = .095$ , 95% CI [.042, .148]). Importantly, we found that despite differences in magnitude, all three types of memory tests showed robust generation effects ( $ts > 8.54$ ,  $ps < .001$ ). Prior work has suggested that different memory tests can be used to assess different types of processing (e.g., recognition tests are sensitive to item-specific processing, and cued recall and free recall are sensitive to cue–target and intertarget relational processing, respectively). Finding a significant generation effect across these three memory tests generally supports the multifactor theory suggesting that self-generation improves memory relative to read controls through both item-specific and relational processing. Given that free recall tests showed a smaller generation effect, however, it may be that intertarget relational processing experiences a smaller boost from generation compared with item-specific and cue–target relational processing.

## Context memory theories

**6. Associative strengthening** To examine the associative strengthening account, we ran a model including the *condition type* (read, generate) variable as the lone moderator, but we restricted this analysis to only include data from context

memory measures ( $k = 349$  estimates). This model showed that generate tasks led to a small but significant increase for context memory compared with read controls, with an estimated mean difference ( $M_{diff}$ ) of .032 (95% CI [.001, .063]),  $t(41) = 2.10$ ,  $p = .042$ . This result suggests that self-generation does improve memory for contextual details relative to controls, supporting the associative strengthening account which claims that generation improves memory for context memory by binding episodic contextual details into a coherent memory representation.

**7. Item–context trade-off** To examine the item–context trade-off account, we ran a model testing the interaction between *condition type* (read, generate) and *memory type* (item, context). This model showed a significant generation effect for both item ( $M_{diff} = .123$ , 95% CI [.108, .138]),  $t(122) = 15.87$ ,  $p < .001$ , and context memory ( $M_{diff} = .032$ , 95% CI [.017, .047]),  $t(122) = 2.11$ ,  $p = .037$ . The Condition Type  $\times$  Memory Type interaction coefficient was significant,  $F(1, 122) = 32.45$ ,  $p < .001$ , indicating that, perhaps unsurprisingly, the generation effect was significantly larger for item memory compared with context memory. Despite the reduced effect, finding a significant generation effect for context memory argues against a strong view of the item–context trade-off account, which posits that generation requires cognitive resources that improve item memory, but as a consequence leave fewer resources to encode contextual information.

**8. Processing account** To assess the processing account, we used a model examining the interaction between *condition type* (read, generate) and *context memory type* (conceptual context, perceptual context). This model showed a significant Condition Type  $\times$  Context Memory Type interaction,  $F(1, 34) = 49.71$ ,  $p < .001$ . Follow-up analyses revealed that for *conceptual* contextual details, generate conditions led to better memory than reading (i.e., a generation effect;  $M_{diff} = .080$ , 95% CI [.046, .114]),  $t(34) = 4.76$ ,  $p < .001$ ; however, for *perceptual* contextual details, the opposite pattern occurred: read controls led to better memory performance than generate tasks did ( $M_{diff} = -.053$ , 95% CI [-.125, .020]),  $t(34) = -5.75$ ,  $p < .001$ . This pattern of results strongly supports the processing account, which claims generation tasks only improve memory for contextual details that are conceptual in nature because of the match in conceptual processing at encoding and test, in accord with transfer-appropriate processing principles.

## Generation constraint

**9a. Generation constraint** Our next model tested the hypothesis that generation constraint affects the magnitude of the memory benefits from self-generation. In this model, we included the *generation constraint* (lower, medium, higher) variable as the lone predictor. This model showed significant

differences between levels of generation constraint,  $F(2, 121) = 4.89, p = .009$ . Follow-up analyses revealed that lower constraint generation led to significantly better memory than both medium-constraint and higher-constraint generation ( $t_s < -2.36, p_s < .020$ ). Additionally, medium-constraint generation led to marginally better memory than higher-constraint generation,  $t(121) = 1.89, p = .06$ . This finding indicates that generation constraint strongly influences memory for self-generated information and is a factor that should be considered in future work.

**9b. Generation constraint by memory test** Research has also shown that the effects of generation constraint may depend on the type of memory test used (Fiedler et al., 1992; McCurdy et al., 2017; McCurdy et al., 2020). Given that different memory tests tap into different types of memory representations, understanding the effects of generation constraint for different memory tests could be important to understanding the memory mechanism(s) underlying this influential factor. Therefore, the next model tested the effect of *generation constraint* (lower, medium, higher) for each *memory test* (recognition, cued recall, free recall) separately. For recognition tests, we found no significant differences between levels of constraint,  $F(2, 113) = 0.42, p = .659$ . This finding is in line with previous work showing generation constraint is less robust on item recognition measures (McCurdy et al., 2017; McCurdy et al., 2020). For cued recall, there were significant differences between levels of constraint,  $F(2, 113) = 3.78, p = .026$ . Follow-up comparisons revealed that lower-constraint generation led to better cued recall memory than both medium-constraint and higher-constraint generation ( $t_s < -2.04, p_s < .044$ ), while the difference between medium-constraint and higher-constraint generation was not significant,  $t(113) = -1.10, p = .27$ . For free recall, we also found significant differences between levels of constraint,  $F(2, 113) = 9.40, p < .001$ . Follow-up comparisons showed that lower constraint led to better free recall memory performance than both medium and higher constraint ( $t_s < -2.61, p_s < .010$ ), while medium constraint led to marginally better memory compared with higher constraint,  $t(113) = -1.74, p = .08$ . Overall, this model indicates that the influence of generation constraint is strongest for measures of cued and free recall. Given that recall tests are thought to be sensitive to relational processing (Burns, 2006; Hirshman & Bjork, 1988; McDaniel et al., 1988), these findings provide some evidence that fewer generation constraints increase the magnitude of the generation effect through enhanced relational processing, in line with prior work (McCurdy et al., 2020).

### Experimental design moderators

To examine the influence of different experimental moderators on the generation effect, we ran a model interacting

*condition type* (read, generate) with each of the 10 remaining moderators included in our coding scheme that were not tested as part of a theory-based model from above (e.g., *learning type, stimuli relation, generation mode*). Our goal with this analysis was to identify important methodological factors that influence the magnitude of the generation effect to help guide future work. Table 2 reports the estimated mean proportions, 95% confidence intervals, and number of comparisons for each model. In this section, we report the statistics only for moderators that were found to significantly influence the magnitude of the generation effect: *learning type, stimuli relation, mode of generation, generation task, presentation pacing, retention delay, and number of stimuli*. For each moderator with more than two levels, the baseline (comparison group) was selected as the most commonly occurring manipulation for that moderator (e.g., semantic was selected as the baseline group for *stimuli relation* because we included  $k = 257$  studies using a semantic stimuli relation, which was more than any other stimuli relation type).

The model of *learning type* (incidental, intentional) showed that incidental learning produced a larger generation effect (generate > read) compared with intentional learning,  $t(121) = 2.37, p = .019$ . The model of *stimuli relation* (semantic, category exemplars, antonym, synonym, rhyme, compound words, definitions, unrelated) revealed that, compared with the most commonly used relation (semantic), rhyming,  $t(64) = -2.33, p = .023$ , compound words,  $t(64) = -3.93, p < .001$ , and unrelated,  $t(64) = -2.19, p = .032$ , all showed significantly smaller generation effects, suggesting that coming up with a semantic associate seems to lead to the largest generation effect. The model of *mode of generation* (verbal/speaking, covert/thinking, writing/typing) revealed that compared with verbal generation, covert generation,  $t(120) = -1.73, p = .085$ , led to marginally smaller, and writing generation procedures,  $t(120) = -2.06, p = .041$ , led to significantly smaller generation effects. The model of *generation task* (word stem, anagram, letter transposition, word fragment, sentence completion, calculation, cue only) revealed that compared with the most commonly used task (word stem), calculation tasks led to significantly larger generation effects,  $t(110) = 6.25, p < .001$ , while letter transposition tasks led to significantly smaller generation effects,  $t(110) = -3.04, p = .003$ . All other tasks (e.g., anagram, word fragment, sentence completion, cue only) were similar in magnitude as word stem tasks ( $t_s < 1.27, p_s > .206$ ). The model of *presentation pacing* (self-paced, timed) showed that self-paced presentation led to a larger generation effect compared with timed presentation,  $t(121) = 3.30, p = .001$ . The model of *retention delay* (immediate, short [5 minutes–1 day], long [ $>1$  day]) revealed that long retention delays ( $>1$  day) led to a larger generation effect compared with immediate and shorter delays ( $t_s > 2.35, p_s < .020$ ). The model examining *number of stimuli* (25 or fewer, 26–50, more than 50) indicated that studies using 25 or less



stimuli per condition led to a significantly larger generation effect than studies using both 26–50 and greater than 50 stimuli ( $t_s > 2.58$ ,  $p_s < .011$ ). Additionally, the generation effect was larger for studies with 26–50 stimuli per condition compared with studies with greater than 50,  $t(120) = 2.15$ ,  $p = .03$ .

## Discussion

In this meta-analysis and theoretical review, we used a mixed-effects multilevel model approach to analyze 126 generation-effect articles to test prominent theories on the generation effect and to examine an underrecognized factor (generation constraint) on the generation effect. This is the largest and most comprehensive meta-analysis on the generation effect to date, and the first meta-analysis to examine context memory within the generation effect. The results of this meta-analysis yielded four main findings. First, we confirmed that the generation effect (better memory for generated compared with read materials) is a robust memory effect that occurs in a wide range of experimental designs. Across all studies, including both item and context memory measures, we found that self-generation provided roughly a 10 percentage point increase in memory performance relative to reading. Second, we tested seven theories that have been proposed to explain the generation effect to assess which theories are supported by the existing data on the generation effect. We found support for four of the seven theories tested, two item memory accounts (the multifactor theory and semantic activation theory), and two context memory accounts (the processing account and associative strengthening account). Third, we examined the extent that generation constraint is a significant moderator of the memory benefits from self-generation, and found that this factor has a sizeable effect on memory for self-generated materials. Fourth, we found evidence of several other experimental moderators that significantly influence the magnitude of the generation effect. In this discussion, we offer our assessment of each generation effect theory tested in this meta-analysis (item memory theories followed by context memory). We start with those theories most strongly supported by our results, followed by the theories that are only partially supported or unsupported. Then, we discuss generation constraint as a critical moderator of the generation effect, followed by a discussion of other experimental moderators shown to influence the magnitude of the generation effect. Additionally, we suggest a future direction for theoretical work on the generation effect.

### Item memory theories

One of the most prominent item memory theories of the generation effect, the multifactor theory (Hirshman & Bjork, 1988; McDaniel et al., 1988), proposes that self-generation improves memory relative to reading via two memory

mechanisms: enhanced item-specific processing (i.e., processing of the features of the target item) and enhanced relational processing (i.e., processing of cue–target or intertarget relational information). This meta-analysis further supports this theory as a reliable explanation of how generation strengthens memory. Early in the history of generation effect research, investigators debated whether either factor alone could account for the generation effect, but as more data accumulates, single-factor theories struggle to fully account for the effect (Begg et al., 1989; Burns, 1990). The results of this meta-analysis provide clear and converging evidence that generation indeed enhances both types of processes (item specific and relational). Interestingly though, we found a larger generation effect for recognition and cued recall tests compared with free recall tests, suggesting that item-specific and cue–target relational processing may be improved to a greater extent than intertarget relational processing. Free recall tests rely primarily on intertarget relationships to perform well, and prior work suggests that the generation effect is often reduced when using free recall measures because most generation tasks do not promote focus on relationships across trials (deWinstanley & Bjork, 1997; Hunt & Einstein, 1981; McDaniel et al., 1990; McDaniel et al., 1988). Overall, the evidence in this meta-analysis supports the idea that generation engages multiple types of processing, which in turn improves memory.

Although the multifactor theory was supported, it still cannot fully account for all of the findings that exist on the generation effect. For example, this theory does not specify that different generation tasks might lead to differences in the relative amount of item-specific or relational processing, leading to differential memory benefits across various generation tasks, as more recent work on the generation effect has shown (McCurdy et al., 2017; McCurdy et al., 2020). Thus, although the multifactor theory represents the best theory to date, there is still variability in memory findings across generation studies left unaccounted for by this theory. It is also important to note that although the multifactor theory garnered the strongest support of the theories included in this meta-analysis, it does not necessarily preclude other theories from having explanatory power for the generation effect. For instance, the multifactor theory relies on encoding task processes to explain the generation effect. Differences in encoding task processes, however, do not account for other boundary conditions of the generation effect, such as the reduced effect for between-subjects designs (selective displaced rehearsal theory).

Both the semantic activation and selective displaced rehearsal theories posit simple methodological conditions that account for variation in the generation effect. The semantic activation theory proposes the generation effect only occurs for meaningful stimuli (Gardiner & Hampton, 1985; McElroy & Slamecka, 1982). In line with this idea, we found no evidence of a generation effect when nonwords (excluding

numbers) were used as stimuli. Interestingly, we found a numerically larger memory benefit for studies that used numbers as stimuli compared with words, suggesting that future investigations studying the generation effect using numbers as stimuli may be a fruitful way to better understand the mechanisms underlying the generation effect. One interesting implication of the semantic activation theory is the application of the generation effect for learning in educational settings. Given that learning involves acquiring information that one did not have before (i.e., is meaningless), some have argued that generation may not be an effective learning strategy for new knowledge (Lutz, Briggs, & Cain, 2003). More recent research attempting to apply the generation effect to educational settings, however, has shown that generation can be beneficial to learning new information, particularly when feedback is given to correct errors (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Metcalfe & Kornell, 2007; Potts & Shanks, 2014; Richland, Bjork, Finley, & Linn, 2005). Specifically, Metcalfe and Kornell (2007) showed that generation improved the learning of new information compared with simply reading the information, even when students (both 6th-grade and college) initially generated an incorrect response. The key is providing feedback to correct these initial errors. Similarly, research on “errorful-generation” shows that generating errors with feedback improves memory for the to-be-learned information over simply reading (or studying) that information (Potts & Shanks, 2014). This work highlights a critical application of the generation effect in educational and other learning environments as a useful learning technique or study strategy. As for the semantic activation theory, the findings on generating errors seems to indicate that even if the material is originally meaningless, if the material can be incorporated into one’s existing knowledge structure, self-generating can provide memory benefits over reading.

We also tested the mental effort theory (Tyler et al., 1979), but found no memory advantage for high-difficulty generation over low-difficulty generation, arguing against mental effort as an explanation of the generation effect. This finding is in line with more recent reports suggesting that the generation effect cannot be explained by mental effort alone (Foley & Foley, 2007; Foley, Foley, Wilder, & Rusch, 1989; Hertel, 1989). Instead, these studies have suggested it is not the *amount* of processing required to generate, but rather the *type* of processing required (cf. the multifactor theory; Hirshman & Bjork, 1988) and its overlap with most memory tests that leads to the memory advantage. In other words, the manner in which a stimulus is encoded seems more predictive of later memory performance compared with the amount of effort involved at encoding, an idea that extends to other memory phenomena as well (Mitchell & Hunt, 1989). Before ruling out the cognitive effort theory completely, however, it should be acknowledged that many of the existing studies examining cognitive effort as an explanation for the generation effect

suffer from the lack of a universal measurement and manipulation of effort. In testing this theory, we relied on the author’s judgment of cognitive effort through difficulty manipulations, but this classification is not without limitations. Kahneman (1973) originally suggested that to examine cognitive effort, researchers should use measures that are independent of the outcome variable, such as performance on a secondary task. The data we examined in this meta-analysis however, often relied on various task difficulty manipulations only assumed to influence the amount of mental effort required. Thus, future research examining the differences of generating versus reading on an independent measure of effort might be necessary to completely rule out the mental effort explanation of the generation effect.

### Context memory theories

Mulligan’s (2011) processing account provides a thoughtful and nuanced explanation of the mixed findings on the generation effect for context memory, proposing that self-generation and reading require different encoding processes, which in turn lead to differential context memory effects (a generation effect for conceptual details like source and cue words that receive more processing when generating, but not perceptual details like font color and font type that receive more processing when reading). In the present study, we found evidence that strongly supports this idea, where generation tasks resulted in better memory for conceptual context details, but reading led to better memory for perceptual details. Based on our context memory findings, the processing account seems to be a viable explanation for the conflicting results found for the generation effect on context memory. Although this seems to be the best explanatory model of context memory, additional work may be needed because there are fewer context memory relative to item memory investigations. Indeed, Nieznański (2011, 2012, 2013) has developed an interesting line of work, studying how processing demands and generation difficulty might affect how self-generation improves context (source) memory. Relatedly, other work investigating the influence of generation constraint on context memory (McCurdy et al., 2017; McCurdy et al., 2020) could be fruitful in advancing understanding of how generation affects memory for contextual details.

We tested two other context memory theories: associative strengthening and item–context trade-off. These contrasting theories developed as the findings in the literature for context memory grew more mixed (Jurica & Shimamura, 1999; Marsh, 2006; Marsh et al., 2001; Mulligan & Lozito, 2004). In this meta-analysis, we found support for the associative strengthening account, but not for the item–context trade-off account. Despite finding support for the associative strengthening theory, our context memory models as a whole suggest that a general associative strengthening or trade-off account is

likely not precise enough to describe the complexity of the generation effect for context memory. Our more nuanced model testing the processing account showed that not all context memory types are improved by generation, arguing against a general enhancement for context memory (as the associative strengthening account suggests). This idea that generation improves some types of context memory, but not others, is supported not only by Mulligan's (2011) work on the processing account, but other empirical work as well (Nieznański, 2012; Overman, Richard, & Stephens, 2017). Thus, future research aimed at clarifying which types of contextual details are improved, not affected, or worsened by self-generation would help advance our understanding the how generation influences context memory.

### Generation constraint

Another major goal of this meta-analysis was to examine the effect of generation constraint on the magnitude of the generation effect. The variety of generation tasks used throughout prior work make this variable important to consider, since it has not been well scrutinized. Indeed, few attempts have been made to provide a framework to understand how differences in generation tasks might influence memory, which is surprising, given the many ways generation has been manipulated in the extant literature (as indicated by the seven types of generation tasks we coded in this meta-analysis). Across all studies included in this meta-analysis (item and context memory), we found that lower-constraint generation tasks led to better memory than both medium-constraint and higher-constraint generation tasks. This finding supports the limited empirical work that has shown fewer generation constraints can improve the generation effect for both item and context memory relative to more highly constrained tasks (Fiedler et al., 1992; Gardiner et al., 1985; McCurdy et al., 2017; McCurdy et al., 2020). Thus, the results of this meta-analysis, along with a growing body of empirical work, advocate for more research considering the influence of differences between tasks, particularly differences in the amount of constraint they impose on participants. One reason why prior work on the generation effect has often avoided truly unconstrained generation tasks is because they invite participants to produce idiosyncratic response that may be somehow differently memorable compared with the words presented in the comparison task (i.e., item selection confounds; Slamecka & Graf, 1978). More recent work on the effects of generation constraint, however, shows evidence that fewer generation constraints increase the generation effect even when controlling for item selection effects (McCurdy et al., 2017; McCurdy, Leach, & Leshikar, 2019; McCurdy et al., 2020). Furthermore, given that we examined generation constraint across several types of generation tasks and studies in this meta-analysis (where it is assumed that item-selection confounds are generally accounted

for), it seems unlikely that the influence of generation constraint is due to item-selection confounds alone.

Beyond simply identifying this phenomenon (that constraint-related differences across tasks influence the generation effect), it is also important to understand potential memory mechanism(s) explaining why constraint might influence the generation effect. The multifactor theory, which we found support for in this meta-analysis, claims that generation often enhances both item-specific and relational processing relative to read controls, resulting in increased memory for generated compared with read materials (Hirshman & Bjork, 1988). It may be possible to understand *generation constraint* through the mechanisms of the multifactor theory. Given that different types of memory tests tap into different types of memory representations (recognition: item-specific; cued recall: cue-target relational; free recall: intertarget relational), we examined the effect of generation constraint for different memory tests. We found the most robust effects of generation constraint occurred for cued recall and free recall memory tests, which prior work suggests are primarily sensitive to cue–target and intertarget relational processing, respectively (Hunt & McDaniel, 1993; Hunt & Seta, 1984). Thus, one possible mechanism is that lower-constraint tasks promote enhanced relational processing relative to higher-constraint generation tasks. This idea is supported by some empirical work showing that lower-constraint generation leads to better memory compared with higher-constraint generation in studies using free recall tests (Fiedler et al., 1992; Gardiner et al., 1985) as well as other work showing that fewer constraints improve memory through enhanced cue–target relational processing (McCurdy et al., 2020). The claim in McCurdy et al. (2020) was based on the finding that reduced constraints improved the generation effect for cued recall, but not recognition, which is the same pattern of results we found in this meta-analysis (lower > medium > higher constraint for cued recall, but no differences for recognition tests). Overall, this meta-analysis shows that generation constraint has the largest effect on memory as measured by recall, potentially implicating enhanced relational processing as a mechanism for how generation constraints influence memory. Continued research aimed at investigating why reduced constraints might influence the type of processing engaged at study may be a particularly useful in advancing our knowledge of the mechanism behind the benefits from lower-constraint generation tasks.

### Experimental design moderators

A final goal of this meta-analysis was to examine the influence of different experimental moderators on the magnitude of the generation effect. We found several significant moderators that may be worth investigating in future research. One particular moderator of note is the length of retention interval between study and test. Our results showed that longer retention times (>1 day) produced a significant increase (approximately 7

percentage points) in the size of generation effect compared with the more standard procedure of shorter (immediate–5 min) and medium (5 min–1 day) length intervals. This finding suggests that generation may act to slow the decay of memory relative to reading. Indeed, the data in Table 2 show that memory for generated materials is relatively stable across the three retention intervals (with a slight increase in performance for longer intervals), while memory for read materials shows a decrease for longer intervals. One possibility is that generation influences consolidation processes, causing memory effects to grow larger over time. Consistent with these findings, Mulligan and Peterson (2015) also found that using a between-subjects design, read items were better remembered than generated items (i.e., a negative generation effect) at an immediate test of free recall. After a 2-day delay, however, generation led to significantly less forgetting than the read condition, and numerically higher recall performance. Together, these results suggest that the benefits of generation on reduced forgetting are genuine and worth further theoretical investigation.

Another notable moderator is the difference between intentional and incidental encoding procedures. We found that incidental encoding procedures led to larger generation effects than did intentional procedures. One reason researchers have proposed that the generation effect is reduced under intentional encoding is due to more effective reading when participants are aware of an upcoming memory test (Begg, Vinski, Frankovich, & Holgate, 1991; Watkins & Sechler, 1988). Specifically, when participants are told of the impending memory test (intentional encoding), they may employ different strategies to help them remember that information during reading, leading to increased memory for read information (and thus a smaller difference in memory performance from generate information). In contrast, generation tasks provide participants with an effective encoding strategy to perform, thus, regardless of intention to remember, generating leads to processing that improves memory (Begg et al., 1991). Our findings support this idea. From Table 2, generation tasks led to nearly identical memory performance for both incidental and intentional procedures, but read tasks showed better memory performance under intentional procedures compared with incidental procedures.

More broadly, finding that incidental procedures lead to larger generation effects suggests that self-generation may be more than a simple memory strategy to help us perform everyday tasks. Instead, these data imply that some of the strongest benefits of self-generation may arise when generation occurs naturally (or without intent to remember), expanding the potential utility and importance of studying this effect. For example, research on inferential processing shows that when people make inferences from texts, these inferences (i.e., self-generated ideas) are often better remembered than explicitly presented information (Ehrlich, 1980; van den Broek, 1994). Better understanding the benefits from naturally occurring generation could have wide-ranging applications, such as

designing more effective educational curriculum and activities aimed at promoting learning and retention. Thus, future research integrating the knowledge about the benefits of explicit and natural uses of self-generation is important to understanding of how the benefits of self-generation can be applied more broadly.

In addition to the retention interval and learning type moderators, we also found several other moderators (e.g., stimuli relation, type of generation task) that significantly influenced the size of the generation effect, and other moderators that did not (e.g., subject population; see Table 2). It is worth noting that the analysis comparing different stimuli relationships showed that rhyme and unrelated words led to smaller generation effects compared with other stimuli relations. Although we did not use this moderator to test the semantic activation theory, this finding provides further support for that account, given that rhyme and unrelated word pairs are not likely to induce semantic processing of the words at encoding. It is also worth noting that we found no differences in the size of the generation effect for younger compared with older adults across the studies included. This finding suggests that generation may be a useful way to improve memory in older adults who often suffer from memory declines with increasing age (Park et al., 2002). Overall, a general conclusion from our analyses of additional moderators is that multiple factors play a role in determining the size of the generation effect, which highlights the need for more nuanced theories that take into account multiple factors simultaneously to more fully capture the variability in generation effects, an idea we discuss further in the following paragraph.

### Limitations and future theoretical directions

As with all meta-analyses, the present study is not without limitations that should be addressed in future empirical work. One issue is that the macrolevel assessment of the theories necessitated by a meta-analytic approach may overlook some of the nuance required to fully capture the complexity of the generation effect. For example, our context memory analyses showed support for the associative strengthening account, indicating that generation generally improves context memory over reading. However, when these contextual details are broken down further, as in our assessment of the processing account, we find that whether or not the generation effect occurs for contextual details critically depends on the type of detail being measured (and likely on additional moderators as well). A similar limitation occurs with many of the other theories we examined in this study as well. For instance, in assessing support for the multifactor theory, our model does not account for the simultaneous influence of other moderators (type of generation task, list structure, etc.). Prior empirical work, however, suggests that the magnitude of the generation effect for recognition, cued recall, or free recall can largely depend on

the categorical structure of the word list (Burns, 1990), the type of instructions given at encoding (deWinstanley & Bjork, 1997), and experimental design (Groszofsky, Payne, & Campbell, 1994). This complexity in generation effects across different experimental procedures could explain the lack of a unified theory on the generation effect. To that end, we encourage future theoretical and empirical work to target the development of models of the generation effect that can account for the simultaneous influence among various moderators that determine the precise magnitude of generation effects under a given situation.

Toward developing a unified theory of generation effect, it is worth noting that the theories we found to be most strongly supported in this meta-analysis were those that simultaneously considered multiple factors to explain the generation effect (multifactor theory, processing account). Looking forward, it may be that future theoretical work should consider various factors in conjunction (beyond what the multifactor and processing account consider) to more effectively account for the generation effect. One such model that attempts to account for multiple factors to make predictions about memory is Jenkins's (1979) tetrahedral model, which states that one must consider four different factors (encoding tasks, memory test, materials, and participant abilities) to fully understand memory phenomena. Jenkins's model further suggests that to truly understand memory outcomes, one must examine how these different factors interact with one another to influence memory. Thus, we suggest that future theoretical work on the generation effect should look at memory as a function of the interactions between various experimental factors, such as those identified by Jenkins (1979). Such work would help develop a richer theoretical understanding of the generation effect.

An important remaining issue in relation to existing and future theories on the generation effect is the measurement of underlying processes in relation to understanding how generation improves memory over reading. A connecting theme in many of the theories supported in this meta-analysis (multifactor theory, processing account) is the match in processing between generation task and memory test, which is the primary principle behind transfer-appropriate processing (Morris et al., 1977). For example, the multifactor theory supported in this meta-analysis suggests that generation tasks improve memory over reading because generation tasks often induce enhanced item-specific and relational processing at encoding (Hirshman & Bjork, 1988). As we have reported, however, there are several factors that influence the size of the generation effect, and thus the type of processing engaged at encoding seems to heavily depend on the experimental context (Burns, 1990). One issue is that the basis of these claims about how different factors influence encoding processes often come from measuring performance on different memory tests (e.g., recognition, cued recall, free recall) that are assumed to be sensitive to different types of processing. What is missing from much of the existing

generation effect literature however, is a measure of encoding processes (e.g., item specific, relational) that is independent of the outcome measure used to assess generation effects (i.e., memory tests). Since the development of the existing generation effect theories tested in this meta-analysis, research has developed independent indices of item-specific and relational processing (Burns, 2006). Including these measures in future work would be a critical step forward in understanding the unique influence of multiple factors (generation tasks, stimuli type, etc.) on the underlying processes engaged at encoding, and should help spur the development of more nuanced theories of the generation effect that can account for multiple factors, as we suggested earlier.

### Beyond the generation effect

Although this meta-analysis focused on the generation effect, our findings could be relevant to other aspects of psychology as well, such as in education (e.g., the testing effect; Dunlosky et al., 2013; Roediger & Karpicke, 2006), or in self-persuasion (e.g., self-generated persuasive arguments; Aronson, 1999; Briñol, McCaslin, & Petty, 2012). For instance, in this meta-analysis, we found evidence that fewer generation constraints increase the magnitude of the generation effect. This finding may be informative in other domains. Specifically, it may be that the memory benefit observed in the testing effect may also be magnified through experimental manipulations that reduce the constraints used during practice tests in the typical testing effect procedure. Such a possibility might be effective in improving memory for instructional material in the classroom. Similarly, self-generated persuasive arguments might be more influential in changing one's attitudes or beliefs when there are fewer constraints on what can be self-generated (Briñol et al., 2012; Janis & King, 1954). Altogether, our findings may also be used to inform research in other domains and contribute to a better understanding of other psychological phenomena beyond the generation effect. Understanding ways to improve memory is an important pursuit (Leach, McCurdy, Trumbo, Matzen, & Leshikar, 2018; Leshikar, Dulas, & Duarte, 2015; Leshikar & Gutchess, 2015; Leshikar et al., 2017; Matzen, Trumbo, Leach, & Leshikar, 2015), and this work contributes to that effort.

### Conclusion

The memory benefits from self-generation have long been recognized. With this meta-analysis, we show the impressive benefits of generating information over reading; however, we also identify areas for further research on the generation effect. In particular, we show that reduced generation constraints

increase memory performance. This finding suggests that the nature of the generation task has a large influence on memory. Investigating factors that influence memory and developing theories that most effectively describe memory effects, such as the ones we have investigated in this meta-analysis, is fundamental to a deeper understanding of memory. Thus, we encourage more work to develop a richer theory to explain the generation effect, perhaps through the consideration of multiple experimental factors. Overall, the generation effect is a versatile memory phenomenon, improving memory in various situations as accounted for by the multifactor theory and processing account.

**Open practices statement** The data and analysis code for this article is available at: <https://osf.io/9pv7a/>. This article was not preregistered.

## Appendix

**Table 3.** Illustration of the coding of the data set structure

Article	Experiment	Sample	Observation	Pairing	Condition
1	1	1	1	1	0
1	1	2	2	1	1
1	2	1	3	2	0
1	2	2	4	2	1
2	1	1	5	1	0
2	1	1	6	1	1
3	1	1	7	1	0
3	1	1	8	1	1
3	1	1	9	1	1
4	1	1	10	1	0
4	1	2	11	1	1
4	1	3	12	2	0
4	1	4	13	2	1
5	1	1	14	1	0
5	1	1	15	1	1
5	1	1	16	2	0
5	1	1	17	2	1
6	1	1	18	1	0
6	1	1	19	1	1
6	1	2	20	2	0
6	1	2	21	2	1
6	1	3	22	3	0
6	1	3	23	3	1

In the table above, we illustrate with some examples the manner in which the structure of the data set was coded. Variable “article” is simply an identifier for each article, whereas “experiment” identifies the unique experiments within each article (in case the same article contains the results from multiple experiments). In this illustration, Article 1 reported the results of two different experiments, while all other articles are single-experiment studies. The “sample” identifier (coded within experiments) indicates which rows correspond to the same sample. For example, the two experiments within Study 1 are both (one-way) between-subjects designs, whereas Study 2 was a (one-way) within-subjects design. Next, the “observation” identifier simply distinguishes the different rows in the data set. A crucial element in such an “arm-based” data structure (where each row corresponds to the results obtained within an individual group and not the contrast between two different conditions) is the “pairing” variable. This identifier indicates which rows within an experiment comprise a particular comparison between generate versus read conditions, where the type of condition within a particular row is reflected by the “condition” dummy (0 = read, 1 = generate). Although in many cases a pairing will only consist of two rows of the data set, there can be situations where more than one generate condition is compared with a read condition. This latter case is illustrated in Article 3, where the same sample completed a read and two different generate conditions (e.g., one with low and one with high generation constraint). To complete this illustration, Articles 4, 5, and 6 exemplify more complex designs, with Articles 4 and 5 representing  $2 \times 2$  factorial designs (with one factor for the read versus generate condition and the other factor representing some other design variable, such as the use of words versus nonwords), where both factors are between-subjects factors in Article 4 and both factors are within-subjects factors in Article 5. These two articles also illustrate how different samples may be involved in the same pairing (as in Article 4) or the same sample may be involved in different pairings (as in Article 5). Finally, Article 6 illustrates a  $2 \times 3$  factorial design, with the two-level within-subjects factor corresponding to the read versus generate condition, and the three-level between-subjects factor to some other design variable (e.g., the type of memory test used, with levels recognition, cued recall, or free recall).

## References

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

- Aronson, E. (1999). The power of self-persuasion. *American Psychologist*, *54*(11), 875–884.
- Begg, I., & Snider, A. (1987). The generation effect: Evidence for generalized inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(4), 553–563.

- Begg, I., Snider, A., Foley, F., & Goddard, R. (1989). The generation effect is no artifact: Generating makes words distinctive. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(5), 977–989.
- \*Begg, I., Vinski, E., Frankovich, L., & Holgate, B. (1991). Generating makes words memorable, but so does effective reading. *Memory & Cognition*, *19*(5), 487–497.
- Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, *35*(2), 201–210.
- Briñol, P., McCaslin, M. J., & Petty, R. E. (2012). Self-generated persuasion: Effects of the target and direction of arguments. *Journal of Personality and Social Psychology*, *102*(5), 925–940.
- \*Burns, D. J. (1990). The generation effect: A test between single-and multifactor theories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(6), 1060–1067.
- \*Burns, D. J. (1992). The consequences of generation. *Journal of Memory and Language*, *31*(5), 615–633.
- \*Burns, D. J. (1996). The item-order distinction and the generation effect: The importance of order information in long-term memory. *The American Journal of Psychology*, *109*(4), 567–580.
- Burns, D. J. (2006). Assessing distinctiveness: Measures of item-specific and relational processing. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 109–130). Oxford, England: Oxford University Press.
- \*Burns, D. J., Curti, E. T., & Lavin, J. C. (1993). The effects of generation on item and order retention in immediate and delayed recall. *Memory & Cognition*, *21*(6), 846–852.
- \*Buyer, L. S., & Dominowski, R. L. (1989). Retention of solutions: It is better to give than to receive. *The American Journal of Psychology*, *102*(3), 353–363.
- \*Carroll, M., Davis, R., & Conway, M. (2001). The effects of self-reference on recognition and source attribution. *Australian Journal of Psychology*, *53*(3), 140–145.
- \*Carroll, M., & Nelson, T. O. (1993). Failure to obtain a generation effect during naturalistic learning. *Memory & Cognition*, *21*(3), 361–366.
- \*Chechile, R. A., & Soraci, S. A. (1999). Evidence for a multiple-process account of the generation effect. *Memory*, *7*(4), 483–508.
- \*Clark, S. E. (1995). The generation effect and the modeling of associations in memory. *Memory & Cognition*, *23*(4), 442–455.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220.
- \*Crutcher, R. J., & Healy, A. F. (1989). Cognitive operations and the generation effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 669–675.
- \*Dew, I. T., & Mulligan, N. W. (2008). The effects of generation on auditory implicit memory. *Memory & Cognition*, *36*(6), 1157–1167.
- \*deWinstanley, P. A., & Bjork, E. L. (1997). Processing instructions and the generation effect: A test of the multifactor transfer-appropriate processing theory. *Memory*, *5*(3), 401–422.
- \*deWinstanley, P. A., Bjork, E. L., & Bjork, R. A. (1996). Generation effects and the lack thereof: The role of transfer-appropriate processing. *Memory*, *4*(1), 31–48.
- \*Donaldson, W., & Bass, M. (1980). Relational information and memory for problem solutions. *Journal of Verbal Learning and Verbal Behavior*, *19*(1), 26–35.
- Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4–58.
- Edger, 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.
- Ehrlich, M. F. (1980). Text comprehension and memory for inferences. *Advances in Psychology* (Vol. 5, pp. 186–195). Amsterdam, Netherlands: Elsevier.
- Fernández-Castilla, B., Maes, M., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2019). A demonstration and evaluation of the use of cross-classified random-effects models for meta-analysis. *Behavior Research Methods*, *51*(3), 1286–1304.
- \*Fiedler, K., Lachnit, H., Fay, D., & Krug, C. (1992). Mobilization of cognitive resources and the generation effect. *The Quarterly Journal of Experimental Psychology*, *45*(1), 149–171.
- \*Flory, P., & Pring, L. (1995). The effects of data-driven and conceptually driven generation of study items on direct and indirect measures of memory. *The Quarterly Journal of Experimental Psychology*, *48*(1), 153–165.
- \*Foley, M. A., & Foley, H. J. (2007). Source-monitoring judgments about anagrams and their solutions: Evidence for the role of cognitive operations information in memory. *Memory & Cognition*, *35*(2), 211–221.
- \*Foley, M. A., Foley, H. J., Durley, J. R., & Maitner, A. T. (2006). Anticipating partners' responses: Examining item and source memory following interactive exchanges. *Memory & Cognition*, *34*(7), 1539–1547.
- Foley, M. A., Foley, H. J., Wilder, A., & Rusch, L. (1989). Anagram solving: Does effort have an effect? *Memory & Cognition*, *17*(6), 755–758.
- \*Froger, C., Sacher, M., Gaudouen, M.-S., Isingrini, M., & Tacconat, L. (2011). Metamemory judgments and study time allocation in young and older adults: Dissociative effects of a generation task. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *65*(4), 269–276.
- \*Gardiner, J. M. (1988). Generation and priming effects in word-fragment completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 495–501.
- \*Gardiner, J. M. (1989). A generation effect in memory without awareness. *British Journal of Psychology*, *80*(2), 163–168.
- \*Gardiner, J. M., & Arthurs, F. S. (1982). Encoding context and the generating effect in multitrial free-recall learning. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *36*(3), 527–531.
- \*Gardiner, J. M., Gregg, V. H., & Hampton, J. A. (1988). Word frequency and generation effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 687–693.
- Gardiner, J. M., & Hampton, J. A. (1985). Semantic memory and the generation effect: Some tests of the lexical activation hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 732–741.
- \*Gardiner, J. M., & Hampton, J. A. (1988). Item-specific processing and the generation effect: Support for a distinctiveness account. *The American Journal of Psychology*, *101*(4), 495–504.
- \*Gardiner, J. M., Smith, H. E., Richardson, C. J., Burrows, M. V., & Williams, S. D. (1985). The generation effect: Continuity between generating and reading. *The American Journal of Psychology*, *98*(3), 373–378.
- \*Geghman, K. D., & Multhaup, K. S. (2004). How generation affects source memory. *Memory & Cognition*, *32*(5), 819–823.
- \*Ghatala, E. S. (1983). When does internal generation facilitate memory for sentences? *The American Journal of Psychology*, *96*(1), 75–83.
- \*Glisky, E. L., & Rabinowitz, J. C. (1985). Enhancing the generation effect through repetition of operations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(2), 193–205.
- \*Graf, P. (1981). Reading and generating normal and transformed sentences. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *35*(4), 293–308.
- \*Greenwald, A. G., & Johnson, M. M. (1989). The generation effect extended: Memory enhancement for generation cues. *Memory & Cognition*, *17*(6), 673–681.

- \*Gregory, M. E., Mergler, N. L., Durso, F. T., & Zandi, T. (1988). Cognitive reality monitoring in adulthood. *Educational Gerontology: An International Quarterly*, 14(1), 1–13.
- \*Grososky, A., Payne, D. G., & Campbell, K. D. (1994). Does the generation effect depend upon selective displaced rehearsal? *The American Journal of Psychology*, 107(1), 53–68.
- \*Guynn, M. J., McDaniel, M. A., Strosser, G. L., Ramirez, J. M., Castleberry, E. H., & Arnett, K. H. (2014). Relational and item-specific influences on generate–recognize processes in recall. *Memory & Cognition*, 42(2), 198–211.
- \*Hara, K., Neumann, E., & Tajika, H. (1989). Effects of word versus nonword rehearsal frequency on the generation effect. *Psychologia*, 32(4), 230–235.
- \*Harbluk, J. L. (1994). *The influence of intervening activities and testing conditions on the accuracy and confidence of source memory* (Doctoral dissertation, The University of Western Ontario, Canada). Retrieved from <https://ir.lib.uwo.ca/digitizedtheses/2379/>
- \*Hertel, P. T. (1989). The generation effect: A reflection of cognitive effort? *Bulletin of the Psychonomic Society*, 27(6), 541–544.
- \*Heth, I. (2001). Two generation tasks: Age-related differences in item and source memory. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, 61(9-B), 5017.
- \*Hicks, J. L., & Marsh, R. L. (1999). Attempts to reduce the incidence of false recall with source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25(5), 1195–1209.
- \*Hirshman, E., & Bjork, R. A. (1988). The generation effect: Support for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(3), 484–494.
- \*Hoffman, H. G. (1997). Role of memory strength in reality monitoring decisions: Evidence from source attribution biases. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(2), 371–383.
- 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.
- Hunt, R. R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory and Language*, 32(4), 421–445.
- Hunt, R. R., & Seta, C. E. (1984). Category size effects in recall: The roles of relational and individual item information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(3), 454–464.
- \*Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, 17(6), 649–667.
- \*Jacoby, L. L. (1983). Remembering the data: Analyzing interactive processes in reading. *Journal of Verbal Learning and Verbal Behavior*, 22(5), 485–508.
- Janis, I. L., & King, B. T. (1954). The influence of role playing on opinion change. *The Journal of Abnormal and Social Psychology*, 49(2), 211–218.
- \*Java, R. I. (1994). States of awareness following word stem completion. *European Journal of Cognitive Psychology*, 6(1), 77–92.
- \*Java, R. I. (1996). Effects of age on state of awareness following implicit and explicit word-association tasks. *Psychology and Aging*, 11(1), 108–111.
- Jenkins, J. J. (1979). Four points to remember: A tetrahedral model of memory experiments. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 429–446). Hillsdale, NJ: Erlbaum.
- \*Johns, E. E., & Swanson, L. G. (1988). The generation effect with nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(1), 180–190.
- Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review*, 88(1), 67–85.
- \*Johnson, M. K., Raye, C. L., Foley, H. J., & Foley, M. A. (1981). Cognitive operations and decision bias in reality monitoring. *The American Journal of Psychology*, 94(1), 37–64.
- \*Johnson, M. M., Schmitt, F. A., & Pietrukowicz, M. (1989). The memory advantages of the generation effect: Age and process differences. *Journal of Gerontology*, 44(3), P91–P94.
- \*Jurica, P. J., & Shimamura, A. P. (1999). Monitoring item and source information: Evidence for a negative generation effect in source memory. *Memory & Cognition*, 27(4), 648–656.
- Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, N.J.: Prentice-Hall.
- \*Kane, J. H., & Anderson, R. C. (1978). Depth of processing and interference effects in the learning and remembering of sentences. *Journal of Educational Psychology*, 70(4), 626–635.
- \*Kelley, M. R., & Nairne, J. S. (2001). von Restorff revisited: Isolation, generation, and memory for order. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1), 54–66.
- \*Kinoshita, S. (1989). Generation enhances semantic processing? The role of distinctiveness in the generation effect. *Memory & Cognition*, 17(5), 563–571.
- Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods*, 2(1), 61–76.
- Leach, R. C., McCurdy, M. P., Trumbo, M. C., Matzen, L. E., & Leshikar, E. D. (2018). Differential age effects of transcranial direct current stimulation on associative memory. *The Journals of Gerontology: Series B*, 74(7), 1163–1173.
- Leshikar, E. D., Dulas, M. R., & Duarte, A. (2015). Self-referencing enhances recollection in both young and older adults. *Aging, Neuropsychology, and Cognition*, 22(4), 388–412.
- Leshikar, E. D., & Gutches, A. H. (2015). Similarity to the self affects memory for impressions of others. *Journal of Applied Research in Memory and Cognition*, 4(1), 20–28.
- Leshikar, E. D., Leach, R. C., McCurdy, M. P., Trumbo, M. C., Sklenar, A. M., Frankenstein, A. N., & Matzen, L. E. (2017). Transcranial direct current stimulation of dorsolateral prefrontal cortex during encoding improves recall but not recognition memory. *Neuropsychologia*, 106, 390–397.
- \*Leynes, P. A. (2012). Event-related potential (ERP) evidence for source-monitoring based on the absence of information. *International Journal of Psychophysiology*, 84(3), 284–295.
- \*Leynes, P. A., Cairns, A., & Crawford, J. T. (2005). Event-related potentials indicate that reality monitoring differs from external source monitoring. *The American Journal of Psychology*, 118(4), 497–524.
- \*Luo, L., Hendriks, T., & Craik, F. I. (2007). Age differences in recollection: Three patterns of enhanced encoding. *Psychology and Aging*, 22(2), 269–280.
- \*Lutz, J., Briggs, A., & Cain, K. (2003). An examination of the value of the generation effect for learning new material. *The Journal of General Psychology*, 130(2), 171–188.
- \*MacLeod, C. M., Pottruff, M. M., Forrin, N. D., & Masson, M. E. (2012). The next generation: The value of reminding. *Memory & Cognition*, 40(5), 693–702.
- \*Marsh, E. J. (2006). When does generation enhance memory for location? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(5), 1216–1220.
- \*Marsh, E. J., Edelman, G., & Bower, G. H. (2001). Demonstrations of a generation effect in context memory. *Memory & Cognition*, 29(6), 798–805.
- Matzen, L. E., Trumbo, M. C., Leach, R. C., & Leshikar, E. D. (2015). Effects of non-invasive brain stimulation on associative memory. *Brain Research*, 1624, 286–296.
- \*McClelland, A. G. R., & Pring, L. (1991). An investigation of cross-modality effects in implicit and explicit memory. *The Quarterly Journal of Experimental Psychology Section A*, 43(1), 19–33.



- \*McCurdy, M. P., Leach, R. C., & Leshikar, E. D. (2017). The generation effect revisited: Fewer generation constraints enhances item and context memory. *Journal of Memory and Language*, *92*, 202–216.
- \*McCurdy, M. P., Leach, R. C., & Leshikar, E. D. (2019). Fewer generation constraints enhances the generation effect for younger, but not older adults. *Open Psychology*, *1*(1), 168–184
- \*McCurdy, M. P., Sklenar, A. M., Frankenstein, A. N., & Leshikar, E. D. (2020). Fewer generation constraints increase the generation effect through enhanced relational memory representations. *Memory*, 1–19. Advance online publication. <https://doi.org/10.1080/09658211.2020.1749283>
- \*McDaniel, M. A., Riegler, G. L., & Waddill, P. J. (1990). Generation effects in free recall: Further support for a three-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(5), 789–798.
- \*McDaniel, M. A., & Waddill, P. J. (1990). Generation effects for context words: Implications for item-specific and multifactor theories. *Journal of Memory and Language*, *29*(2), 201–211.
- \*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.
- \*McElroy, L. A. (1987). The generation effect with homographs: Evidence for postgeneration processing. *Memory & Cognition*, *15*(2), 148–153.
- \*McElroy, L. A., & Slamecka, N. J. (1982). Memorial consequences of generating nonwords: Implications for semantic-memory interpretations of the generation effect. *Journal of Verbal Learning and Verbal Behavior*, *21*(3), 249–259.
- \*McFarland, C. E., Frey, T. J., & Rhodes, D. D. (1980). Retrieval of internally versus externally generated words in episodic memory. *Journal of Verbal Learning and Verbal Behavior*, *19*(2), 210–225.
- \*McFarland, C. E., Warren, L. R., & Crockard, J. (1985). Memory for self-generated stimuli in young and old adults. *Journal of Gerontology*, *40*(2), 205–207.
- \*McNamara, D. S., & Healy, A. F. (1995). A procedural explanation of the generation effect: The use of an operand retrieval strategy for multiplication and addition problems. *Journal of Memory and Language*, *34*(3), 399–416.
- \*McNamara, D. S., & Healy, A. F. (2000). A procedural explanation of the generation effect for simple and difficult multiplication problems and answers. *Journal of Memory and Language*, *43*(4), 652–679.
- Metcalf, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin & Review*, *14*(2), 225–229.
- Mitchell, D. B., & Hunt, R. R. (1989). How much “effort” should be devoted to memory? *Memory & Cognition*, *17*(3), 337–348.
- Moeyaert, M., Ugille, M., Natasha Beretvas, S., Ferron, J., Bunuan, R., & Van den Noortgate, W. (2017). Methods for dealing with multiple outcomes in meta-analysis: A comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology*, *20*(6), 559–572.
- Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine*, *151*(4), 264–269.
- 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.
- \*Mulligan, N. W. (2001). Generation and hypermnnesia. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(2), 436–450.
- \*Mulligan, N. W. (2002a). The effects of generation on conceptual implicit memory. *Journal of Memory and Language*, *47*(2), 327–342.
- \*Mulligan, N. W. (2002b). The emergent generation effect and hypermnnesia: Influences of semantic and nonsemantic generation tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 541–554.
- \*Mulligan, N. W. (2002c). The generation effect: Dissociating enhanced item memory and disrupted order memory. *Memory & Cognition*, *30*(6), 850–861.
- \*Mulligan, N. W. (2004). Generation and memory for contextual detail. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 838–855.
- \*Mulligan, N. W. (2011). Generation disrupts memory for intrinsic context but not extrinsic context. *The Quarterly Journal of Experimental Psychology*, *64*(8), 1543–1562.
- \*Mulligan, N. W., & Duke, M. D. (2002). Positive and negative generation effects, hypermnnesia, and total recall time. *Memory & Cognition*, *30*(7), 1044–1053.
- Mulligan, N. W., & Lozito, J. P. (2004). Self-generation and memory: The psychology of learning and motivation. *Advances in Research and Theory*, *45*, 175–214.
- \*Mulligan, N. W., & Lozito, J. P. (2006). An asymmetry between memory encoding and retrieval: Revelation, generation, and transfer-appropriate processing. *Psychological Science*, *17*(1), 7–11.
- \*Mulligan, N. W., Lozito, J. P., & Rosner, Z. A. (2006). Generation and context memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 836–846.
- \*Mulligan, N. W., & Peterson, D. (2008). Assessing a retrieval account of the generation and perceptual-interference effects. *Memory & Cognition*, *36*(8), 1371–1382.
- \*Mulligan, N. W., & Peterson, D. J. (2015). The negative testing and negative generation effects are eliminated by delay. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(4), 1014–1025.
- \*Muntean, W. J., & Kimball, D. R. (2012). Part-set cueing and the generation effect: An evaluation of a two-mechanism account of part-set cueing. *Journal of Cognitive Psychology*, *24*(8), 957–964.
- \*Naime, J. S., Pusen, C., & Widner, R. L. (1985). Representation in the mental lexicon: Implications for theories of the generation effect. *Memory & Cognition*, *13*(2), 183–191.
- \*Naime, J. S., Riegler, G. L., & Serra, M. (1991). Dissociative effects of generation on item and order retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(4), 702–709.
- \*Naime, J. S., & Widner, R. L. (1988). Familiarity and lexicality as determinants of the generation effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 694–699.
- \*Nicolas, S., Ehrlich, M.-F., & Facci, G. (1996). Implicit memory and aging: Generation effect in a word-stem completion test. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, *15*(5), 513–533.
- \*Nicolas, S., & Tardieu, H. (1996). The generation effect in a word-stem completion task: The Influence of conceptual processes. *The European Journal of Cognitive Psychology*, *8*(4), 405–424.
- \*Nieznański, M. (2011). Generation difficulty and memory for source. *The Quarterly Journal of Experimental Psychology*, *64*(8), 1593–1608.
- \*Nieznański, M. (2012). Effects of generation on source memory: A test of the resource tradeoff versus processing hypothesis. *Journal of Cognitive Psychology*, *24*(7), 765–780.
- Nieznański, M. (2013). Effects of resource demanding processing on context memory for context-related versus context-unrelated items. *Journal of Cognitive Psychology*, *25*(6), 745–758.
- \*Nieznański, M. (2014a). Context reinstatement and memory for intrinsic versus extrinsic context: The role of item generation at encoding or retrieval. *Scandinavian Journal of Psychology*, *55*(5), 409–419.
- \*Nieznański, M. (2014b). The role of reinstating generation operations in recognition memory and reality monitoring. *Polish Psychological Bulletin*, *45*(3), 363–371.

- \*Olofsson, U., & Nilsson, L.-G. (1992). The generation effect in primed word-fragment completion reexamined. *Psychological Research*, 54(2), 103–109.
- \*Overman, A. A., Richard, A. G., & Stephens, J. D. (2017). A positive generation effect on memory for auditory context. *Psychonomic Bulletin & Review*, 24(3), 944–949.
- Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N. S., Smith, A. D., & Smith, P. K. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology and Aging*, 17(2), 299–320.
- \*Payne, D. G., Neely, J. H., & Burns, D. J. (1986). The generation effect: Further tests of the lexical activation hypothesis. *Memory & Cognition*, 14(3), 246–252.
- \*Pesta, B. J., Sanders, R. E., & Murphy, M. D. (1999). A beautiful day in the neighborhood: What factors determine the generation effect for simple multiplication problems? *Memory & Cognition*, 27(1), 106–115.
- \*Pesta, B. J., Sanders, R. E., & Nemeč, R. J. (1996). Older adults' strategic superiority with mental multiplication: A generation effect assessment. *Experimental Aging Research*, 22(2), 155–169.
- Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. *Journal of the American Medical Association*, 295(6), 676–680.
- \*Peynircioğlu, Z. F., & Mungan, E. (1993). Familiarity, relative distinctiveness, and the generation effect. *Memory & Cognition*, 21(3), 367–374.
- Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, 143(2), 644–667.
- R Core Team. (2018). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- \*Rabinowitz, J. C. (1989). Judgments of origin and generation effects: Comparisons between young and elderly adults. *Psychology and Aging*, 4(3), 259–268.
- \*Rabinowitz, J. C., & Craik, F. I. (1986). Specific enhancement effects associated with word generation. *Journal of Memory and Language*, 25(2), 226–237.
- Raudenbush, S. W., & Bryk, A. S. (1987). Examining correlates of diversity. *Journal of Educational Statistics*, 12(3), 241–269.
- \*Reardon, R., Durso, F. T., Foley, M. A., & McGahan, J. R. (1987). Expertise and the generation effect. *Social Cognition*, 5(4), 336–348.
- Richland, L. E., Bjork, R. A., Finley, J. R., & Linn, M. C. (2005). *Linking cognitive science to education: Generation and interleaving effects*. Paper presented at the Proceedings of the twenty-seventh annual conference of the Cognitive Science Society.
- \*Riefer, D. M., Chien, Y., & Reimer, J. F. (2007). Positive and negative generation effects in source monitoring. *The Quarterly Journal of Experimental Psychology*, 60(10), 1389–1405.
- Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255.
- Rohatgi, A. (2015). WebPlotDigitizer (Version 3.9) [Computer program]. Retrieved from <http://arohatgi.info/WebPlotDigitizer>
- \*Rosburg, T., Johansson, M., & Mecklinger, A. (2013). Strategic retrieval and retrieval orientation in reality monitoring studied by event-related potentials (ERPs). *Neuropsychologia*, 51(3), 557–571.
- \*Rosner, Z. A. (2012). *The generation effect and memory* (Doctoral dissertation, University of California, Berkeley). Retrieved from [http://digitalassets.lib.berkeley.edu/etd/ucb/text/Rosner\\_berkeley\\_0028E\\_12868.pdf](http://digitalassets.lib.berkeley.edu/etd/ucb/text/Rosner_berkeley_0028E_12868.pdf)
- \*Rosner, Z. A., Elman, J. A., & Shimamura, A. P. (2013). The generation effect: Activating broad neural circuits during memory encoding. *Cortex*, 49(7), 1901–1909.
- Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2006). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. New York, NY: John Wiley & Sons.
- Salanti, G., Higgins, J. P., Ades, A., & Ioannidis, J. P. (2008). Evaluation of networks of randomized trials. *Statistical Methods in Medical Research*, 17(3), 279–301.
- \*Schmidt, S. R. (1990). A test of resource-allocation explanations of the generation effect. *Bulletin of the Psychonomic Society*, 28(1), 93–96.
- \*Schmidt, S. R. (1992). Evaluating the role of distinctiveness in the generation effect. *The Quarterly Journal of Experimental Psychology*, 44(2), 237–260.
- \*Schmidt, S. R., & Cherry, K. (1989). The negative generation effect: Delineation of a phenomenon. *Memory & Cognition*, 17(3), 359–369.
- \*Schwartz, B. L., & Metcalfe, J. (1992). Cue familiarity but not target retrievability enhances feeling-of-knowing judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(5), 1074–1083.
- \*Schweickert, R., McDaniel, M. A., & Riegler, G. (1994). Effects of generation on immediate memory span and delayed unexpected free recall. *The Quarterly Journal of Experimental Psychology Section A*, 47(3), 781–804.
- \*Serra, M., & Naime, J. S. (1993). Design controversies and the generation effect: Support for an item-order hypothesis. *Memory & Cognition*, 21(1), 34–40.
- \*Slamecka, N. J., & Fevreiski, J. (1983). The generation effect when generation fails. *Journal of Verbal Learning and Verbal Behavior*, 22(2), 153–163.
- \*Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, 4(6), 592–604.
- \*Slamecka, N. J., & Katsaiti, L. T. (1987). The generation effect as an artifact of selective displaced rehearsal. *Journal of Memory and Language*, 26(6), 589–607.
- \*Smith, E. R., & Branscombe, N. R. (1988). Category accessibility as implicit memory. *Journal of Experimental Social Psychology*, 24(6), 490–504.
- \*Smith, R. W., & Healy, A. F. (1998). The time-course of the generation effect. *Memory & Cognition*, 26(1), 135–142.
- \*Soraci, S. A., Carlin, M. T., Chechile, R. A., Franks, J. J., Wills, T., & Watanabe, T. (1999). Encoding variability and cuing in generative processing. *Journal of Memory and Language*, 41(4), 541–559.
- \*Soraci, S. A., Franks, J. J., Bransford, J. D., Chechile, R. A., Belli, R. F., Carr, M., & Carlin, M. (1994). Incongruous item generation effects: A multiple-cue perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(1), 67–78.
- Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, 5(1), 60–78.
- Steffens, M. C., & Erdfelder, E. (1998). Determinants of positive and negative generation effects in free recall. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 51(4), 705–733.
- Stram, D. O. (1996). Meta-analysis of published data using a linear mixed-effects model. *Biometrics*, 52(2), 536–544.
- \*Tacconat, L., Baudouin, A., Fay, S., Clarys, D., Vanneste, S., Toumelle, L., & Isingrini, M. (2006). Aging and implementation of encoding strategies in the generation of rhymes: The role of executive functions. *Neuropsychology*, 20(6), 658–665.
- \*Tacconat, L., Froger, C., Sacher, M., & Isingrini, M. (2008). Generation and associative encoding in young and old adults: The effect of the

- strength of association between cues and targets on a cued recall task. *Experimental Psychology*, 55(1), 23–30.
- \*Taconnat, L., & Isingrini, M. (2004). Cognitive operations in the generation effect on a recall test: Role of aging and divided attention. *Journal of Experimental Psychology Learning Memory and Cognition*, 30, 827–837.
- Tyler, S. W., Hertel, P. T., McCallum, M. C., & Ellis, H. C. (1979). Cognitive effort and memory. *Journal of Experimental Psychology: Human Learning and Memory*, 5(6), 607–617.
- van den Broek, P. (1994). Comprehension and memory of narrative texts: Inferences and coherence. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 539–588). San Diego, CA: Academic Press.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48.
- Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, 37(5), 360–363.
- \*Watkins, M. J., & Sechler, E. S. (1988). Generation effect with an incidental memorization procedure. *Journal of Memory and Language*, 27(5), 537–544.
- \*Weldon, M. S., & Colston, H. L. (1995). Dissociating the generation stage in implicit and explicit memory tests: Incidental production can differ from strategic access. *Psychonomic Bulletin & Review*, 2(3), 381–386.
- \*Whiting, W. L. (1998). *Effects of elaboration on age differences in memory performance*. (Doctoral Dissertation, Georgia Institute of Technology).
- \*Widner, R. L. (1995). Associative spread as a mediating variable in the generation effect. *Memory*, 3(1), 1–19.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.