Fewer Generation Constraints Increase the Generation Effect for Item and Source Memory through Enhanced Relational Processing

Matthew P. McCurdy, Allison M. Sklenar, Andrea N. Frankenstein, and Eric D. Leshikar

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

Author Note

This is an Accepted Manuscript of an article published by Taylor & Francis in *Memory* on April 20, 2020, available online: http://www.tandfonline.com/10.1080/09658211.2020.1749283

The data supporting the findings of this study are openly available at https://osf.io/cqhfy/

Corresponding Author:

Eric D. Leshikar
1007 West Harrison Street (M/C 285)
Chicago, IL 60607
(312) 355-2739
leshikar@uic.edu

Abstract

Memory is often better for information that is self-generated versus read (i.e., the generation effect). Theoretical work attributes the generation effect to two mechanisms: enhanced item-specific and relational processing (i.e., the two-factor theory). Recent work has demonstrated that the generation effect increases when generation tasks place lower, relative to higher, constraints on what participants can self-generate. This study examined whether the effects of generation constraint on memory might be attributable to either mechanism of the two-factor theory. Across three experiments, participants encoded word pairs in two generation conditions (lower- and higher-constraint) and a read control task, followed by a memory test for item memory and two context memory details (source and font color). The results of these experiments support the idea that lower-constraint generation increases the generation effect via enhanced relational processing, as measured through both recognition and cued recall tasks. Results further showed that lower-constraint generation improves context memory for conceptual context (source), but not perceptual context (color), suggesting that this enhanced relational processing may extend to conceptually related details of an item. Overall, these results provide more evidence that fewer generation constraints increase the generation effect and implicate enhanced relational processing as a mechanism for this improvement.

Keywords: Generation effect, Generation constraint, Item memory, Context memory

Word count: 12,046

Fewer Generation Constraints Increase the Generation Effect for Item and Source Memory Through Enhanced Relational Processing.

The commonly observed memory benefit for self-generated materials over those that are simply read is known as the *generation effect* (Slamecka & Graf, 1978). A wealth of evidence suggests the generation effect is robust across a variety of procedures and memory tests (Bertsch, Pesta, Wiscott, & McDaniel, 2007; McCurdy, Viechtbauer, Frankenstein, Sklenar, & Leshikar, Under Review). Despite its robustness, some research has shown that certain experimental factors can influence the size of this effect (e.g., between-subject vs. within-subject designs; Begg & Snider, 1987; Grosofsky, Payne, & Campbell, 1994; Hirshman & Bjork, 1988; Kinoshita, 1989; Slamecka & Katsaiti, 1987). Understanding how different experimental factors influence the generation effect is important because it can provide insights about the mechanisms underlying memory improvements. In the present study, we examine the influence of *generation constraint* (i.e., information given to participants limiting what they can self-generate) on the magnitude of the generation effect.

In prior work, many commonly used generation tasks (e.g., solving anagrams from cue words, open - csole; Foley & Foley, 2007; Foley, Foley, Wilder, & Rusch, 1989) highly constrain participants to generate a single, correct response. Limited research, however, suggests that tasks with fewer constraints lead to larger generation effects compared to tasks with higher constraints (Fiedler, Lachnit, Fay, & Krug, 1992; Gardiner, Smith, Richardson, Burrows, & Williams, 1985; McCurdy, Leach, & Leshikar, 2017, 2019), suggesting this factor is important to consider in advancing our understanding of the generation effect. Most recently, research directly examining the influence of generation constraint compared memory for materials produced in a *lower-constraint* task (freely generating from a cue word: open – _____) relative to a *higher-*

constraint task (solving an anagram: open – csoel; filling in missing letters: open – cl*s*), and found that in many instances lower-constraint generation led to increased memory (McCurdy et al., 2017, 2019). Given that generation constraint has not been well scrutinized, we aim to expand on this prior work by investigating potential mechanisms accounting for the memory improvements often found for lower- compared to higher-constraint generation tasks.

A theoretical framework, the two-factor theory (also known as the multi-factor theory; Hirshman & Bjork, 1988; McDaniel, Waddill, & Einstein, 1988), posits that self-generation improves *item memory* (i.e., memory for the generated target) relative to read controls through two memory mechanisms: 1) increased item-specific processing (i.e., processing of the features that are unique to the target item) and 2) increased relational processing of information presented with the item (e.g., strengthening of the relationship between the cue and generated target word; Donaldson & Bass, 1980; Hirshman & Bjork, 1988).[1] One way to assess the contribution of these two factors on memory is through the use of different memory tests. Research suggests that item recognition memory tests are especially sensitive to differences in item-specific processing (Burns, 2006; Einstein & Hunt, 1980), whereas cued recall tests are more sensitive to relational information, especially the cue-target relationship (Donaldson & Bass, 1980; Hirshman & Bjork, 1988).

Different types of memory tests have been used in prior work to investigate the role of item-specific and relational processing in generation effects. For instance, some research suggests that for item recognition tests (i.e., distinguishing between previously seen items and distractor items), self-generating makes items more distinct relative to reading, leading to greater

[1] It should be noted that other research, most notably McDaniel et al. (1988), has shown that in some situations generation can also enhance "whole-list" relational processing of the shared features across all items encoded (i.e., the multi-factor theory), however, the present study does not aim to interrogate this type of processing, thus we have restricted this review of the literature to research pertaining to "cue-target" relational processing.

recognition of generated items among distractor words (Begg, Snider, Foley, & Goddard, 1989; Gardiner & Hampton, 1988; Glisky & Rabinowitz, 1985; Slamecka & Graf, 1978). For cued recall tests (i.e., remembering that *this* cue word led to the generation of *that* target), findings show that self-generating also promotes a strengthening of the association between the target word and the cue used to generate the target word (i.e., enhanced cue-target relational processing; Donaldson & Bass, 1980; Hirshman & Bjork, 1988; Jacoby, 1978), leading to improved cued recall performance over reading.

Although limited work has shown that memory performance differs for items produced under varying levels of generation constraint (Fiedler et al., 1992; Gardiner et al., 1985; McCurdy et al., 2017, 2019), it is unknown what memory mechanism(s) might contribute to differences between lower- and higher-constraint generation tasks. McCurdy et al. (2017) found that lower-constraint generation led to better memory than higher-constraint for measures of cued recall, but both types of task often led to equivalent memory improvements (over read controls) for measures of recognition, suggesting enhanced memory via relational processing for lower-constraint tasks. In that study however, the cued recall test occurred immediately following an item recognition test, potentially confounding conclusions about differences in relational processing between lower- and higher-constraint. Therefore, the present study aimed to extend this prior work by testing recognition and cued recall performance independently in order to investigate the extent the memory benefits from lower- and higher-constraint generation could be tied to item-specific and relational processing.

In addition to item memory generation effects, other research has examined the generation effect for context memory (i.e., memory for extraneous episodic details associated with studied items). Prior studies on the generation effect for context memory have resulted in

mixed findings, with some showing context memory benefits (Geghman & Multhaup, 2004; Greenwald & Johnson, 1989; Marsh, 2006; Marsh, Edelman, & Bower, 2001), but others showing no benefit or even worse context memory for generating relative to reading (Mulligan, 2004, 2011; Mulligan, Lozito, & Rosner, 2006; Nieznański, 2014). A deeper analysis of these findings, however, shows that various types of contextual details have been examined across these studies. A theoretical perspective called the *processing account* (Jacoby, 1983; Mulligan, 2004, 2011) provides a framework to understand these mixed findings.

The processing account relies on transfer-appropriate processing principles (Morris, Bransford, & Franks, 1977) to predict which types of context memory details will be improved or impaired from generating compared to reading. For example, Jacoby (1983) suggests that the act of generation often requires conceptual processing (i.e., processing of conceptually related information to generate a target from a cue; e.g., antonym: open – c____), whereas reading often leads to perceptual processing (i.e., processing of visual features of the target item). Similarly, context memory details may also be categorized into conceptually-based details (e.g., source), and perceptually-based details (e.g., font color). The processing account predicts that generating (relative to reading) will improve memory for conceptually-based details because there is a greater overlap in processing between generation (conceptual processing) and recalling conceptual information (e.g., source). For perceptual details, however, the processing account predicts that generating will not improve memory for perceptually-based details because there is less overlap in processing between generating (conceptual processing) and recalling perceptual information (e.g., font color). This account has been supported by a number of studies showing that generation often improves context memory for conceptual-based details, but that this effect is often reduced or sometimes reversed (i.e., reading greater than generating) for perceptual-

based details (Geghman & Multhaup, 2004; Marsh, 2006; Marsh et al., 2001; Mulligan, 2004, 2011; Mulligan et al., 2006; Nieznański, 2011, 2012, 2014).

Interestingly, the limited work examining generation constraint has shown that the most robust and consistent effects of constraint (lower greater than higher) are for context memory (McCurdy et al., 2017, 2019). Specifically, McCurdy et al. (2017, 2019) had participants report in which task words were encoded (lower-constraint, higher-constraint, read) as their context (source) memory measure and consistently found a robust effect of generation constraint (lower-constraint greater than higher-constraint) across several experiments. Both of these studies, however, examined only one type of contextual detail (source), which could be classified as a conceptual detail according to the processing account view (Jacoby, 1983; Mulligan, 2004, 2011). Given that prior work has primarily used higher-constraint tasks in studying the generation effect for context memory, it remains unknown the extent that memory for other types of contextual details (e.g., perceptual details) may be affected by generation constraint. Thus, another aim of the present study was to examine how generation constraint influences the generation effect for different contextual details by comparing memory performance from a lower-constraint task to a higher-constraint task for both a conceptual (source) and perceptual (font color) detail. This design should provide clarity among two competing hypotheses. If the processing account (Jacoby, 1983; Mulligan, 2004, 2011) is supported, we should observe that lower-constraint generation improves context memory over higher-constraint generation *only* for conceptual details (e.g., source), but not perceptual details (e.g., font color). Alternatively, it is plausible that the enhanced memory benefits from lower-constraint generation extend to multiple types of contextual details (both conceptual and perceptual). Thus, finding that lower-constraint generation improves memory over higher-constraint generation for *both* conceptual (source) and

perceptual (font color) details would argue against the processing account, and instead support the idea that lower-constraint generation yields robust context memory support for a variety of contextual details.

Experiment 1

Experiment 1 was designed to test whether the item memory benefits for words produced in a lower- versus higher-constraint generation task are driven by the item-specific processing memory mechanism of the two-factor theory. In this experiment, we compared the memory benefits from a lower-constraint generation task to a commonly used higher-constraint generation task, and a read control task, using an item recognition test, which prior work suggests is sensitive to item-specific processing (Burns, 2006; Einstein & Hunt, 1980). Additionally, we tested memory for source (conceptual) and font color (perceptual) to examine the effects of generation constraint on memory for different types of contextual details.

We made three predictions for this experiment: First, for item recognition, we expected to find the standard generation effect (i.e., both generation tasks improve memory over reading), based on a large body of research suggesting that a variety of generation tasks (including ones of the type used in this study; Foley & Foley, 2007; McCurdy et al., 2017, 2019) improve item memory over read controls (Bertsch et al., 2007; McCurdy et al., Under Review). Second, for generation constraint, we expected that both generation tasks (lower- and higher-constraint) would yield similar memory as measured by item recognition, based on prior work suggesting that the effects of generation constraint are weaker using item recognition measures (McCurdy et al., 2017). Such a finding would suggest that both generation tasks (lower- and higher-constraint) lead to a similar increase in item-specific processing over read controls. Third, for context memory, we made separate predictions based on the type of context detail. For source context

memory (a conceptual detail), we expected to find a standard generation effect, based on the processing account view that generation improves memory for conceptual details over reading (Jacoby, 1983; Mulligan, 2004, 2011). In addition, we expected that lower-constraint generation would lead to better source memory compared to higher-constraint generation based on prior work (McCurdy et al., 2017, 2019). For color context memory (a perceptual detail), we expected to find no generation effect for the higher-constraint generation task, in line with the processing account view that generation often does not improve memory for perceptual details (Jacoby, 1983; Mulligan, 2004, 2011). For lower-constraint generation, however, we saw two possible outcomes: Based on the processing account, we expected that lower-constraint generation would provide no memory benefit for font color relative to read controls. Alternatively, given that prior work has shown stronger benefits for context memory from lower-constraint compared to higher-constraint generation tasks, we expected that lower-constraint generation might lead to better memory for font color compared to higher-constraint generation and read controls. This outcome would suggest that the enhanced memory benefits from lower-constraint generation extends to perceptual context memory. We aimed to distinguish between these two competing hypotheses.

Method

Participants

Twenty-seven adults (mean age: 19.6, *SD*: 1.1, 14 females) were recruited from the University of Illinois at Chicago introductory psychology subject pool and greater Chicago community. Three participants were excluded from all analyses for failing to follow instructions, leaving a total N of 24. An a priori power analysis (G*Power; Faul, Erdfelder, Buchner, & Lang, 2009) revealed that for a medium effect size (d = .55, based on effect sizes we have found in previous pair-wise comparisons; McCurdy et al., 2017, 2019), and a power level (1 – err prob.)

of .80, the recommended sample size was 22 for a one-tailed test. All participants gave their written informed consent in accordance with the University of Illinois at Chicago Institutional Review Board prior to participation. Participants were either given course credit or paid for their participation.

Materials

A total of 96 cue-target word pairs were selected from The University of South Florida word association norms (Nelson, McEvoy, & Schreiber, 2004). Word pairs were selected based on their forward cue-to-target association strength (FSG) such that all word pairs were highly associated (FSG > .50, Mean FSG = .624), similar to stimuli used in prior work (McCurdy et al., 2017, 2019). Prior work has shown that the generation effect is influenced by the initial strength of association between the cue and target word, often showing that low-associated pairs lead to larger generation effects, while high-associated pairs lead to smaller generation effects (Hirshman & Bjork, 1988; Taconnat, Froger, Sacher, & Isingrini, 2008). Therefore, we chose to use all highly-associated word pairs to reduce any strength of association effects that might have clouded our ability to investigate the primary variable of interest (generation constraint). All cue and target words contained between 4-7 letters and there was no clear relationship across any of the word pairs in the list. Across participants, word pairs were counterbalanced to create eight different word list versions such that each word pair occurred exactly once in each condition (generate, scramble, read, new) and as a green and red word in each of those four conditions.

Procedure

The experiment consisted of two phases, an encoding phase followed by a retrieval phase. Stimuli for both phases were presented using E-Prime software (version 2.0.8; Psychology Software Tools, 2012). Participants were trained on shortened versions of both

phases prior to the experiment to ensure they understood the task instructions. After training, participants started the encoding phase. During encoding, participants were shown 72 word pairs across three conditions, or tasks (generate, scramble, read; 24 words in each task). Half of the word pairs in each condition were presented in red font on a black background, whereas the other half were presented in green font (see **Figure 1** for a schematic of this phase). In the generate task (i.e., *lower-constraint generation*), participants were given a cue word followed by a blank line (e.g., option – _____). Participants were instructed to generate the first word that came to mind when they saw the cue word, and then to say both the cue and generated target aloud to the experimenter. Participants were told their responses in this task were subjective and there were no right or wrong answers. The experimenter recorded each response in the generate task to include these target items in the recognition test. In the scramble task (i.e., *higher-constraint generation*), participants were shown a cue word followed by a target that was scrambled (e.g., reply – asewrn). In this task, participants were instructed to use the cue to help them unscramble the target word, and then say both words aloud to the experimenter. The first letter of the scrambled target word was always in its correct position to reduce the number of skipped trials,[2] as done previously (Foley & Foley, 2007; Foley et al., 1989; McCurdy et al., 2017). In the read task (i.e., control), participants were shown a cue word followed by a fully intact target word (e.g., brief – short) and were instructed to read both words aloud. The tasks were performed in six blocks of 12 encoding trials each. All trials within a block were of the same task (e.g., scramble). Between blocks, participants were notified which task they were about to perform for 3000ms, "Get ready to do the Generate / Scramble / Read task." The order of the blocks was

[2] Pilot testing showed that using a scramble task where all letters of the target word were scrambled (reply – snarwe) led to significantly more "skipped" trials than a scramble task where the first letter was always in the correct place but the remaining letters scrambled (reply – asnrew).

randomized with the caveat that the same task was not performed in consecutive blocks. All trials were self-paced and each trial was separated by a 500ms fixation cross.

Following the encoding phase, participants filled out a short demographics questionnaire while the experimenter uploaded their responses from the generate task (lower-constraint) into E-Prime for the recognition test. This process and filler task took approximately 2 minutes. After the filler task, participants then completed the retrieval phase. In this phase participants were shown a total of 96 words (72 "old" *target* words from the encoding phase, 24 "new" distractor words not shown at encoding) in a random order that was different from the order of the encoding phase. The distractors were unrelated to the target words. For each recognition trial, participants were shown either an "old" target word (seen at encoding) or a "new" word and were asked to make three self-paced judgments corresponding to our three memory measures (item, source context, and color context; see **Figure 1**). First, participants judged whether the word was old, new, or whether they did not know, which served as the item recognition measure. Second, following a 500ms fixation, participants then judged whether they encountered that word in the generate, scramble, or read condition, or whether they did not know (source context recognition). Third, participants judged whether the word was presented in a red or green font at encoding, or whether they did not know (color context recognition). For items that were recognized as "new", participants were instructed to make "don't know" responses for the second and third retrieval decision, as done before (Leshikar & Duarte, 2012, 2014; Leshikar, Dulas, & Duarte, 2015; Leshikar & Gutchess, 2015; Leshikar et al., 2017; Leshikar, Park, & Gutchess, 2015; McCurdy et al., 2017). The "don't know" response was included as an option to reduce guessing (Duarte, Henson, & Graham, 2008; Duarte, Henson, Knight, Emery, & Graham,

2010). All memory decisions were made by pressing the V, B, N, and M keys (standard QWERTY keyboard) that corresponded to different response options.

Results

Participants gave responses on 100% of the trials in the generate and read tasks, and 98% of the scramble trials. Unsuccessful scramble trials (e.g., participant could not unscramble the word) were removed from all analyses. Response rates for item, source context, and color context recognition, as well as responses to new items, are reported in **Table 1**. We conducted two sets of analyses on this data (subject-based, item-based). First, using *subjects* as the unit of analysis, we conducted a one-way repeated-measures analysis of variance (ANOVA) for each memory measure (item recognition, source context, color context) separately to examine differences in memory between the three tasks (generate, scramble, read). Item recognition was calculated as the percentage of items seen at encoding correctly identified as "old" for each task. Source context and color context performance were calculated using the conditional source identification measure (CSIM; Murnane & Bayen, 1996), which reduces the influence of item memory performance on context recognition. Specifically, source context memory performance was calculated as the proportion of items correctly identified as "old" in item recognition (i.e., item hits) that were also correctly identified with their source (source correct / item correct). Similarly, color context memory performance was calculated as the proportion of items correctly identified as "old" that were also correctly identified with their encoded font color (color correct / item correct). Estimates from the subject-based analyses using these calculations for item recognition, source context, and color context are graphed in **Figure 2**. All subject-based ANOVAs and subsequent follow-up analyses were performed using JASP statistical software (JASP Team, 2018).

Researchers have often used highly constrained generation tasks in prior work to combat against what are known as item-selection confounds (i.e., differences among the items being encoded that can influence memorability of that item). We recognize that this type of confound may be particularly problematic in our design because the lower-constraint task allows participants to respond freely, potentially producing more idiosyncratic responses (e.g., generating "open – door", instead of the normed response of "open – close"), compared to the higher-constraint task. Thus, in each experiment we conducted an additional analysis using each *trial* as the unit of analysis, to control for and examine the influence of item-selection confounds on our analyses comparing the lower- and higher-constraint generation task. Specifically, we conducted a logistic (logit) regression analysis for each memory test, that predicts the likelihood of correctly remembering an item, source, or color, based on the type of task (generate, scramble, or read) and including as a control variable whether or not the subject-generated target matched the normed expected target (non-matched, matched).[3] This analysis allowed us to examine the effect of generation constraint (lower- versus higher-constraint) on memory performance, while controlling for (and measuring) the influence of item-selection confounds in the data. We report the results of these trial-based analyses after the subject-based analyses in the results of each experiment.[4] All regression analyses were done in R 3.5.2 (R Core Team, 2018), using the *lme4* (version 1.1-21; Bates, Mächler, Bolker, & Walker, 2014) package.

[3] In Experiment 1, participants generated the expected normed response 58.4% (*SD* = 14%) of the time across all generate trials.

[4] Another method of controlling for item-selection confounds is to perform ANOVAs on the trial-level data (analogous to the subject-based ANOVAs) and removing any trial where the participant did not generate expected target item. These trial-level ANOVAs (not reported) led to similar results as the logistic regression analysis, therefore, we chose to report the logistic regression because it does not require removing any trials, and provides a way to examine and control for any differences in memory performance between non-matched and matched trials.

Subject-Based Analyses

For item recognition, we found significant differences between the three tasks, $F(2, 46) = 60.11$, $MSE = 0.01$, $p < .001$, $\eta^2_p = .72$. Planned follow-up analyses revealed that both the generate ($M = .81$, $SD = .15$), $t(23) = 8.51$, $p < .001$, $d = 1.74$, and scramble ($M = .86$, $SD = .11$), $t(23) = 9.60$, $p < .001$, $d = 1.96$, tasks led to better memory than read controls ($M = .52$, $SD = .16$), consistent with the standard generation effect. A paired sample *t*-test comparing the generate and scramble tasks showed no significant difference between the tasks, $t(23) = -1.55$, $p = .14$, $d = -0.32$.[5]

For source context memory, we found significant differences between the tasks, $F(2, 46) = 22.08$, $MSE = 0.02$, $p < .001$, $\eta^2_p = .49$. Planned follow-up analyses revealed that source recognition was better in both the generate ($M = .83$, $SD = .16$), $t(23) = 5.92$, $p < .001$, $d = 1.21$, and scramble ($M = .69$, $SD = .16$), $t(23) = 2.78$, $p < .011$, $d = .57$, tasks compared to the read task ($M = .57$, $SD = .20$), supporting a standard generation effect for source context memory. Additionally, we found that source recognition was significantly better in the generate task compared to the scramble task, $t(23) = 4.85$, $p < .001$, $d = .99$.

For color context memory, we found significant differences between the tasks, $F(2, 46) = 4.27$, $MSE = 0.02$, $p = .02$, $\eta^2_p = .16$. Planned follow-up tests revealed that the generate task ($M = .42$, $SD = .20$) led to significantly better color recognition compared to the read task ($M = .32$, $SD = .22$), $t(23) = 2.72$, $p = .012$, $d = .56$. There were no differences between the scramble ($M = .38$, $SD = .18$) and read tasks, $t(23) = 1.41$, $p = .17$, $d = .29$, or between the generate and scramble tasks, $t(23) = 1.75$, $p = .09$, $d = .36$.

[5] Traditional frequentist statistics cannot be used to assess the likelihood of the null hypothesis, therefore, we also analyzed this data comparing the generate and scramble tasks by estimating a Bayes factor ($BF_{01} = 1.58$). This analysis indicates that these data are roughly 1.5 times more likely to occur under the null hypothesis compared to the alternative hypothesis.

Trial-Based Analyses

The coefficient summary tables from the logistic regression analyses are shown in **Table 2**. For item recognition, the model provided a significantly better fit than the null model (a model with no predictors), x^2 (3) = 184.7, $p < .001$. Results showed that both the generate and scramble tasks led to a significantly greater likelihood of remembering an item relative to read controls, x^2 (2) = 158.0, $p < .001$. A Wald chi-square test comparing the generate task to the scramble task showed that the scramble task led to significantly greater likelihood of recognition, x^2 (1) = 4.2, p = .039, when controlling for whether the subject-generated item matched the expected normed target (i.e., item-selection confounds). Although we did not see this result in the subject-based analyses, the subject-level data showed a similar pattern of results (higher-constraint greater than lower-constraint), but did not come out as statistically significant. The *norm matched target* coefficient of the model is important because it provides some evidence about the influence of item-selection confounds in the data. In this analysis, the coefficient was not significant, x^2 (1) = 0.33, p = .57, indicating that there was no difference in the likelihood to remember an item for the subject-generated responses that did not match the normed target items compared to responses that did match the normed target.

For source context memory, the model was a significantly better fit than the null model, x^2 (3) = 65.39, $p < .001$. Results showed that both the generate and scramble tasks led to a greater likelihood of correctly remembering the source of an item relative to read controls, x^2 (2) = 31.8, $p < .001$. A Wald chi-square test showed participants were significantly more likely to remember the source of an item in the generate task compared to the scramble task, x^2 (1) = 11.8, $p < .001$, confirming our subject-based findings even when controlling for whether the participant generated the expected normed target. The norm matched target coefficient was significant in

this model, x^2 (1) = 4.1, p = .042, indicating that participants were less likely to remember the source of an item when the subject-generated response matched the normed expected target.

For color context memory, the model was only a marginally better fit than the null model, x^2 (3) = 6.35, p = .095. The results showed that the generate task led to significantly greater likelihood of remembering the color relative to reading, x^2 (1) = 4.8, p = .028, while the scramble task did not, x^2 (1) = 2.0, p = .15, confirming our subject-based results. The *norm matched target* coefficient was not significant, x^2 (1) = 0.00, p = .98, indicating that item-selection confounds did not strongly influence our color memory findings.

Experiment 1 Discussion

In Experiment 1, there were three major findings. First, both generation tasks led to better item memory compared to read controls, in line with the standard generation effect (Bertsch et al., 2007; McCurdy et al., Under Review). For generation constraint, the subject-based analyses revealed that item recognition did not differ between the lower- and higher-constraint generation tasks (while the trial-based analyses showed better recognition for higher- compared to lower-constraint). Although null results should be interpreted cautiously, equivalent memory performance between lower- and higher-constraint generation suggests that both generation tasks enhanced item-specific processing equivalently over read controls. This finding is consistent with our predictions, and is in line with prior work showing that varying generation constraint often does not lead to reliable differences in measures of item recognition (McCurdy et al., 2017, 2019). Thus, it seems likely that the memory benefits from lower-constraint generation cannot be tied to enhanced item-specific processing relative to higher-constraint generation. Indeed, after controlling for item-selection confounds in our trial-based analyses, we instead found that higher-constraint generation led to a small advantage over lower-constraint generation,

indicating that it may be more likely that higher constraints enhance item-specific processing relative to lower constraints. Future work might examine this idea further using other direct indices of item-specific processing (e.g., items per category, cumulative recall; Burns, 2006) to corroborate our trial-based findings.

Second, for source context we found evidence of a generation effect for both generation tasks. Additionally, we found that source context was better in the lower-constraint task over the higher-constraint task, even after controlling for item-selection confounds, in line with prior work showing lower-constraint generation tasks lead to robust source memory benefits over higher-constraint generation tasks (McCurdy et al., 2017, 2019). Research suggests that self-generation may improve memory for contextual details by binding relational details (e.g., the target, the cue used to generate that word, source, etc.) into a single memory representation (Greenwald & Johnson, 1989; Marsh, 2006; Marsh et al., 2001), which in turn improves accessibility to those details at retrieval (Bower, 1970; Hunt & Einstein, 1981). If this is the case, greater relational processing at encoding should lead to increased binding of these contextual details. Our source memory findings support this idea and suggest that fewer constraints may act to enhance relational processing at encoding (a proposition we aimed to further test in Experiment 2).

Third, our color context analysis showed that lower-constraint generation significantly improved color recognition over reading, but that higher-constraint did not. In contrast to the processing account (Jacoby, 1983; Mulligan, 2004, 2011), our results provide some evidence that self-generation may lead to enhanced color memory (a perceptual context detail) compared to reading when a lower-constraint task is used. Similar to source memory, it could be argued that font color is another detail that can be bound to the memory representation of the word pair, and

thus is improved by tasks that induce greater relational encoding (Greenwald & Johnson, 1989; Marsh et al., 2001). Furthermore, in conjunction with our source memory findings, this color memory effect suggests that multiple contextual details may be improved under lower- compared to higher-constraint generation. Perhaps not surprisingly, color recognition performance was notably lower than the other memory measures. It could be that color is a feature that is not as easily bound into a memory representation as other contextual details (e.g., source). That is, it might be easy to remember the cognitive process you engaged in to produce an item (i.e., source memory), but more difficult to integrate color detail into memory.

Overall, Experiment 1 generally supported our prediction of equivalent memory improvement attributable to item-specific processing and hinted that lower-constraint generation may enhance relational processing to a greater extent than higher-constraint as implied by the source and color memory findings. We further examined the possibility that lower-constraint generation enhances relational processing in our second experiment.

Experiment 2

Experiment 2 was designed to test whether the item memory benefits for materials produced in a lower- versus higher-constraint generation task may be attributed to increased relational processing, the second memory mechanism of the two-factor theory. We replaced the item recognition test used in Experiment 1 with a cued recall test, which prior work suggests is primarily sensitive to relational processing (Donaldson & Bass, 1980; Hirshman & Bjork, 1988; Underwood & Schulz, 1960). Additionally, we tested memory for source and color context to extend our findings of Experiment 1 comparing the effects of generation constraint on memory for different types of contextual details.

We made three predictions for Experiment 2. First, for cued recall, we expected to find the standard generation effect across both generation tasks, based on prior work showing generation effects are often robust using cued recall measures (Donaldson & Bass, 1980; Hirshman & Bjork, 1988). Second, for generation constraint we expected that a lower-constraint task would lead to better cued recall performance compared to higher-constraint, which would support the idea that lower-constraint generation increases the generation effect via enhanced relational processing relative to higher-constraint tasks. We made these predictions based on prior work showing reliable memory benefits for lower- compared to higher-constraint generation tasks when measured by recall tests (Fiedler et al., 1992; Gardiner et al., 1985; McCurdy et al., 2017). Third, for context memory, we made the same predictions as Experiment 1, as we aimed to substantiate our conclusions about the influence of generation constraint on multiple contextual details.

Method

Participants

Twenty-seven adults (mean age: 19.1, *SD* = 1.6, 10 females) were recruited from the University of Illinois introductory psychology subject pool. Like Experiment 1, an a priori power analysis (*d* = .55; power = .80; one-tailed test) recommended a sample size of 22. All participants gave their informed written consent in accordance with the University of Illinois at Chicago Institutional Review Board and were given course credit for participation.

Materials

The identical 96 cue-target word pairs and counterbalanced word lists were used as in Experiment 1 to minimize the differences between the two experiments. Words occurring in the

"new" condition for each counterbalanced list version were not presented, given that no distractor items were used in this experiment.

Procedure

The procedure for Experiment 2 was identical to Experiment 1 with two changes at retrieval. First, after encoding participants were given a paper and pencil cued recall test that replaced the recognition test of Experiment 1. Second, participants were trained on the encoding phase, but were not trained on the cued recall phase, thus participants were unaware of the impending memory test (i.e., an incidental memory task) and subsequently the details they would be asked to remember, which we considered to be a stricter test of whether context memory is influenced by generation constraint. This procedure differed from Experiment 1, where participants were trained on the recognition phase, and thus, knew which details they would be asked to remember (i.e., intentional memory procedures).

For cued recall, participants were given a sheet of paper that listed the 72 cue words from the encoding phase in a random order that was different from the encoding phase. Each cue word was followed by a blank space. To the right of the blank space, each of the three conditions were listed (generate, scramble, read), and to the right of that, both font colors were listed (red, green; see **Supplemental Figure 1** for an example of the cued recall test format). Participants were instructed to write the word that was paired with the cue word from the encoding phase onto the blank line (cued recall). Then they were asked to circle which task they encoded the word pair in (source context recognition), as well as the font color the pair was presented in at encoding (color context recognition). Participants were instructed to leave the line blank, and to not circle the source or the font color if they were unsure of any of these details to reduce guessing. The

instructions to leave unknown items blank were designed to parallel the "don't know" response option used in the recognition test in Experiment 1.

Results

Participants gave responses on 100% of the encoding trials in the generate and read tasks, and 95% of the scramble trials. Unsuccessful scramble trials were removed from all analyses. Raw response rates for cued recall, source context and color context are reported as a function of encoding condition in **Table 1**. As in Experiment 1, we conducted two sets of analyses (subject-based, trial-based). First, using *subjects* as the unit of analysis, we conducted a one-way repeated measures ANOVA for each memory test separately (cued recall, source context, color context). Cued recall scores were calculated as the proportion of items seen at encoding correctly recalled for each task. Source and color context recognition were calculated using the same CSIM procedure described in Experiment 1. Estimates from the subject-based analyses using these calculations for cued recall, source context, and color context are graphed in **Figure 3.** Second, as in Experiment 1, we conducted a logistic regression using *trials* as the unit of analysis to examine and control for potential item-selection confounds. We ran a separate model for each memory test that included the encoding task and norm matched target[6] variables as predictors. The coefficient summary tables from these analyses are shown in **Table 2**.

Subject-Based Analyses

Results of the cued recall ANOVA revealed significant differences between tasks, $F(2, 52) = 48.45$, $MSE = 0.01$, $p < .001$, $\eta^2_p = .65$. Planned follow-up analyses showed that recall was significantly higher in the generate ($M = .88$, $SD = .12$), $t(26) = 8.65$, $p < .001$, $d = 1.66$, and

[6] In Experiment 2, participants generated the expected normed response 54.9% (SD = 14%) of the time across all generate trials.

scramble tasks (M = .82, SD = .17), $t(26) = 6.60$, $p < .001$, $d = 1.27$, compared to the read task (M = .60, SD = .19), in line with the standard generation effect. Recall was also significantly higher in the generate task compared to the scramble task, $t(26) = 2.50$, $p = .02$, $d = 0.48$, indicating that lower-constraint generation (generate) significantly improved cued recall over a higher-constraint task (scramble).

The source context memory analysis revealed our data violated the equal variances assumption, therefore, we used Greenhouse-Geisser corrected degrees of freedom. The results showed significant differences between the tasks, $F(1.60, 41.66) = 23.42$, $MSE = 0.03$, $p < .001$, $\eta^2_p = .47$. Planned follow-up analyses revealed that source memory was greater for both the generate (M = .95, SD = .05), $t(26) = 8.87$, $p < .001$, $d = 1.71$, and scramble (M = .80, SD = .21), $t(26) = 2.86$, $p = .008$, $d = .55$, tasks compared to the read task (M = .65, SD = .20). Further, the generate task led to better source memory compared to the scramble task, $t(26) = 3.48$, $p = .002$, $d = .67$.

The color context memory analysis revealed no significant differences between the three tasks, $F(2, 52) = 0.97$, $MSE = 0.008$, $p = .39$, $BF_{01} = 4.48$, indicating that generation did not improve memory for color in this experiment, regardless of constraint.

Trial-Based Analyses

For cued recall, the model provided a significantly better fit than the null model (a model with no predictors), x^2 (3) = 164.2, $p < .001$. Results showed a robust generation effect for both the generate and scramble tasks in the likelihood to recall an item correctly relative to read controls, $x^2(2) = 131.2$, $p < .001$. A Wald chi-square test confirmed our subject-based finding that the generate task led to a greater likelihood of an item being recalled compared to the scramble task, x^2 (1) = 18.1, $p < .001$, when controlling for whether the subject-generated

response matched the normed target or not. The *norm matched target* coefficient was, however, significant, x^2 (1) = 12.5, $p < .001$, indicating that there was a greater likelihood of correct recall for the subject-generated responses that did not match the normed target items compared to those that did match the expected target. Despite this finding, we still found that the generate task was greater than five times more likely to be recalled compared to the scramble task when controlling for this factor (see **Table 2** for odds ratios), indicating that although item-selection confounds may have influenced our subject-based findings to some degree, the effect of generation constraint was still robust.

For source context memory, the model was a significantly better fit than the null model, x^2 (3) = 136.08, $p < .001$. Results showed that both the generate and scramble tasks led to a greater likelihood of correctly remembering the source of an item relative to read controls, $x^2(2) = 73.6$, $p < .001$. A Wald chi-square test revealed that participants were significantly more likely to remember the source of an item in the generate task compared to the scramble task, $x^2(1) = 13.3$, $p < .001$, when controlling for whether or not the participant generated the expected normed target. The *norm matched target* coefficient was again significant in this model, $x^2(1) = 4.1$, $p = .040$. As in Experiment 1, the coefficient was negative, indicating that participants were less likely to remember the source of an item when the subject-generated response matched the normed expected target.

For color context memory, the model was not a better fit than the null model, $x^2(3) = 1.70$, $p = .636$, supporting our subject-based results showing no differences in the likelihood to remember an item's font color across the three conditions.

Cross-Experiment Analysis

A key finding from the analyses of Experiments 1 and 2 is the different effects of generation constraint on item memory based on the type of memory test used (recognition, cued recall). In Experiment 1 we found no significant differences between the generate (lower-constraint generation) and scramble (higher-constraint generation) tasks using a recognition test, but in Experiment 2 we found that the lower-constraint (generate) task led to better item memory compared to the higher-constraint (scramble) task using a cued recall test. Given the importance of these different effects to our claim regarding the mechanism underlying the item memory benefits from lower-constraint generation, we followed up these analyses with a 2 (task: lower-constraint, higher-constraint) x 2 (memory test: recognition, cued recall) mixed ANOVA comparing item memory performance between the lower-constraint (generate) and higher-constraint (scramble) task based on the type of memory test (recognition, cued recall). The results of this analysis revealed no main effects for task, $F(1, 49) = 0.72$, $p = .79$, or memory test, $F(1, 49) = 0.26$, $p = .62$, but critically, the task by memory test interaction was significant, $F(1, 49) = 7.71$, $MSE = .009$, $p = .008$, $\eta^2_p = .14$. This interaction is depicted in **Figure 4**, which shows that the effect of generation constraint was significantly different between the recognition test and cued recall test. Specifically, lower-constraint generation led to better memory than higher-constraint on the cued recall test, but not on the recognition test. This result corroborates the general conclusions from Experiments 1 and 2 that lower-constraint generation increases the generation effect through enhanced relational processing, but not item-specific processing. It should be noted, however, that the results of this cross-experiment analysis should be interpreted with some caution. These experiments were designed to detect pairwise differences, and thus are slightly underpowered to detect a cross-experiment interaction, increasing the likelihood of false discovery of this effect.

Experiment 2 Discussion

The results of Experiment 2 led to three major findings. First, for item memory, our cued recall test revealed a standard generation effect (both lower- and higher-constraint produced better memory than read controls), but importantly showed that lower-constraint generation led to better recall than higher-constraint generation. This finding adds to the growing body of research indicating the effects of generation constraint are reliable when measured by cued recall (McCurdy et al., 2017), and extends this previous work by confirming the effect is robust even when measured independently from an earlier recognition test, and after controlling for item-selection confounds. Given that cued recall tests are primarily sensitive to relational information (i.e., the relationship between the cue and target word; Donaldson & Bass, 1980; Hirshman & Bjork, 1988; Underwood & Schulz, 1960), our finding of improved recall for lower- over higher-constraint generation suggests that fewer generation constraints may serve to improve relational processing over and above higher-constraint generation tasks. This is an important finding because it implicates a potential mechanism by which reduced generation constraints increase the size of the generation effect.

Second, we again found an effect of generation constraint on source context memory, where lower-constraint generation increased source memory accuracy relative to higher-constraint, consistent with Experiment 1 and adding to the growing body of work on memory effects for this detail (McCurdy et al., 2017, 2019). Additionally, we also found the standard generation effect (i.e., lower- and higher-constraint > read) for this detail, providing more evidence that enhanced relational processing from self-generation may support the binding of conceptual context details into a single memory representation, making that item and other

conceptually-related details more likely to be retrieved (Greenwald & Johnson, 1989; Marsh, 2006; Marsh et al., 2001).

Third, for color context we found no effects across the three conditions. This result contrasts with our findings from Experiment 1, where lower-constraint generation led to improved memory for font color. This discrepancy between experiments may be explained by notably poor performance for this measure in Experiment 2, perhaps due to the incidental nature of the memory test procedure used in this experiment. Unlike Experiment 1, participants were not trained on the cued recall test prior to the encoding phase as they were in Experiment 1, making Experiment 2 an incidental memory task. The difference in intentional versus incidental encoding has been recognized as a factor that can influence the magnitude of the generation effect (Bertsch et al., 2007), and may explain why our findings were different for this memory measure between these two experiments. Although inconsistent with Experiment 1, the findings in Experiment 2 (no generation effect for font color) are in line with the processing account (Jacoby, 1983; Mulligan, 2004, 2011) which suggests generation provides no benefit for perceptual context details because there is little overlap between the encoding and retrieval processes.

Experiment 3

In Experiments 1 and 2 we found evidence that lower-constraint generation improves the generation effect through enhanced relational processing, while item-specific processing seemed generally uninfluenced by generation constraint. In a third experiment, we aimed to substantiate these findings by examining the influence of generation constraint on *distinctiveness* (Hunt, 2012; Hunt & Worthen, 2006). Distinctiveness has been conceptualized as the combined effect of item-specific *and* relational processing (Burns, 2006; Hunt, 2006) and can be used as an

explanatory concept for how we make accurate memory judgments. Specifically, relational information is used to bring to mind similar items that were encoded in a given episode. Then, after bringing to mind the relevant episode, item-specific information is used to specify (or distinguish) the particular target item within that episode in order to make accurate memory judgments (Hunt, 2006, 2012). One way to measure distinctiveness is using a recognition test with related distractor items, because when related items are used as distractors, both relational and item-specific information become critical to make accurate memory judgments (Burns, 2006; Hunt & Seta, 1984). In contrast, when unrelated distractors are used (as in Experiment 1), accurate memory judgments can be made from item-specific information alone (because there is less use for relational information when the target and distractor items are unrelated). Prior work using recognition tests with related distractor items has shown that increased relational processing at encoding often leads to higher hit rates (i.e., correctly identifying an "old" item as being presented at encoding), but also increases false alarm rates (i.e., incorrectly endorsing a related distractor item as being presented at encoding; Huff & Bodner, 2013; Hunt, Smith, & Dunlap, 2011). Thus, if lower-constraint generation improves the generation effect through relational processing, but not item-specific processing, as our first two experiments suggest (i.e., there is equivalent item-specific processing for both lower- and higher-constraint), we expected that lower-constraint generation would lead to increased item hits and *greater* false alarms compared to the higher-constraint task. Alternatively, if fewer generation constraints increase distinctiveness (i.e., increases in *both* relational and item-specific processing, and not just one or the other), we expected that lower-constraint generation would lead to increased hit rates and *fewer* false alarms compared to the higher-constraint task. Therefore, in this experiment we examined both item hits and false alarms between the lower-constraint and higher-constraint

tasks to test these competing hypotheses and substantiate our conclusions from Experiments 1 and 2.

A second goal of Experiment 3 was to clarify our findings on context memory for font color. In the first two experiments, we found mixed findings for font color memory, which could be attributed to differences in intentional versus incidental memory procedures. Given the important implication of finding a generation task that might improve context memory for a perceptual detail, in this third experiment we aimed to reconcile the contrasting results from Experiments 1 and 2 by further examining the influence of generation constraint on memory for font color using intentional encoding procedures.

Method

Participants

Thirty-six adults (mean age: 18.9, *SD* = 1.9, 26 females) were recruited from the University of Illinois introductory psychology subject pool. Four participants were excluded from all analyses, two for failing to follow instructions and two for overall memory performance (across all tasks) greater than three standard deviations below the sample mean, leaving a total N of 32. As in Experiments 1 and 2, an a priori power analysis (*d* = .55, power = .80, one-tailed test) recommended a sample size of 22. We increased our sample size from the previous experiments given our prior work has not examined effects in false alarm data, which may be smaller than the effects we have seen previously. All participants gave their informed written consent in accordance with the University of Illinois at Chicago Institutional Review Board and were given course credit for their participation.

Materials

In this experiment, our design required related distractors, thus we used category-exemplar stimulus pairs (e.g., type of vegetable – carrot). Specifically, 48 unique categories were selected from the Van Overschelde, Rawson, and Dunlosky (2004) category norms. The two most highly associated exemplars from each of the 48 categories were used as target and distractor items for the recognition test. Across participants, categories and their exemplars were counterbalanced to create six total list versions so that each category occurred in each condition (generate, scramble, read) and both the first and second exemplars served as the target and distractor item exactly once in each condition. For example, for the category "type of fruit", the two strongest associated exemplars were "apple" and "orange". In one list, participants studied "type of fruit – apple", with "orange" serving as the distractor item. In a second list, participants studied "type of fruit – orange", with "apple" serving as the distractor item at recognition.

Procedure

Experiment 3 consisted of two phases (encoding, retrieval). Stimuli for both phases were presented using E-Prime software (version 2.0.8; Psychology Software Tools, 2012). Participants were trained on shortened versions of both phases before starting the experiment to ensure that they understood the instructions and that they were aware of the details they would be asked to remember (i.e., intentional encoding). After training, participants began the encoding phase. In the encoding phase, participants were shown a total of 48 category-exemplar pairs across the three conditions (generate, scramble, read; 16 categories in each). Half (8) of the category-exemplar pairs were presented in green font, and the other half (8) in red font, in a random order. In the generate task (i.e., *lower-constraint generation*), participants were given a category followed by a blank line (e.g., A unit of time – _____). Participants were instructed to think of a single word that fit the category shown, and then say both the category and the target word aloud

to the experimenter. The experimenter recorded each response in the generate task to include these items in the recognition test. In the scramble task (i.e., *higher-constraint generation*), participants were shown a category followed by a word that fit the category that was scrambled (e.g., An occupation – dcorot). Participants were instructed to unscramble the word that fit the category, and say both the category and word aloud. As in the first two experiments, the first letter of the scrambled target word was in its correct position to reduce the number of skipped trials. In the read task (i.e., control), participants were shown a category and an intact target word that fit the category (e.g., A kitchen utensil – fork), and were instructed to read both the category and word aloud. The tasks were performed in six blocks of 8 encoding trials each. As in the first two experiments, all trials within a block were of the same task, and before each block participants were instructed which task to perform for 3000ms ("Get ready to do the Generate / Scramble / Read task"). The order of the blocks was randomized with the caveat that the same task was not performed consecutively. All trials were self-paced and separated by a 500ms fixation cross.

Following the encoding phase, participants were given a non-verbal filler task (digit symbol substitution) for approximately 2 minutes while the experimenter uploaded their responses from the generate task into the E-Prime software for the recognition test. After the filler task, participants began the retrieval phase. Participants were shown a total of 96 words, one at a time, in a random order that was different from the encoding phase. Half (48) of the words, were "old" *target* words from the encoding phase, and the other half (48) were "new" *distractor* words. The critical difference between this recognition test and the one used in Experiment 1, was that the *distractor* items were related exemplars from the same 48 categories the participants saw at encoding. Participants were shown a single word on the screen, and asked

to make three forced-choice recognition judgments (item, source, color). For the item memory judgment, participants were asked to decide whether the word shown was encountered during the encoding phase or not (old, new). If the participant responded "old" to the word, they were asked to make two additional judgments on the source and color associated with that item at encoding. For the source memory judgment, participants decided which condition they had encoded the word (generate, scramble, read). For the color memory judgment, participants decided which font color the target word and category were encoded in (green, red). All memory decisions were made by pressing the V, B, and N keys (standard QWERTY keyboard) corresponding to different response options. After the retrieval phase, participants filled out a demographics questionnaire and were debriefed.

Results

Participants gave responses on 96% of the encoding trials in the generate task, 94% of trials in the scramble task, and 100% of the trials in the read task. Unsuccessful generate and scramble trials were removed from all analyses. Additionally, four categories (a family relative, a female first name, a male first name, a type of ship/boat) led to significantly greater number of non-normed and skipped trials in both the generate and scramble tasks (p's < .01), and therefore all trials from these four categories were removed from all analyses (subject-based and trial-based) across all participants and conditions (leaving a total of 44 categories included in the analyses). Hits and false alarm rates for item recognition, and raw response rates for source context and color context are reported as a function of encoding condition in **Table 1**. As in the first two experiments, we conducted two sets of analyses (subject-based, trial-based). First, using *subjects* as the unit of analysis, we conducted a one-way repeated measures ANOVA for each memory measure separately (item memory hits, item memory false alarms, source context, color

context). Item memory hits were calculated as the proportion of correctly identified target words out of the 16 targets studied for each task. Item memory false alarms were calculated as the proportion of distractor words falsely identified as "old" out of the 16 total distractors per task. Target and distractor words were paired one-to-one, thus for each target word there was a corresponding distractor word from the same category. This allowed us to classify each distractor word to a particular task (generate, scramble, read), giving a unique false alarm rate for each task that we used in the false alarm analysis. Source and color context recognition were calculated using the same CSIM procedure as in Experiments 1 and 2. Estimates from the subject-based analyses using these calculations for source context, and color context are graphed in **Figure 5.** Then, as with Experiments 1 and 2, we performed a logistic regression on each of the memory measures (item hits, false alarms, source context, color context) using the trial-level data to control for and examine potential item-selection confounds. The regression model included the encoding task and norm matched target variables as predictors.[7] The coefficient summary table for the trial-based analyses are reported in **Table 2**.

Subject-Based Analyses

For item recognition hits, we found significant differences between the tasks, $F(1.5, 47.6) = 70.05$, $MSE = 0.73$, $p < .001$, $\eta^2_p = .69$. Follow-up analyses revealed item hits were better for both the generate ($M = .96$, $SD = .06$), $t(31) = 10.31$, $p < .001$, $d = 1.82$, and scramble ($M = .92$, $SD = .12$), $t(31) = 7.97$, $p < .001$, $d = 1.41$, tasks compared to the read task ($M = .66$, $SD = .17$). A comparison of the two generation tasks showed the generate task led to significantly better item recognition compared to the scramble task, $t(31) = 2.11$, $p = .04$, $d = .37$.

[7] In Experiment 3, participants generated the expected normed response 49.9% (SD = 10%) of the time across all generate trials.

For false alarms, however, we found no significant differences between the three tasks, $F(1.56, 48.50) = 1.28$, $p = .282$, $\eta^2_p = .04$, $BF_{01} = 3.59$, indicating that generation (regardless of constraint) did not reduce false alarms relative to reading. It should be noted that false alarms were at floor, however, potentially obscuring our ability to detect effects between tasks.

For source context memory, we found significant differences between the tasks, $F(1.5, 44.9) = 8.73$, $MSE = 0.16$, $p = .002$, $\eta^2_p = .22$. Follow-up analyses revealed source memory was better for both the generate ($M = .97$, $SD = .07$), $t(31) = 3.61$, $p = .001$, $d = 0.64$, and scramble ($M = .92$, $SD = .08$), $t(31) = 2.27$, $p = .03$, $d = 0.40$, tasks compared to the read task ($M = .85$, $SD = .18$). Comparison of the two generation tasks showed that the generate task led to better source memory compared to scramble task, $t(31) = 2.49$, $p = .02$, $d = 0.44$, providing more evidence that lower-constraint generation improves memory over higher-constraint generation for conceptual context details.

For color context memory we found no significant differences between the three tasks, $F(2, 62) = 1.94$, $MSE = 0.02$, $p = .15$, $BF_{01} = 3.75$.

Trial-Based Analyses

For item recognition, the model provided a significantly better fit than the null model, $x^2 (3) = 178.4$, $p < .001$. Results showed a robust generation effect for both the generate and scramble tasks in the likelihood to correctly remember an item as being studied at encoding, relative to read controls, $x^2 (2) = 108.0$, $p < .001$. A Wald chi-square test found that the generate task led to a marginally greater likelihood of an item being remembered compared to the scramble task, $x^2 (1) = 3.6$, $p = .058$, when controlling for whether the subject-generated response matched the normed target or not. The *norm matched target* coefficient, however, was

not significant, x^2 (1) = 2.1, p = .15, indicating item-selection confounds did not strongly influence this data.

For false alarms, the model fit marginally better than the null model, x^2 (3) = 6.90, p = .075. The results showed that the generate task led to a significantly higher likelihood of false alarms, compared to both the read task, x^2 (1) = 5.1, p = .024, and the scramble task, x^2 (1) = 5.7, p = .017. There was no difference in the likelihood of false alarms between scramble and read, x^2 (1) = 0.05, p = .82. The *norm matched target* coefficient was not significant, x^2 (1) = 2.5, p = .11.

For source context memory, the model was a significantly better fit than the null model, x^2 (3) = 32.67, p < .001. Results showed a different pattern of results for the generate and scramble tasks. The generate task led to a greater likelihood of correctly remembering the source of an item relative to read controls, x^2 (1) = 16.6, p < .001, and the scramble task, x^2 (1) = 4.5, p = .035. The scramble task showed the opposite pattern, where participants were less likely to remember the source of an item compared to the read task, x^2 (1) = 9.8, p = .002. The *norm matched target* coefficient was not significant in this model, x^2 (1) = 0.03, p = .86, indicating that item-selection confounds did not strongly influence source memory performance in this experiment.

For color context memory, the model was not a better fit than the null model, x^2 (3) = 3.67, p = .30, indicating that there were no differences in the likelihood to remember an item's font color across the three conditions.

Experiment 3 Discussion

Experiment 3 had three main findings. First, we found that lower-constraint generation led to better item memory hits than higher-constraint generation. We also found some evidence that the lower-constraint generation task led to increased rates of false alarms when controlling

for item-selection effects (trial-based analyses). Taken together, these findings suggest that lower-constraint generation did not enhance distinctiveness, but rather led to enhanced relational processing, while leaving item-specific processing relatively unaffected compared to higher-constraint generation. These results substantiate our overall thesis that lower-constraint generation increases the generation effect through enhanced relational processing relative to higher-constraint generation. Second, we found additional evidence that generation (regardless of constraint) improves source memory (a conceptual detail) over read controls, but that lower-constraint generation improves source memory to a greater extent than higher-constraint generation. This finding further promotes the idea that fewer constraints enhance relational processing, resulting in greater binding of conceptually related details. Third, we found no evidence of a generation effect for color context memory, and no differences based on generation constraint. This finding offers some clarity to our contrasting findings in the first two experiments, and supports the idea that generation (regardless of constraint) does not seem to enhance memory for perceptual context details. Overall, these context memory results are in line with the processing account (Jacoby, 1983; Mulligan, 2004, 2011), which suggests generation improves memory for conceptual context details, but not perceptual context details.

General Discussion

There were three primary aims of this set of experiments investigating the impact of generation constraint on item and context memory. First, we examined the generation effect for item, source and color memory. We consistently found a standard generation effect for item and source context memory measures (but not color context, in line with the processing account) in all experiments. Second, we examined whether the effect of generation constraint on item memory performance could be attributed to two mechanisms, item-specific and relational

processing (two-factor theory; Hirshman & Bjork, 1988). Our results indicate that lower-constraint generation is associated with enhanced relational processing, but similar item-specific processing, compared to higher-constraint generation. These results provide a potential explanation for why generation tasks with fewer constraints often lead to larger generation effects. Third, we investigated the impact of generation constraint on context memory. We examined the memory benefits for lower- and higher-constraint generation for both a conceptual (source) and perceptual (font color) context detail, and found that lower-constraint generation improved source memory compared to higher-constraint generation, but this benefit did not reliably extend to font color. Together, our context memory findings support the processing account of the generation effect for context memory (Jacoby, 1983; Mulligan, 2004, 2011).

Overall, our item memory findings provide two important contributions to the generation effect literature. First, data in these experiments provide evidence that fewer generation constraints can increase the generation effect, which adds to a growing body of work suggesting that generation constraints influence the memory benefits from self-generation (Fiedler et al., 1992; Gardiner et al., 1985; McCurdy et al., 2017, 2019). Importantly, constraint has not garnered much attention in past work on the generation effect. One reason prior work has eschewed the study of lower-constraint tasks is due to the potential confound of item-selection effects – the idea that lower-constraint generation leads to the generation of items that are inherently different from the items being studied in the comparison conditions (higher-constraint generation, read). Our findings show, however, that even when item-selection effects are accounted for, generation constraint still influences memory performance. Thus, the findings from the current experiments strongly suggest generation constraint is a factor that should be closely considered in future work. Particularly, our findings suggest that future studies should

more carefully consider the type of generation task to use in an experiment, perhaps based on the type of criterial test being used. Second, and perhaps more importantly, data from these experiments implicate enhanced relational processing as a memory mechanism accounting for past and present findings that fewer constraints increase the magnitude of the generation effect. This finding is important because it suggests that the current theoretical understanding of the generation effect, the two-factor theory, may no longer be the best account for what is known about the generation effect. It is becoming increasingly clear that all generation tasks are not equal in the type of processing they induce which in turn leads to variations in memory enhancement due to self-generation (McDaniel et al., 1988). Our present findings further promote the idea that future theoretical reasoning should consider how the processing induced by the generation task interacts with other experimental factors (e.g., memory test, stimuli, subject abilities), in order to more fully account for the variation observed across generation effect investigations (Jenkins, 1979; McDaniel & Butler, 2011; Roediger, 2008).

Another goal of this study was to examine the context memory benefits of lower-constraint generation for two types of context details, source memory (conceptual detail) and font color memory (perceptual detail). For source memory, across all experiments our results revealed that a lower-constraint generation task provided robust and consistent source memory benefits compared to a higher-constraint task and read controls, bolstering prior findings on this detail (McCurdy et al., 2017, 2019). Prior work has supported the idea that the enhanced relational processing from a generation task (usually of the cue-target relationship) can extend to other details as well (Geghman & Multhaup, 2004; Greenwald & Johnson, 1989; Marsh et al., 2001). This enhanced relational processing is thought to lead to a binding of multiple details into a single memory representation. Our source memory findings support this claim, and provide

more evidence that fewer generation constraints induce greater relational processing, in turn resulting in better memory for details conceptually associated with the item.

For color context memory, although we found some evidence that lower-constraint generation improved memory for font color in Experiment 1 (subject-based results), generally we found equivalent memory performance for font color across the three encoding tasks (lower-constraint, higher-constraint, and read controls). This finding importantly identifies another potential boundary condition (along with item recognition tests) of lower-constraint generation. Identifying the limiting features of lower-constraint generation is an important step in advancing our understanding of the experimental conditions when generation provides the largest benefits to memory, and should continue to be a focus of future work on the generation effect. It is also worth noting that across all three experiments, memory performance for this color detail was substantially lower than for item and source context memory measures. In Experiments 1 and 2, there were a large number of "don't know" responses suggesting that these findings may have only considered judgments that participants had high confidence in. Therefore, it is possible that participants' conservative judgments limited our ability to see any potential effects. In Experiment 3, however, when participants were forced to make a color judgment we still found no reliable differences in font color memory between any of the encoding tasks. Thus, it seems unlikely that conservative judgments in Experiments 1 and 2 influenced our color memory conclusions. Perhaps the more likely explanation for such low font color memory performance is that font color is a detail that is more difficult to integrate into a memory representation (which could also rationalize the high number of "don't know" responses in Experiments 1 and 2 for this detail). In this set of experiments, participants were required to remember multiple details (target item, source, and color). Some research suggests that due to capacity limits of cognitive

resources, generating selectively improves memory for information based on task demands and goals (Begg et al., 1989; Begg, Vinski, Frankovich, & Holgate, 1991). Therefore, it could be that font color was deemed the least critical component compared to the item and source memory measures, and thus was less likely to be encoded. The present study was designed to compare the memory effects between two types of contextual details (conceptual and perceptual), however future work might consider examining the effects of generation constraint on perceptual details independently, especially given our suspicion that the conceptual detail attracted a greater proportion of the finite cognitive resources available at encoding.

As a whole, our context memory findings may be best explained by the *processing account* (Jacoby, 1983; Mulligan, 2004, 2011). This account suggests that it is the match in processing between encoding and retrieval that determines which context details are improved by generation (in accord with transfer-appropriate processing principles; Morris et al., 1977). At encoding, generation is thought to require conceptual processing, or thinking about the relationship between the cue (and other information) in order to generate a target word. In contrast, reading is thought to lead to perceptual processing, or thinking about the data-driven (visual) features of the target word (Jacoby, 1983). Similarly, contextual details of an episode can be broadly classified as either conceptual details and perceptual details. Thus, in line with processing account (and transfer-appropriate processing), generating a word should improve memory for conceptual, but not perceptual, context details because of the overlap in processing between generation and conceptual context memory details. We generally found support for this account, where generation tasks improved source memory (a conceptual detail), but not font color memory (a perceptual detail).

This set of experiments provides multiple elements of evidence showing that lower-constraint generation enhances relational processing. Thus, a final question perhaps worth considering in light of the present finding is: *Why* does lower-constraint generation lead to additional relational processing over higher-constraint generation? To speculate, we consider the nature of the operations required by a lower- versus higher-constraint task. In the lower-constraint task, participants are given a cue and are able to freely generate a target word, with less guidance or information about what should be generated. Perhaps this freedom (i.e., reduced constraint) leads to more relational processing because the relation between the cue and target is not given to the participant (as it often is in higher-constraint tasks), and they must form the relation between the cue and target themselves. This freedom may in turn lead participants to activate a greater range of conceptually related information before selecting a single answer. These related concepts could potentially serve as retrieval cues, facilitating later retrieval of the target item (see Carpenter, 2011; Pyc & Rawson, 2010, for a related idea in the testing effect). Another potential outcome of greater activation of related concepts is inflated false alarms for associated information (Huff & Bodner, 2013; Hunt et al., 2011) which we found some evidence for in our data, strengthening the case for this speculation. Future work directly testing the influence of generation constraint on false memories could help to advance our theoretical understanding of how fewer constraints increase the generation effect, while also examining a potentially harmful effect of lower-constraint generation on memory.

Finding ways to improve memory is an important endeavor (Leach, McCurdy, Trumbo, Matzen, & Leshikar, 2018; Leshikar, Dulas, et al., 2015; Leshikar et al., 2017), and the present findings contribute to that pursuit. Decades of research have shown that self-generated materials are often better remembered than materials that are simply read. Here however, we show that not

all generation tasks are equivalent in the memory improvements they induce. Specifically, we show that materials produced under fewer constraints magnify the size of the generation effect and importantly we tie this added memory benefit to enhanced relational processing. Overall, this work provides compelling evidence that level of constraint in generation tasks influence the magnitude of the generation effect for both item and source memory and should be taken into account in future work on the generation effect.

Acknowledgments

The authors would like to thank Dena Giannakopoulos for her assistance in data collection, and Thomas Griffin, PhD, for his comments that helped shape this study.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

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.

Bower, G. H. (1970). Organizational factors in memory. *Cognitive Psychology, 1*(1), 18-46.

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, NY: Oxford University Press.

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(6), 1547-1552.

Donaldson, W., & Bass, M. (1980). Relational information and memory for problem solutions. *Journal of Verbal Learning and Verbal Behavior, 19*(1), 26-35.

Duarte, A., Henson, R. N., & Graham, K. S. (2008). The effects of aging on the neural correlates of subjective and objective recollection. *Cerebral Cortex, 18*(9), 2169-2180.

Duarte, A., Henson, R. N., Knight, R. T., Emery, T., & Graham, K. S. (2010). Orbito-frontal cortex is necessary for temporal context memory. *Journal of Cognitive Neuroscience, 22*(8), 1819-1831.

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

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149-1160.

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.

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., Wilder, A., & Rusch, L. (1989). Anagram solving: Does effort have an effect? *Memory & Cognition, 17*(6), 755-758.

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.

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.

Greenwald, A. G., & Johnson, M. M. (1989). The generation effect extended: Memory enhancement for generation cues. *Memory & Cognition, 17*(6), 673-681.

Grosofsky, 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.

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.

Huff, M. J., & Bodner, G. E. (2013). When does memory monitoring succeed versus fail? Comparing item-specific and relational encoding in the DRM paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(4), 1246-1256.

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and Memory* (pp. 3-25). Oxford, NY: Oxford University Press.

Hunt, R. R. (2012). Distinctive processing: The co-action of similarity and difference in memory. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 56, pp. 1-46). San Diego, CA: Academic Press.

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., & 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.

Hunt, R. R., Smith, R. E., & Dunlap, K. R. (2011). How does distinctive processing reduce false recall? *Journal of Memory and Language, 65*(4), 378-389.

Hunt, R. R., & Worthen, J. B. (2006). *Distinctiveness and memory*. Oxford, NY: Oxford University Press.

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.

JASP Team. (2018). JASP (Version 0.9. 0.1) [Computer software]: Amsterdam.

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.

Kinoshita, S. (1989). Generation enhances semantic processing? The role of distinctiveness in the generation effect. *Memory & Cognition, 17*(5), 563-571.

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., & Duarte, A. (2012). Medial prefrontal cortex supports source memory accuracy for self-referenced items. *Social Neuroscience, 7*(2), 126-145.

Leshikar, E. D., & Duarte, A. (2014). Medial prefrontal cortex supports source memory for self-referenced materials in young and older adults. *Cognitive, Affective, & Behavioral Neuroscience, 14*(1), 236-252.

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., & Gutchess, 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.

Leshikar, E. D., Park, J. M., & Gutchess, A. H. (2015). Similarity to the Self Affects Memory for Impressions of Others in Younger and Older Adults. *The Journals of Gerontology: Series B, 70*(5), 737-742.

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.

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 constraints enhance the generation effect for source memory in younger, but not older adults. *Open Psychology, 1*(1), 168-184.

McCurdy, M. P., Viechtbauer, W., Frankenstein, A. N., Sklenar, A. M., & Leshikar, E. D. (Under Review). *Theories of the generation effect and the impact of generation constraint: A meta-analytic review*. Manuscript submitted for publication.

McDaniel, M. A., & Butler, A. C. (2011). A contextual framework for understanding when difficulties are desirable. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 175-198). New York, NY: Psychology Press.

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.

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. (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., Lozito, J. P., & Rosner, Z. A. (2006). Generation and context memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(4), 836-846.

Murnane, K., & Bayen, U. J. (1996). An evaluation of empirical measures of source identification. *Memory & Cognition, 24*(4), 417-428.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36*(3), 402-407.

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. (2014). 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.

Psychology Software Tools, I. (2012). E-Prime 2.0.8.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*(6002), 335-335.

R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Roediger, H. L. (2008). Relativity of remembering: Why the laws of memory vanished. *Annual Review of Psychology, 59*, 225-254.

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.

Taconnat, 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.

Underwood, B. J., & Schulz, R. W. (1960). *Meaningfulness and verbal learning*. Philadelphia, PA: Lipponcott.

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language, 50*(3), 289-335.

Table 1

Raw (uncorrected) Means and Standard Deviations (in parentheses) for Item, Source, and Color Memory Judgments for Each Encoding Condition and New Words by Experiment

Experiment 1

Item Recognition			
Task	**Old**	**New**	**DK**
Generate	.81 (.15)	.14 (.13)	.05 (.04)
Scramble	.86 (.11)	.10 (.10)	.04 (.05)
Read	.52 (.16)	.37 (.16)	.11 (.11)
New	.16 (.16)	.70 (.19)	.14 (.13)

Source Recognition				
Task	**Generate**	**Scramble**	**Read**	**DK**
Generate	.68 (.22)	.07 (.07)	.05 (.07)	.20 (.17)
Scramble	.10 (.09)	.60 (.17)	.12 (.07)	.18 (.12)
Read	.07 (.07)	.12 (.08)	.33 (.17)	.48 (.18)
New	.03 (.05)	.05 (.07)	.10 (.10)	.82 (.17)

Color Recognition			
Task	**Correct**	**Incorrect**	**DK**
Generate	.28 (.19)	.29 (.19)	.43 (.28)
Scramble	.35 (.17)	.26 (.18)	.39 (.25)
Read	.21 (.16)	.14 (.12)	.65 (.21)
New	--	.12 (.15)	.88 (.15)

Experiment 2

Cued Recall			
Task	**Correct**	**Incorrect**	**Blank**
Generate	.88 (.12)	.03 (.05)	.09 (.10)
Scramble	.82 (.17)	.01 (.02)	.17 (.15)
Read	.60 (.19)	.04 (.07)	.36 (.22)

Source Recognition				
Task	**Generate**	**Scramble**	**Read**	**Blank**
Generate	.87 (.11)	.02 (.03)	.01 (.03)	.10 (.10)
Scramble	.06 (.18)	.68 (.21)	.08 (.08)	.18 (.15)
Read	.07 (.08)	.11 (.08)	.42 (.20)	.40 (.23)

Color Recognition			
Task	**Correct**	**Incorrect**	**Blank**
Generate	.21 (.21)	.22 (.24)	.57 (.39)
Scramble	.22 (.20)	.22 (.19)	.56 (.38)
Read	.16 (.17)	.16 (.18)	.69 (.34)

Experiment 3

Item Recognition (Related Distractors)		
Task	**Hits**	**FA**
Generate	.96 (.06)	.05 (.09)
Scramble	.92 (.12)	.03 (.04)
Read	.60 (.19)	.04 (.07)

Source Recognition			
Task	**Generate**	**Scramble**	**Read**
Generate	.95 (.09)	.03 (.05)	.02 (.05)
Scramble	.05 (.08)	.87 (.11)	.08 (.07)
Read	.14 (.11)	.21 (.15)	.65 (.21)

Color Recognition		
Task	**Correct**	**Incorrect**
Generate	.44 (.12)	.56 (.12)
Scramble	.48 (.14)	.52 (.14)
Read	.45 (.12)	.55 (.12)

Note. "Generate" = lower-constraint generation task; "Scramble" = higher-constraint generation task; "DK" = don't know; "FA" = False alarm rate of related distractors endorsed as "old". Values reported in this table represent the uncorrected response rates for each memory measure. In the analyses, manuscript, and in Figures 2, 3, and 4 we use conditional measures of source recognition and color recognition. In Experiment 1, the row corresponding to "new" items reflects the response rates across all 24 new items not presented at encoding (i.e., responses to new words regardless of whether participants correctly judged them as new).

Table 2. *Logistic (Logit) Regression Summary Tables by Memory Type and Experiment.*

Experiment 1

Item Recognition Hits					Conditional Source					Conditional Color				
Predictor	**β (*SE*)**	**Wald $x^2(1)$**	***p***	**e^{β} OR**	**Predictor**	**β (*SE*)**	**Wald $x^2(1)$**	***p***	**e^{β} OR**	**Predictor**	**β (*SE*)**	**Wald $x^2(1)$**	***p***	**e^{β} OR**
Intercept	0.21 (.23)	0.8	.370	1.23	**Intercept**	0.82 (.29)	12.5	< .001	2.27	**Intercept**	-0.73 (.23)	10.3	.001	0.48
Generate	1.31 (.16)	63.0	< .001	3.69	**Generate**	1.14 (.20)	47.4	< .001	3.11	**Generate**	0.39 (.18)	4.8	.028	1.48
Scramble	1.68 (.15)	135.4	< .001	5.41	**Scramble**	0.47 (.15)	10.6	.001	1.61	**Scramble**	0.22 (.15)	2.0	.150	1.25
NMT	-0.12 (.22)	0.3	.570	0.88	**NMT**	-0.44 (.27)	4.1	.042	0.64	**NMT**	0.01 (.19)	0.0	.980	1.00

Experiment 2

Cued Recall					Conditional Source				
Predictor	**β (*SE*)**	**Wald $x^2(1)$**	***p***	**e^{β} OR**	**Predictor**	**β (*SE*)**	**Wald $x^2(1)$**	***p***	**e^{β} OR**
Intercept	-0.52 (.27)	3.7	.055	0.59	**Intercept**	1.69 (.48)	12.1	< .001	5.44
Generate	2.10 (.22)	89.0	< .001	8.20	**Generate**	1.94 (.26)	57.6	< .001	6.96
Scramble	1.11 (.13)	70.6	< .001	3.05	**Scramble**	0.98 (.16)	35.6	< .001	2.67
NMT	0.92 (.26)	12.5	< .001	2.52	**NMT**	-0.97 (.47)	4.2	.040	0.38

Experiment 3

Item Recognition Hits					Item Recognition False Alarms					Conditional Source				
Predictor	**β (*SE*)**	**Wald $x^2(1)$**	***p***	**e^{β} OR**	**Predictor**	**β (*SE*)**	**Wald $x^2(1)$**	***p***	**e^{β} OR**	**Predictor**	**β (*SE*)**	**Wald $x^2(1)$**	***p***	**e^{β} OR**
Intercept	1.41 (.55)	6.6	.01	4.09	**Intercept**	-4.19 (.54)	60.7	< .001	0.02	**Intercept**	1.62 (.57)	8.2	.004	5.05
Generate	2.28 (.31)	53.3	< .001	9.79	**Generate**	0.85 (.38)	5.1	.024	2.33	**Generate**	1.70 (.42)	16.6	< .001	5.46
Scramble	1.64 (.19)	71.8	< .001	5.14	**Scramble**	-0.09 (.40)	0.05	.82	0.91	**Scramble**	0.79 (.25)	9.8	.002	2.20
NMT	-0.78 (.54)	2.1	.15	0.46	**NMT**	0.74 (.46)	2.5	.11	2.10	**NMT**	0.10 (.54)	0.03	.86	1.10

Note. "Intercept" = Read control task (reference group); "Generate" = lower-constraint generation task; "Scramble" = higher-constraint generation task; "NMT" = norm matched target; *SE* = Standard Error; OR = odds ratio. Betas (β) represent log odds of correctly remembering an item, source, or font color (or log odds of false alarming), relative to the Intercept (read control). All Wald chi-square tests have *degrees freedom* = 1. Conditional color summary tables not reported in Experiment 2 and 3 because model did not fit the data better than a null model (with no predictors).

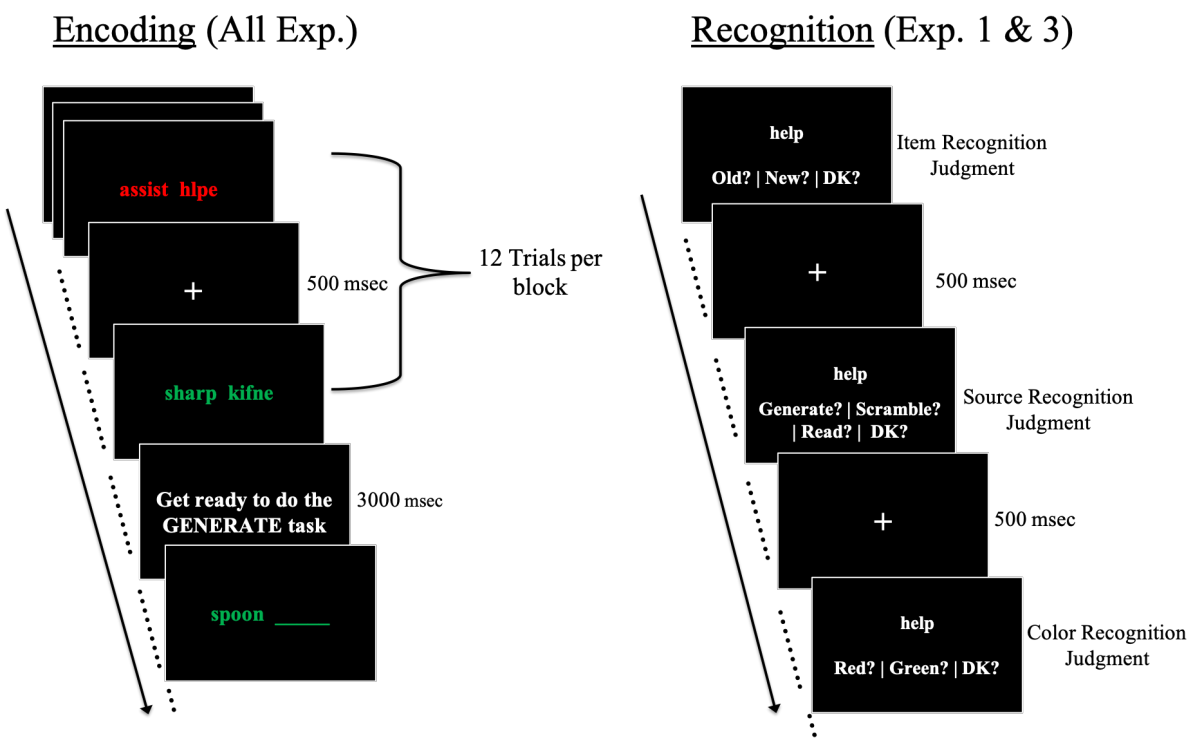

Figure 1. Trial schematic for encoding phase of all experiments, and recognition phase of Experiments 1 and 3. See **Supplemental Figure 1** for a sample depiction of the cued recall test in Experiment 2.

Note. Experiment 3 did not provide a "don't know" (DK) response option at recognition.

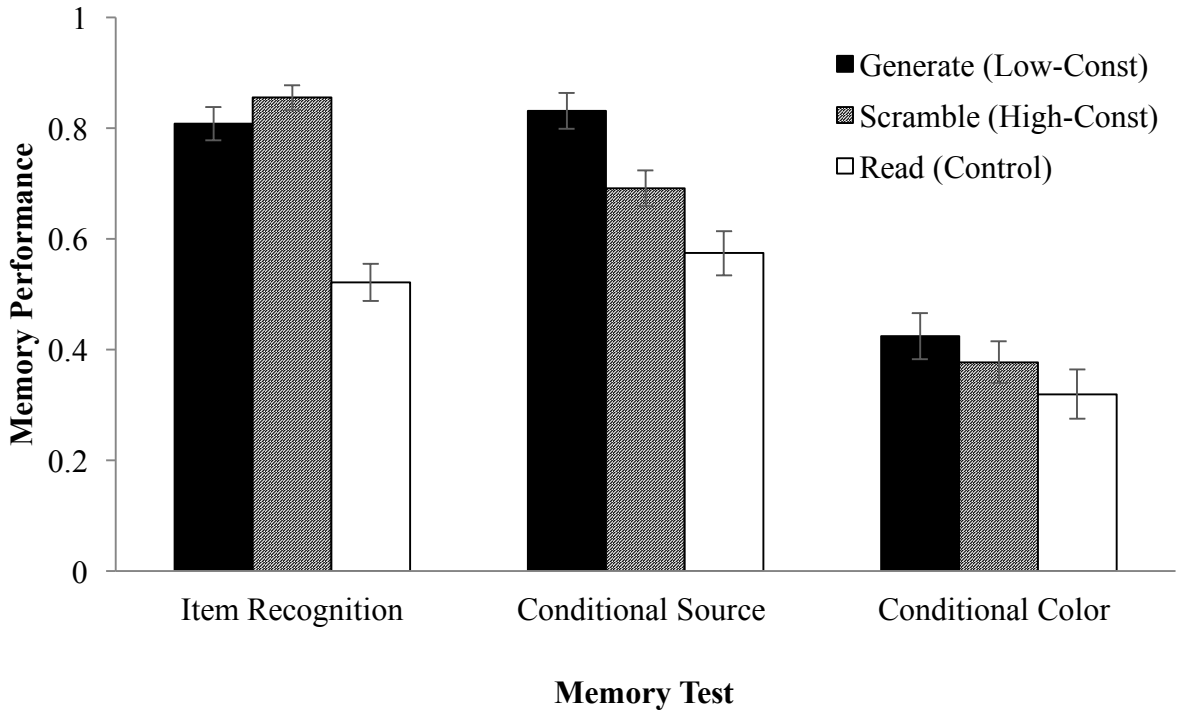

Figure 2. Item recognition (item memory), conditional source (context), and conditional color (context) memory performance by encoding task in **Experiment 1**. Source and color context recognition are conditional measures and reflect response rates of source and color recognition (respectively) only for items that were correctly recognized in the item recognition measure. "Generate" = lower-constraint generation task; "Scramble" = higher-constraint generation task. Error bars represent standard error.

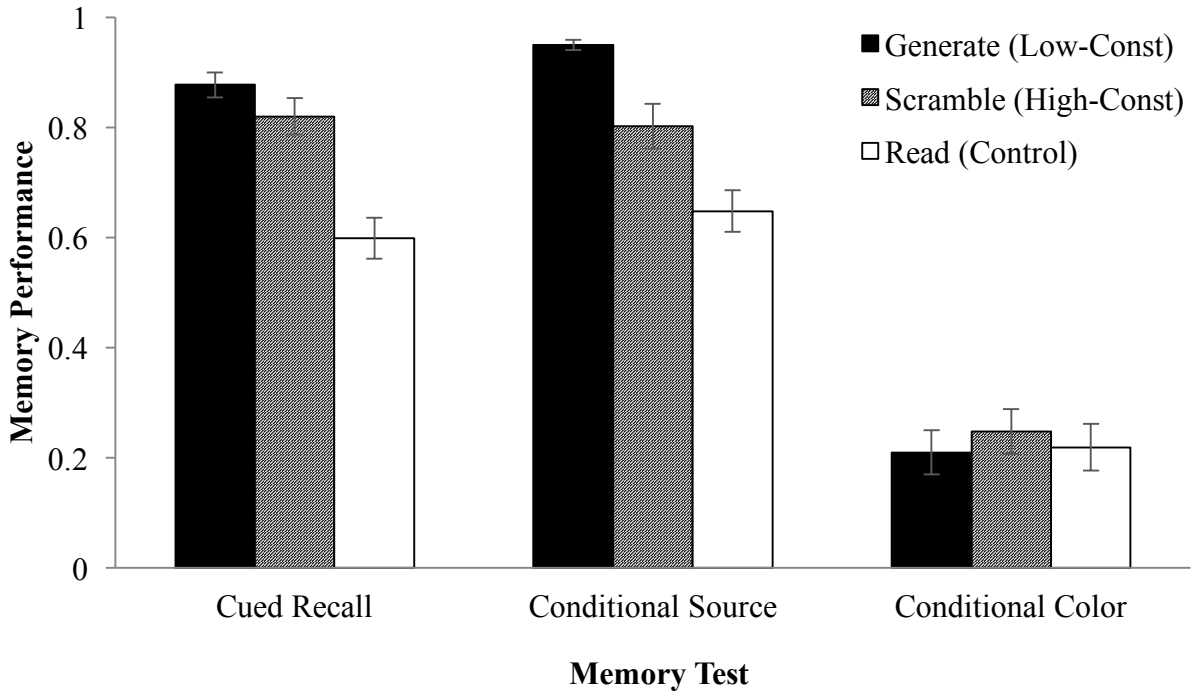

Figure 3. Cued recall (item memory), conditional source (context), and conditional color (context) memory performance by encoding task in **Experiment 2**. Source and color recognition are conditional measures and reflect percentage of correct response for source and color recognition (respectively) only out of items that were correctly recalled in the cued recall measure. "Generate" = lower-constraint generation task; "Scramble" = higher-constraint generation task. Error bars represent standard error.

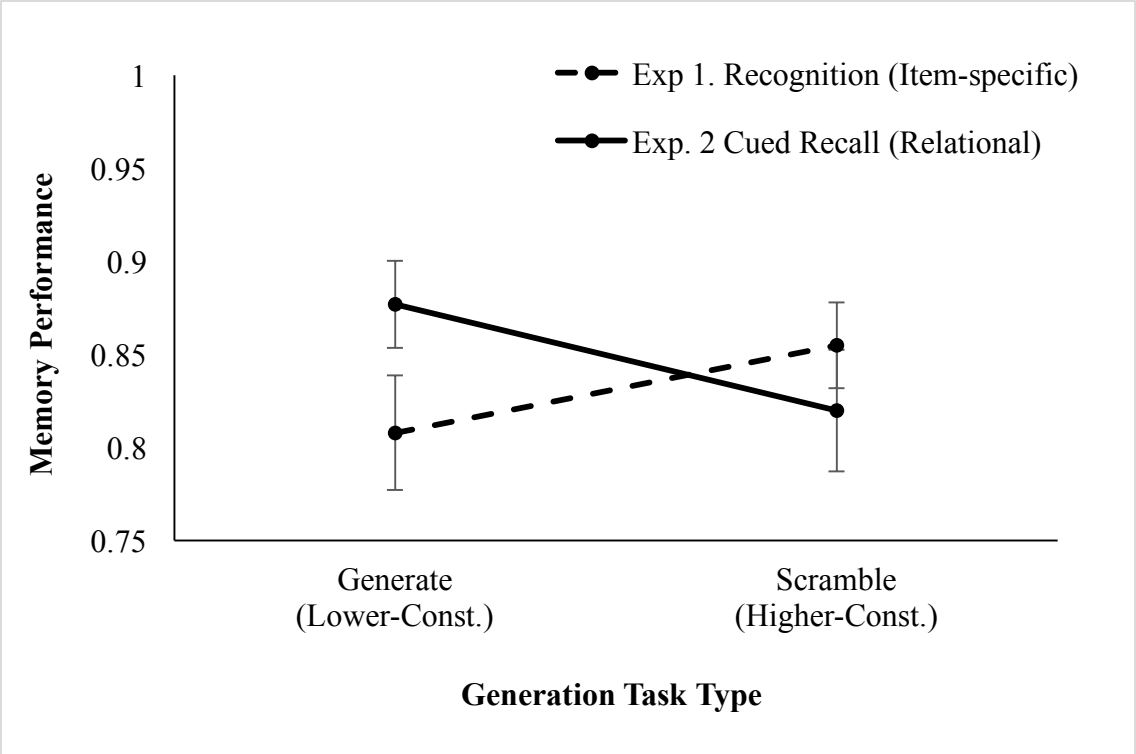

Figure 4. Cross experiment interaction plot. Item memory performance by generation task type and memory test. Error bars represent standard error.

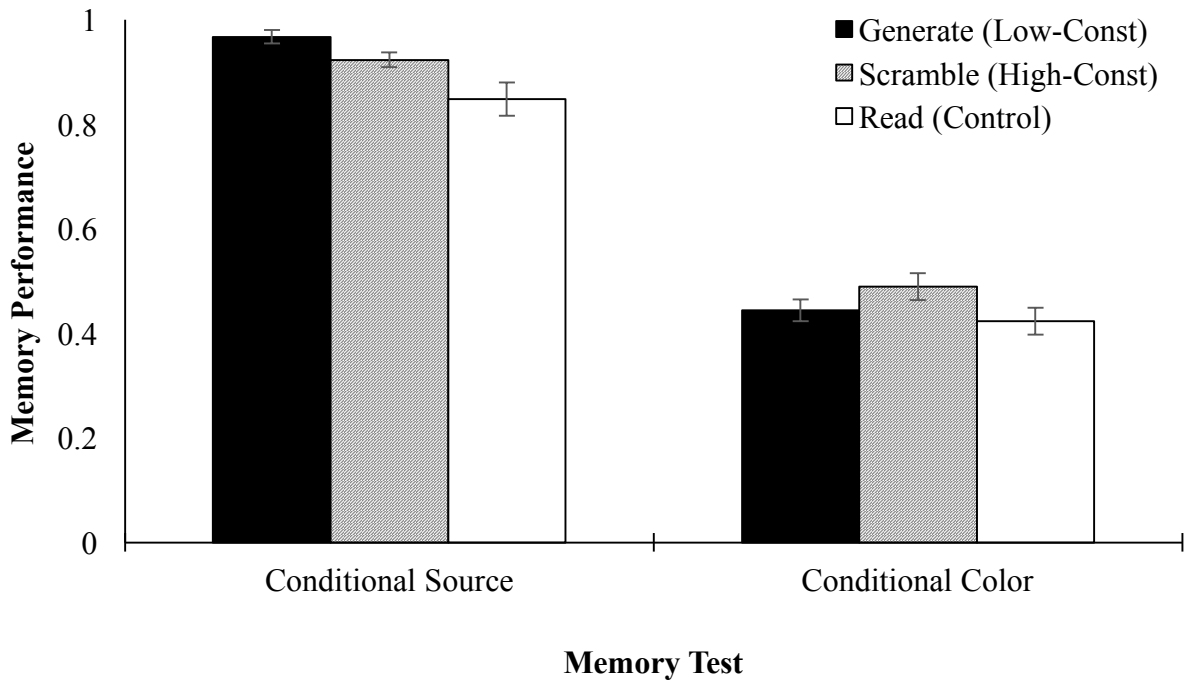

Figure 5. Conditional source (context), and conditional color (context) memory performance by encoding task in **Experiment 3**. Item hit rate and false alarm rate are reported in Table 1. "Generate" = lower-constraint generation task; "Scramble" = higher-constraint generation task. Error bars represent standard error.

For each of the following words, write the word that was paired with this word from the first part of the experiment. Then circle the condition the word appeared in, followed by the color the word appeared in.

best	_____	Generate | Scramble | Read	Red | Green
attire	_____	Generate | Scramble | Read	Red | Green
circle	_____	Generate | Scramble | Read	Red | Green
bacon	_____	Generate | Scramble | Read	Red | Green
hinge	_____	Generate | Scramble | Read	Red | Green
touch	_____	Generate | Scramble | Read	Red | Green
roar	_____	Generate | Scramble | Read	Red | Green
⋮			

Supplemental Figure 1. Example of the cued recall test format used in Experiment 2.