# Don't believe the font: Sans Forgetica is not desirable for learning

Jason Geller[1], Sara D. Davis[2], & Daniel Peterson[2]

[1] University of Iowa
[2] Skidmore College

Abstract

Do students learn better with material that is perceptually harder-to-process? While evidence is equivocal on the matter, recent claims suggest that placing materials in Sans Forgetica font, which is perceptually hard-to-process, has positive effects on student learning. Given the weak evidence for perceptual disfluency effects, this led us to examine the mnnmonic effects of Sans Forgetica more closely. In three preregistered experiments, we tested if Sans Forgetica is really unforgetable. In Experiment 1 ($N$ = 233), participants studied weakly realted cue-target pairs with targets presented in either Sans Forgetcia or with missing letters (e.g., G_RL). Cued recall performance showed a robust generation effect, but no Sans Forgetica memory benefit. In Experiment 2 ($N$=528), participants read a passage about ground water with select sentences presented in either Sans Forgetcia, yellow highlighting, or unmodified. Cued recall for select words were better for pre-highlighted information than when unmodified. Critically, presenting sentences in Sans Forgetica did not produce better cued recall than pre-highlighted sentences or sentences presented unchanged. In Experiment 3 ($N$ = 60), indiviudals did not have better discriminabiliy for Sans Forgetic in an old-new recognition test. Our findings suggest that Sans Forgetica is really forgeticable.

*Keywords:* Disfluency, Recall, Desirable Difficulty, Learning and Memory
Word count: 4458

Students want to remember more and forget less. Being able to recall and apply previously learned information is key for successful learning. Decades of research in the laboratory

Jason Geller, Department of Psychology and Brain Sciences, University of Iowa, W113 Seashore Hall, Iowa City, IA, 52242;

Correspondence concerning this article should be addressed to Jason Geller, Department of Psychological and Brain Science, W113 Seashore Hall, Iowa City, IA, 52242. E-mail: jason-geller@uiowa.edu

and in the classroom have put forth the paradoxical idea that making learning harder (not easier) should have the desirable effect of improving long-term retention of material–called the desirable difficulty principle (Bjork & Bjork, 2011). Notable examples of desirable difficulties include having participants generate information from word fragments instead of passively reading intact words (Bertsch, Pesta, Wiscott, & McDaniel, 2007), spacing out study sessions instead of massing them (Carpenter, 2016), and having participants engage in retrieval practice after studying instead of simply restudying the information (Kornell & Vaughn, 2016). Another simple strategy that has gained some attention is to make material more perceptually disfluent. This can be done by changing the material's perceptual characteristics. Visual material that is masked (Mulligan, 1996), inverted (Sungkhasettee, Friedman, & Castel, 2011), presented in an atypical font (Diemand-Yauman, Oppenheimer, & Vaughan, 2011; French et al., 2013), blurred (Rosner, Davis, & Milliken, 2015), or even in handwritten cursive (Geller, Still, Dark, & Carpenter, 2018) have all been shown to produce memory benefits. The desirable effect of perceptual disfluency on memory is called the disfluency effect (Bjork & Yue, 2016).

Although appealing as a pedagogical strategy due to the relative ease of implementation, there have been several experiments that failed to find memorial benefits for perceptually disfluent materials (e.g., Magreehan, Serra, Schwartz, & Narciss, 2016; Rhodes & Castel, 2008, @Rhodes2009; Rummer, Schweppe, & Schwede, 2016; Yue, Castel, & Bjork, 2013), casting doubt upon the robustness of the disfluency effect. Corroborating this, A recent meta-analysis by Xie, Zhou, and Liu (2018) with 25 studies and 3,135 participants found a small, non-significant, effect of perceptual disfluency on recall and ($d$ = -0.01) and transfer ($d$ = 0.03). Despite having no mnemonic effect, perceptual disfluency produced longer reading times ($d$ = 0.52) and lower judgments of learning ($d$ = -0.043). In the laboratory, Geller et al. (2018) and Geller & Still (2018) manipulated several boundary conditions (e.g., level of degradation, type of judgement of learning, retention interval, and testing expectancy) and found you can get positive memory effects from perceptual disflunet materials (in recognition), but it is not robust. Taken together, the evidence is weak for perceptual disfluency being a desirable difficulty.

Despite the weak evidence, perceptual disfluency is still being touted as a viable learning tool, especially in the popular press. Recently, reputable news sources like the Washington Post (https://www.washingtonpost.com/business/2018/10/05/introducing-sans-forgetica-font-designed-boost-your-memory/) and National Public Radio (NPR; https://www.npr.org/2018/10/06/655121384/sans-forgetica-a-font-to-remember) claimed that a new font called Sans Forgetica could enhance memory, despite only unpublished evidence being available at the time (Earp, 2018). It is thought that the mnemonic benefit is due to the characteristics of the font. Sans Forgteica is a variation of a sans-serif typeface that consists of intermittent gaps in letters that are back slanted (see fig. 1). Since the release of those news articles, the Sans Forgetica font is available on all operating systems (download the font file), some browsers (e.g., Chrome), and can be downloaded on your phone.

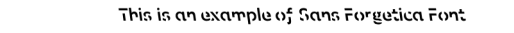


*Figure 1*. Example of Sans Forgetica font.

## Current Studies

Given the weak evidence for the disfluency effect, we thought it pertinent to empirically examine whether Sans Forgetica produces more durable learning. The question of whether Sans Forgetica produces a mnnmenomic benefits has clear practical implications. In the educational domain, it would be relatively quick and easy to use place materials in Sans Forgetica font. However, in order for the Sans Forgetica to be useful, it is important to note and understand both its successes and failures. To the authors' knowledge, there is only one peer-reviewed paper (Eskenazi & Nix, 2020) examining the effectiveness of Sans Forgetica in generating a desirable difficulty. In one experiment Eskenazi and Nix (2020) found that words and definitions in Sans Forgetica font lead to better orthographic discriminabity (i.e., choosing the correct spelling of the word) and semantic acquisition (i.e., retrieving the definition of a word), but only if participants were good spellers. As the Eskenazi and Nix (2020) study focused on lexical acquisition (orthographic and semantic features of a word), it is not clear if the benefits of Sans Forgetica font extends to other memory processes. Given this, we felt it was pertinent to examine the effectiveness of Sans Forgetica in two different memory experiments. To this end, we conducted to two high-powered preregistered experiments examining whether (1) recall is better in Sans Forgetica font and (2) how it compares with other notable learning techniques–generation (Experiment 1) and pre-highlighting (Experiment 2). Comparing Sans Forgetica to other study techniques allows us to examine the mechanisms underlying the effect, if any.

## Experiment 1

In Experiment 1 we were interested in answering two questions. First, is Sans Forgetica more memorable than a normal, fluent, font (e.g., Arial)? Second, is the Sans Forgetica effect on memory similar in magnitude to the generation effect? While very is known about Sans Forgetica, one of the most intuitively appealing theories for why Sans Forgetica font benefits memory is that of mental effort. It is believed that reading materials in Sans Forgetica requires more effort than simply reading a normal font. Essentially, the intermittent gaps of Sans Forgetica requires readers to generate or fill in the missing pieces producing a memory advantage. This mechanism of action is similar to that of the generation effect, wherein information is better remembered when generated or filled-in compared to if it is simply read. In Experiment 1 we examined the mnemonic benefit of Sans Forgetica and generation by looking at cued recall performance with weakly related pairs. If Sans Forgetica does produce a mnemonic benefit, we should observe better cued recall performance for targets in Sans forgetica font compared to Arial font. Futhrer, if it is similar to the generation effect, the magnitude of the memory benefit between the two should be similar.

## Method

**Participants.** Two-hundred and thirty people from Amazon's Mechanical Turk Service participated for money. Sample size was based on a priori power analyses conducted using PANGEA v0.2 (Westfall, 2015). Sample size was calculated based on the smallest effect of interest (SEOI; Lakens & Evers, 2014). In this case, we were interested in powering our study to detect a medium effect size interaction effect ($d = .35$). We choose this effect size as our SESOI due in part to the small effect sizes seen in actual classroom studies (Butler, Marsh, Slavinsky, & Baraniuk, 2014). Therefore, assuming an alpha of .05 and a desired power of 90%, a sample size of 230 is required to detect whether an effect size of .35 differs from zero. After excluding participants who 1) did not complete every phase of the experiment, 2) started the experiment multiple times, 3) reported experiencing technical problems did not indicate that they were fluent in English [^2]: This question was not asked during the experiment., or 5) reported seeing our stimuli before, we were left with 115 participants per group.

**Materials.** The preregistration for Experiment 1 can be found here: https://aspredicted.org/3ai98.pdf. All materials, data, and analysis scripts for both Experiment 1 can be found here (https://osf.io/d2vy8/). The results contained herein are computationally reproducible by going to the primary author's github repository for the paper (https://github.com/jgeller112/SF_Expt2) and clicking on the binder button.

Participants were presented with 22 weakly related cue-target pairs taken from Carpenter, Pashler, and Vul (2006))[^1]: Two cue-target pairs (e.g., range-rifle and train-plane) had to be thrown out as they were not presented due to a coding error. The cue-target pairs were all nouns, 5–7 letters and 1–3 syllables in length, and high in concreteness (400–700) and frequency (at least 30 per million). Free association norms (Nelson, McEvoy, & Schreiber, 2004) were used to create 22 weakly associated pairs of similar forward and backward strength. Two counterbalanced lists were created for each difficulty type group(generation and Sans Forgetica) so that each item could be presented in each disfluency conditions without repeating any items for an individual participant.

**Design and Procedure.** Disfluency (fluent vs. disfluent) was manipulated within-subjects and within-items and difficulty type (Generation vs. Sans Forgetcia) was manipulated between participants. For half the participants, targets were presented in Sans Forgetica while the other half were presented in Arial font; for the other half of participants, targets were presented with missing letters (vowels were replaced by underscores) and the other half were intact (Arial font). After a short 2 minute distraction task (anagram generation), they completed a cued recall test. During cued recall, participants were presented 24 cues one at a time and asked to provide the target word. After they were thanked and debriefed.

Participants completed the experiment on-line via the Qualtrics survey platform hosted on Amazon Mechanical Turk. After reading and consenting, participants were randomly assigned to one of two conditions: The generation condition or the Sans Forgetica condition. Participants were told to study word pairs so that later they could recall second word (target) when cued with the first word (cue). The experiment began with the presentation of 22 word pairs, shown one at a time, for 2 seconds each. The cue word always

appeared on the left and the target always on the right. Immediately proceeding this, participants did a short 2 minute distraction task (anagram generation). Finally participants completed a cued recall test. During cued recall, participants were presented 22 cues one at a time and asked to provide the target word. Responses were self-paced. Once completed, participants clicked on a button to advance to the next question. At the end, participants were asked several demographic questions.

**Scoring.** Spell checking was automated with the hunspell package in R (Ooms, 2018) using spellCheck.R. Because participants were recruited in the United States, we used the American English dictionary. A nice walk through on how to use this package can be found in Buchanan, De Deyne, and Montefinese (2019). Using this package, each response was corrected for misspellings. Corrected spellings are provided in the most probable order, therefore, the first suggestion was always selected as the correct answer. As a second pass, we manually examined the output to catch incorrect suggestions. If the response was close to the correct response, it was marked as correct.

**Analysis.** Although we pre-registered a traditional ANOVA approach, in Experiment 1, 2, and 3we opted for a more powerful mixed modeling approach that better represents the structure of the data (Hoffman & Rovine, 2007; Locker, Huffman, & Bovaird, 2007). All data were anayzed in R (vers. 3.5.0; R Core Team, 2019), with models fit using the lme4 package (vers. 2.3.1; Bates, Mächler, Bolker, & Walker, 2015) and the brms package [vers. 2.11.0; Bürkner (2018)). All figures were created with ggplot (Wickham, 2016). In all three experiments, logistic mxied-effects models were used to model binary outcomes. All models were analyzed using maximal random effects structures with random slopes where allowed (**???**). All figures were created with ggplot (Wickham, 2016). In cases were null effects arose, a Bayes Factor was dervided by fitting the maximal model and using the hypothesis function in brms (Bürkner, 2018).

## Results and Discussion

We fit a generalzied linear mixed model (logit link) to predict cued recall accuracy with difficulty type (generation vs. sans forgetcia) and disfluency (fluent vs. disfluency) as categorical predictors. Each categorcal preictor was devaition coded to assess each main effect and inteaction independently of all the other predictors in the model. This is the final model used: formula: glmer(acc~difftypedisflu + (1+disflu|ResponseID) + (1+difftype|target), family=binomial, data=data). Each categorcal preictor was devaition coded (0.5, -0.5) to assess each main effect and inteaction independently of all the other predictors in the model. Effect sizes (Cohen's $d$) were labelled following Chinn (2000)'s recommendations. There was no difference in cued recall between Generation and Sans Forgetica groups, $\beta$ = -0.09, $SE$ = 0.11, 95% CI [-0.30, 0.13], $p = 0.431$, $d = 0.05$). Individuals recalled more disfluent target words than fluney target words, $\beta = 0.21$, $SE = 0.06$, 95% CI [0.09, 0.33], p < .001, $d$ = 0.12). This was qualified by an interaction between difficulty type and disfluency, $\beta = 0.22$, $SE = 0.04$, 95% CI [0.14, 0.30], $p < .001$, $d = 0.11$). As seen in Fig. 2, the magnitude of the generation effect was larger than the Sans Forgetica effect.

```
## Warning: Missing column names filled in: 'X1' [1]
```

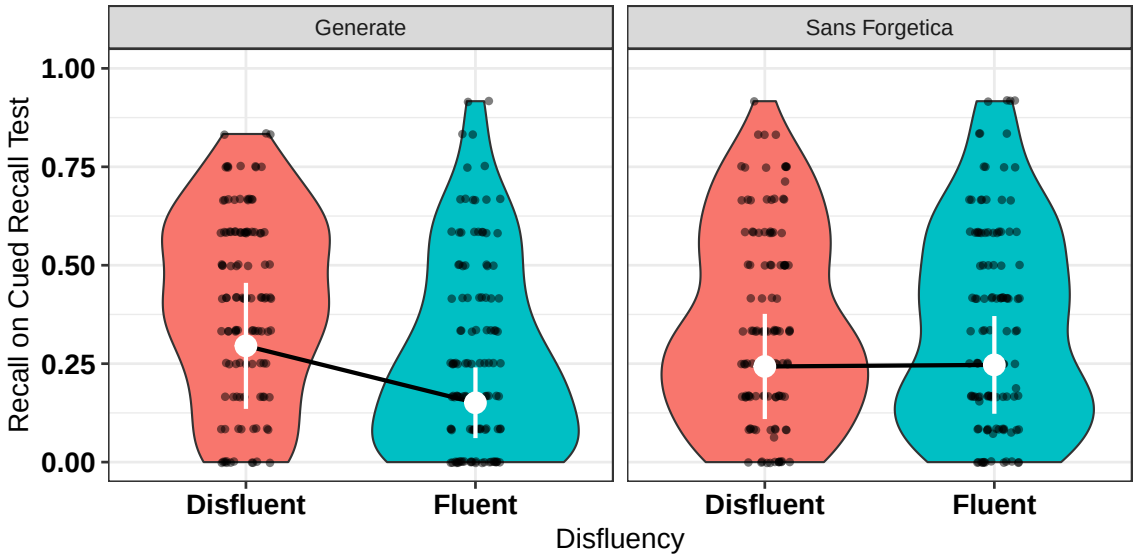


*Figure 2*. Accuracy on cued recall test. Violin plots represent the kernal density of avearge accuracy(black dots) with the fixed effect mean (white dot) and 95% CIs derived from the GLMER model.

The results for Experiment 1 are clear-cut. Cued recall for items presented intact and in Sans Forgetica font were equivocal. That is, we did not observe a memory benefit for Sans Forgetica font. We did, however, observe greater recall for generated items, which replicates decades of literature (Bertsch et al., 2007). This suggests that (1) presenting materials in Sans Forgetica does not lead to better memory and (2) the Sans Forgetica effect is most likely not a desirable difficulty.

## Experiment 2

Experiment 1 failed to find a memory benefit for Sans Forgetica effect. A limitation of Experiment 1 is that simple stimulus-response learning lacks educational realism. To remedy this, Experiment 2 tested the mnemonic effects of Sans Forgetica using more realistic materials. Whereas Experiment 1 tested whether Sans Forgetica is driven by the generative process of retrieval, Experiment 2 examined whether the Sans Forgetcia effect might exert its mnemonic benefit by making material more distinctive. Specifically, Sans Forgetica may make the marked portion of text more memorable because it stands out from the surrounding text. This is similar to the effects of pre-highlighting on learning. Indeed, some evidence supports this type of role for highlighting: When students read pre-highlighted passages, they recall more of the highlighted information and less of the non-highlighted information compared to students who receive an unmarked copy of the same passage (Fowler & Barker, 1974; Silvers & Kreiner, 1997). To this end, Experiment 2 compared cued recall performance on a passage where some of the sentences were either presented in: Sans Forgetica, pre-highlighted in yellow, or unmodified. We hypothesized that if the Sans Forgetica effect is mainly driven by distinctiveness, words presented in Sans Forgetica

should benefit more from the disfluency than the passage presented unmodified. Further, the benefit for Sans Forgetica should be similar in magnitude to the pre-highlighting condition as both manipulations serve to increase the distinctiveness of the text.

## Method

The pre-registration form for Experiment 2, which includes hypotheses, planned analyses,exclusion criteria, and sample size justification, can be found at: https://aspredicted.org/3jz3z.pdf.

**Participants.** Five hundred and twenty-eight undergraduates ($N = 528$) participated for partial completion of course credit. Sample size was based on a priori power analyses conducted using PANGEA v0.2. Sample size was calculated based on the smallest effect of interest (Lakens & Evers, 2014). Similar to Experiment 1, we were interested in powering our study to detect a medium-sized effect size ($d = .35$). Therefore, assuming an alpha of .05 and a desired power of 90%, a sample size of 170 per group is required to detect whether an effect size of .35 differs from zero. After excluding participants based on our preregistered exclusion criteria, we were left with unequal group sizes. Because of this, we ran six more participants per group, giving us 176 participants in each of the three conditions.

**Materials.** All materials used for this experiment can be found on our OSF page (https://osf.io/d2vy8/) under our Expt 2 Stims folder. Participants read a passage on ground water (856 words) taken from from the U.S. Geological Survey (see Yue, Storm, Kornell, & Bjork, 2014). Eleven critical phrases[1] each containing a different keyword, were selected from the passage (e.g., the term *recharge* was the keyword in the phrase: Water seeping down from the land surface adds to the ground water and is called recharge water.) and were either presented in SF, highlighted, or unmodified. Then, 11 fill-in-the blank questions were created from these phrases by deleting the keyword and asking participants to provide it on the final test (e.g., Water seeping down from the land surface adds to the ground water and is called _____ water). There was 1 manipulation check question: "What was the passage you read on?."

**Design and Procedure.** Participants were randomly assigned to either the pre-highlighted condition, Sans Forgetica condition, or unmodified condition. Our design manipulated three difference types of passages between-subjects: pre-highlighting, Sans Forgetica, and unmodified.

Participants completed the experiment on-line via the Qualtrics survey platform. After reading and signing a consent form, participants were randomly assigned to one of three conditions: pre-highlighting, Sans Forgetica, or unmodified. Participants read a passage on ground water. All participants were instructed to read the passage as though they were studying material for a class. After 10 minutes, all participants were given a brief questionnaire (2 questions) asking them to indicate their metacognitive beliefs afte reading

[1] originally we had 12 critical phrases but a pilot test showed that one of the questions was repeated twice so we removed one of them and also added a manipulation check question to sure participants were paying attention

the passage. The two questions were: "Do you feel that the presentation fo the material helped you remember" and "How likely is it that you will be able to recall material from the passage you just read on a scale of 0 (not likely to recall) to 100 (likely to recall) in 5 minutes?" Participants were then given a short distraction task (anagrams) for 3 minutes. Finally, all participants were given 12 fill-in-the-blank test questions, presented one at a time.

**Scoring.** Spell checking was automated with the same procedure as Experiment 1.

## Results and Discussion

For congruence with Experiment 1, we fit a logistic mixed model in a similar fashion. We fit a model with the fixed effect of passage type and random intercepts for participants ($N$=528) and questions ($N$=11): (formula: acc=glmer(auto_acc~passage_type+(1|Participant) + (1|Question), data=data, family="binomial"). Passage type was coded using treatment coding. We hypothesized that recall for pre-highlighted and Sans Forgetica sentences would be better remembered than normal sentences and that there would be no recall differences between the highlighted and sans forgetia sentences. Our hypotheses were partially supported (see Fig. 2). Results indicated that pre-highlighted sentences were better remembered than sentences presented normally, $\beta = 0.38$, $SE = 0.17$, 95% CI [0.05, 0.71], $p < .05$, $d = 0.21$, and were marginally better remembered than sentences presented in Sans Forgetcia, $\beta$ = -.317, *exp(B)* = 1.37, $SE$ = .168, $z$ = -1.89, $p$ = .059, $d$ = -0.18. Critically, there was no difference between sentences presented normally and in Sans Forgetcia, $\beta = 0.06$, $SE = 0.17$, 95% CI [-0.26, 0.39], $p = 0.700$, $d = .03$. A Bayes factor indicated strong evidence for no effect between the two conditions (BF = 7.47).

## Exploratory Analysis

In Experiment 2 we also asked students about their metacognitive awareness of the manipulatons. Specifically we asked participants: "How likely is it that you will be able to recall material from the passage you just read on a scale of 0 (not likely to recall) to 100 (likely to recall) in 5 minutes?" Initial analyses suggest that the normal passages were given higher JOLs ($M$ = 57.4, $SE$ = 1.97) than the pre-highlighted passage ($M$ = 50.3, $SE$ = 1.97), t(525) = -7.08, $p$ = .023. There were no reliable differences between the pre-highlighted passage and Sans Forgetica ($M$ = 53.8, $SE$ = 1.97), $t(525)$ = -3.52, $p$ = .415 or between the passage in Sans Forgetica and the passage presented normally, $t(525)$ = 3.56, $p$ = .406.

| contrast | estimate | SE | df | t.ratio | p.value |
|---|---|---|---|---|---|
| Pre-highlighted - Unmodified | -7.079546 | 2.7792 | 525 | -2.547332 | 0.0299152 |
| Pre-highlighted - Sans Forgetica | -3.517046 | 2.7792 | 525 | -1.265488 | 0.4153929 |
| Unmodified - Sans Forgetica | 3.562500 | 2.7792 | 525 | 1.281844 | 0.4060534 |

Words presented in Sans Forgetica did not lead to better recall than words left unmodified or pre-highlighted. We did, however, observe better memory for pre-highlighted information compared to words presented unmodified or in a Sans Forgetica font.

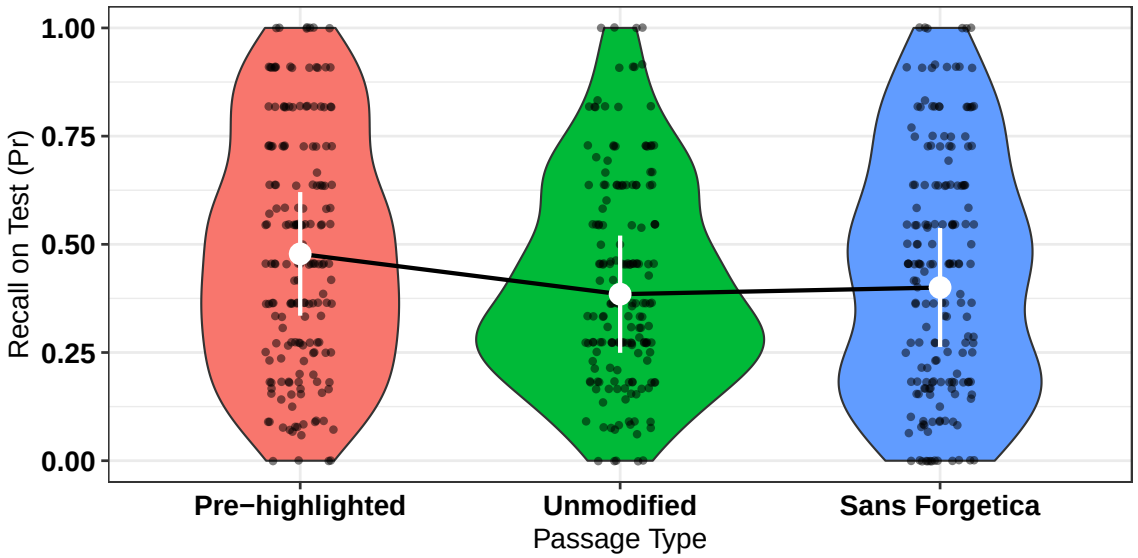


*Figure 3*. Passage accuracy as a function of passage type. Violin plots represent the kernal density of avearge accuracy(black dots) with the fixed effect mean (white dot) and 95% CIs derived from the GLMER model.

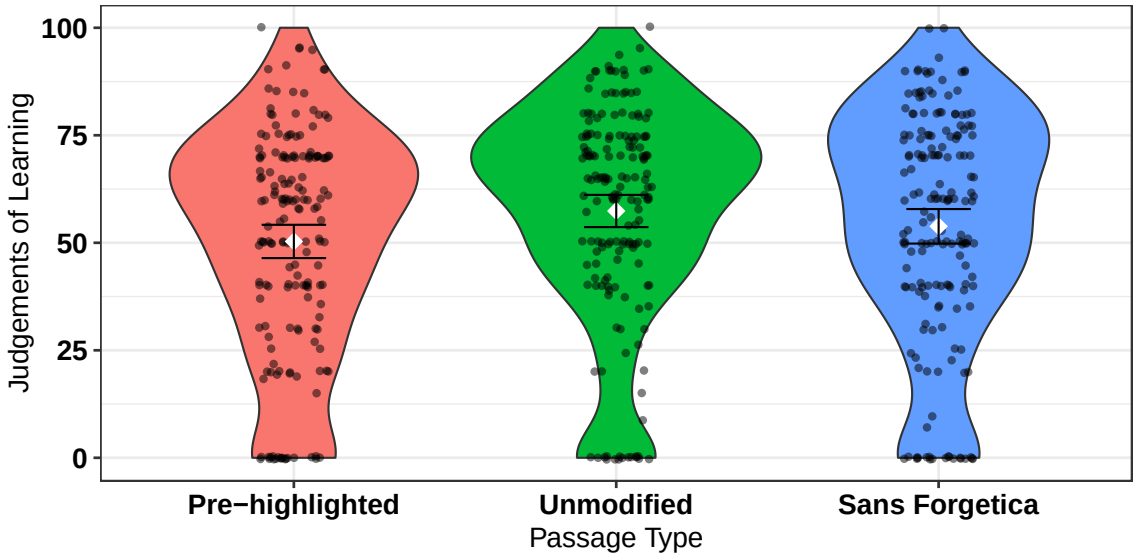


*Figure 4*. Judgements of learning as a function of passage type.

Examining metamemory judgments, we showed that a passage in Sans Forgetica font does not produce lower judgement of learning compared to unmodified or pre-highlighted passages. Interestingly, individuals gave lower JOLs to pre-highlighted information compared to materials presented in a normal font. One potential reason for pre-highlighted information receiving lower JOLs than the normal passage is that pre-highlighted information served to focus participants attention specific parts of the passage. Given the question, participants might have thought this would hinder them if tested over the passage as a whole.

## Experiment 3

The pre-registration form for Experiment 3, which includes hypotheses, planned analyses, exclusion criteria, and sample size justification, can be found at: https://osf.io/ekqh5.

**Participants.** Sixty participants ($N = 60$) participated for partial completion of course credit. Saample size was determined by a similar procedure to the above experiments. No participnats had to be thrown out for failing to meet the exclusion criteraia noted above.

**Materials.** The full set of stimuli can be found at https://osf.io/dsxrc/

Stimuli were 188 nouns taken from Geller et al. (2018). ALl words were from the English Lexicon Project database (Balota et al., 2007). Both frequency (all words were high frequency; mean log HAL frequency = 9.2) and length (all words were four letters in length) were controlled.

### Design and Procedure

The experiment employed a within-subject design. The factor of script type (Arial vs. Sans Forgetica) was manipulated within-subjects. We employed 188 words, 94 at study (47 in each script condition) and 188 at test (94 old and 94 new). This resulted in four counterbalanced lists. Lists were assigned to participants so that across participants each word occurred equally often in the four possible conditions: Arial-old, Arial-new, Sans Forgetica-old, Sans Forgetica-new.

Word order was completely randomized, such that Arial and Sans Forgetica words were randomly intermixed in the study phase, and Arial and Sans Forgetica old and new words were randomly intermixed in the test phase. All old words were presented at test in the same manner in which they were presented at study; that is, Arial words during study were presented in Arial font at test, and Sans Forgetica words during study were presented in Sans Forgetica font at test.

The experiment was created and conducted using the Gorilla Experiment Builder ((Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020); http://www.gorilla.sc). The experiment protocol and all tasks are available to preview and copy from Gorilla Open Materials at http://www.gorilla.sc/openmaterials/36778.

After reading and signing a consent form, participants first completed a study phase. During the study phase, a fixation cross appeared at the center of the screen for 500 ms.

The fixation cross was immediately replaced by a word in the same location. To continue to the next trial, participants pressed the continue button at the bottom of the screen. Each trial was self-paced. After the study phase, a short 3-minute distractor task was administered in which participants wrote down as many United States capitals as they could. Afterward, participants took an old-new recognition test. At test, a word appeared in the center of the screen that either had been presented during study ("old") or had not been presented during study ("new"). Old words occurred in their original script, and following the counterbalancing procedure, each new word was presented in Arial font or Sans Forgetica font. For each word presented, participants chosee from one of two boxes displayed on the screen: a box labeled "old" to indicate that they had named the word during study, and a box labeled "new" to indicate they did not remember naming the word. Words stayed on the screen until participants gave an "old" or "new" response. All words were individually randomized for each participant during both the study and test phases. After the experiment, participants were debriefed. The entire experiment took about 30 minutes to complete.

## Results and dicussion

In recognition memory, signal detection theory has proven to be a very informative and efficient approach to analyzing binary accuracy data. However, considering the deficiency in precision and power in traditional analyses compared to mixed effects analyses, it is worth considering a generalized linear mixed effect approach to signal detection theory (**???**). In its simplest from, SDT models are probit regressions. To estimate the SDT paramater of interest (d'), we fit a logiftical mixed model (with a probit link) to participant responses (their actual response (sayold; whether they responded with old vs. new)) with fixed effects for actual status of the item (isold; whether the item was old vs. new) and condition (Arial vs. Sans Forgetica) and the interaction between the two with random intercepts for participants ($N$=60) and targets ($N$=188): oldnew=glmer(sayold~isold*condition+(1|Participant)+ (1|target)). The variables isold and condition were contrast coded (0.5, -0.5) to allow for the estimation of the interaction between isold and condition. Within this mode, the fixed effect of condition is the difference in c between groups, and the interaction term isold:condition would describe the difference in d' between conditions.

Hit rates and false alarm rates can be seen in Fig. 3. The results are staright-forward. Indivausls were more biased to say Sans Forgetica stimuli were old, $\beta = 0.26$, $SE = 0..026$, 95% CI [0.05, 0.71], $p < .005$, $d = 0.21$. However, there was no difference in d-prime between the two conditions, $\beta = 0.033$, $SE = 0.05$, 95% CI [0.05, 0.71], $p = .519$, $d = 0.21$. The Bayes factor indicated strong evidence that the effect was zero, BF = 13.68.

## Discussion

Taken together, these results suggests that Sans Forgetica might not be a desirable difficulty. While it has been reported that Sans Forgetica font can enhance performance (see Eskenazi & Nix, 2020), we report results from two high-powered experiments arguing against this claim. Specifically, we demonstrated that Sans Forgetica does not enhance

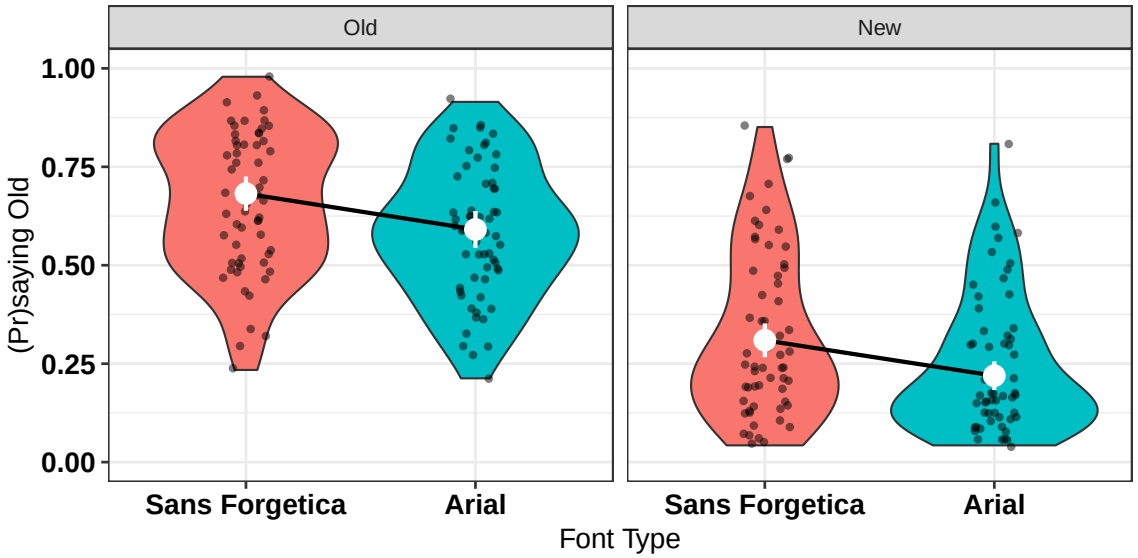


*Figure 5*. Mean proportions of "old" responses for Experiment 3. Violin plots represent the kernal density of average probability (black dots) with the fixed effect mean (white dot) and 95% CIs derived from the glmer model.

recall for cue-target pairs (Experiment 1) or words embedded in sentences from a passage (Experiment 2). This adds to the increasing literature showing that perceptual disfluency has very little impact on actual memory performance (e.g., Magreehan et al. (2016); Rhodes and Castel (2008); Rhodes and Castel (2009); Rummer et al. (2016); Xie et al. (2018); Yue et al. (2013)), While Sans Forgetica did not produce a memory benefit, we did observe a memory advantage for items that had to be generated (Experiment 1) and that were pre-highlighted, thereby replicating previous results ().

## Limitations

**Transfer-appropriate processing.** In both experiments we looked at cued recall. A recent transfer-appropriate processing (TAP) framework has contextualized when difficulties are desirable and when they are not (McDaniel & Butler, 2011). Its essence emphasizes the qualitative mismatch of the evoked encoding processes of the applied difficulty (and by the material) with respect to the required retrieval processes of the memory test. Thus, one important aspect postulated by this framework denotes the specific encoding processes stimulated by the type of difficulty applied. For example, generating incomplete word-fragments within a text intensifies the processing of the word cue and the word surroundings that help to identify the word, thus enhancing proposition-specific encoding. In contrast, creating sentence coherency in a text with randomized sentences intensifies the processing of the relationships of information in the text, thus enhancing relational encoding (McDaniel, Hines, Waddill, & Einstein, 1994). Consequently, the generation-task, which required word-generation, led to improved verbatim recall, but it was not desirable for relational test questions (and vice versa). These differently evoked encoding processes

(proposition-specific versus relational) by the generation task predicted different memory effects. It is thus possible that the Sans Forgetica effect arises only under certain memory paradigms (e.g., free recall or recognition). It is hard to test this however, as the mechanisms that give rise to the effect are unclear and currently there is not strong evidence that the Sans Forgetica effect is reliable. Future research should explore different testing conditions.

**Processing Difficulty.** One criticism put forth when examining perceptual disfluency is that studies do not objectively test (e.g., by using RTs) that stimuli are in fact perceptually disfluent [see Geller et al. (2018)). Given that the two experiments contained herein were presented on-line using the Qualtircs platform, it was difficult to test this assumption. However, recently, a eye-tracking study by (Eskenazi & Nix, 2020) provided evidence that Sans Forgetica is perceptually disfluent. In their study, as better spellers had longer gaze duration and spent more total time on words presented in Sans Forgetica than poor spellers. This suggests that Sans Forgetica is perceptually disfluent as long as you are a good speller.

## Conclusion

The two experiments herein present evidence against claims put forth by its creators and the media [also see Eskenazi and Nix (2020)). We concede that our conclusions of no effect might be a bit premature. It is possible that there is an effect of Sans Forgetica, but the effect size might be smaller than we could detect across our two studies. We powered our studies to detect a medium-sized effect. Further, as noted by (Eskenazi & Nix, 2020) and others [Geller2018; Geller2019] there are important moderating factors of the disfluency effect that should be considered. Once more research is published, a meta-analysis can be conducted to determine the effect size and any moderating factors of the Sans Forgetica effect. Regardless, it is our conclusion that Sans Forgetica lives up its name. Students looking to remember more and forget less should use other "power tools" shown to enhance learning.

## References


Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

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

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

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In *Psychology and the real world: Essays illustrating fundamental contributions to society.* (pp. 56–64). New York, NY, US: Worth Publishers.

Bjork, R. A., & Yue, C. L. (2016). Commentary: Is disfluency desirable? Springer New York LLC. https://doi.org/10.1007/s11409-016-9156-8

Buchanan, E. M., De Deyne, S., & Montefinese, M. (2019). A practical primer on processing semantic property norm data. *Cognitive Processing*. https://doi.org/10.1007/s10339-019-00939-6

Butler, A. C., Marsh, E. J., Slavinsky, J. P., & Baraniuk, R. G. (2014). Integrating Cognitive Science and Technology Improves Learning in a STEM Classroom. *Educational Psychology Review*, *26*(2), 331–340. https://doi.org/10.1007/s10648-014-9256-4

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Carpenter, S. K. (2016). Spacing effects on learning and memory. In *The curated reference collection in neuroscience and biobehavioral psychology* (pp. 465–485). Elsevier Science Ltd. https://doi.org/10.1016/B978-0-12-809324-5.21054-7

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin and Review*, *13*(5), 826–830. https://doi.org/10.3758/BF03194004

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, *19*(22), 3127–3131. https://doi.org/10.1002/1097-0258(20001130)19:22<3127::AID-SIM784>3.0.CO;2-M

Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2011). Fortune favors the: Effects of disfluency on educational outcomes. *Cognition*, *118*(1), 111–115. https://doi.org/10.1016/j.cognition.2010.09.012

Earp, J. (2018). Q&A: Designing a font to help students remember key information.

Eskenazi, M. A., & Nix, B. (2020). Individual Differences in the Desirable Difficulty Effect During Lexical Acquisition. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/xlm0000809

Fowler, R. L., & Barker, A. S. (1974). Effectiveness of highlighting for retention of text material. *Journal of Applied Psychology*, *59*(3), 358–364. https://doi.org/10.1037/h0036750

French, M. M., Blood, A., Bright, N. D., Futak, D., Grohmann, M. J., Hasthorpe, A., ... Tabor, J. (2013). Changing fonts in education: How the benefits vary with ability and dyslexia. *Journal of Educational Research*, *106*(4), 301–304. https://doi.org/10.1080/00220671.2012.736430

Geller, J., Still, M. L., Dark, V. J., & Carpenter, S. K. (2018). Would disfluency by any other name still be disfluent? Examining the disfluency effect with cursive handwriting. *Memory and Cognition*, *46*(7), 1109–1126. https://doi.org/10.3758/s13421-018-0824-6

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods*, *39*(1), 101–117. https://doi.org/10.3758/BF03192848

Kornell, N., & Vaughn, K. E. (2016). How Retrieval Attempts Affect Learning: A Review and Synthesis. *Psychology of Learning and Motivation - Advances in Research and Theory*, *65*, 183–215. https://doi.org/10.1016/bs.plm.2016.03.003

Lakens, D., & Evers, E. R. K. (2014). Sailing From the Seas of Chaos Into the Corridor of Stability: Practical Recommendations to Increase the Informational Value of Studies. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *9*(3), 278–292. https://doi.org/10.1177/1745691614528520

Locker, L., Huffman, L., & Bovaird, J. A. (2007). On the use of multilevel modeling as an alternative to items analysis in psycholinguistic research. *Behavior Research Methods*, *39*(4), 723–730. https://doi.org/10.3758/BF03192962

Magreehan, D. A., Serra, M. J., Schwartz, N. H., & Narciss, S. (2016). Further boundary conditions for the effects of perceptual disfluency on judgments of learning. *Metacognition and Learning*, *11*(1), 35–56. https://doi.org/10.1007/s11409-015-9147-1

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

McDaniel, M. A., Hines, R. J., Waddill, P. J., & Einstein, G. O. (1994). What Makes Folk Tales Unique: Content Familiarity, Causal Structure, Scripts, or Superstructures? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(1), 169–184. https://doi.org/10.1037/0278-7393.20.1.169

Mulligan, N. W. (1996). The effects of perceptual interference at encoding on implicit memory, explicit memory, and memory for source. *Journal of Experimental Psychology:*

*Learning Memory and Cognition*, *22*(5), 1067–1087. https://doi.org/10.1037/0278-7393.22.5.1067

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. Psychonomic Society Inc. https://doi.org/10.3758/BF03195588

Ooms, J. (2018). *Hunspell: High-performance stemmer, tokenizer, and spell checker*. Retrieved from https://CRAN.R-project.org/package=hunspell

Rhodes, M. G., & Castel, A. D. (2008). Memory Predictions Are Influenced by Perceptual Information: Evidence for Metacognitive Illusions. *Journal of Experimental Psychology: General*, *137*(4), 615–625. https://doi.org/10.1037/a0013684

Rhodes, M. G., & Castel, A. D. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin and Review*, *16*(3), 550–554. https://doi.org/10.3758/PBR.16.3.550

Rosner, T. M., Davis, H., & Milliken, B. (2015). Perceptual blurring and recognition memory: A desirable difficulty effect revealed. *Acta Psychologica*, *160*, 11–22. https://doi.org/10.1016/j.actpsy.2015.06.006

Rummer, R., Schweppe, J., & Schwede, A. (2016). Fortune is fickle: null-effects of disfluency on learning outcomes. *Metacognition and Learning*, *11*(1), 57–70. https://doi.org/10.1007/s11409-015-9151-5

Silvers, V. L., & Kreiner, D. S. (1997). The effects of pre-existing inappropriate highlighting onreading comprehension. *Reading Research and Instruction*, *36*(3), 217–223. https://doi.org/10.1080/19388079709558240

Sungkhasettee, V. W., Friedman, M. C., & Castel, A. D. (2011). Memory and metamemory for inverted words: Illusions of competency and desirable difficulties. *Psychonomic Bulletin and Review*, *18*(5), 973–978. https://doi.org/10.3758/s13423-011-0114-9

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Xie, H., Zhou, Z., & Liu, Q. (2018). Null Effects of Perceptual Disfluency on Learning Outcomes in a Text-Based Educational Context: a Meta-analysis. *Educational Psychology Review*, *30*(3), 745–771. https://doi.org/10.1007/s10648-018-9442-x

Yue, C. L., Castel, A. D., & Bjork, R. A. (2013). When disfluency is-and is not-a desirable difficulty: The influence of typeface clarity on metacognitive judgments and memory. *Memory and Cognition*, *41*(2), 229–241. https://doi.org/10.3758/s13421-012-0255-8

Yue, C. L., Storm, B. C., Kornell, N., & Bjork, E. L. (2014). Highlighting and Its Relation to Distributed Study and Students' Metacognitive Beliefs. *Educational Psychology Review*, *27*(1), 69–78. https://doi.org/10.1007/s10648-014-9277-z