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 really is 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 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

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

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 shown to be a desirable difficulty is to make material perceptually disfluent by changing the material's perceptual characteristics. In a now infamous paper, Diemand-Yauman, Oppenheimer, and Vaughan (2011) demonstrated this by placing to-be studied materials in atypical fonts (e.g., comic sans). This resulted in better memory than if the material was in a common font. This finding has been found with other perceptual manipulations such as masking (Mulligan, 1996), inversion (Sungkhasettee, Friedman, & Castel, 2011), blurring (Rosner, Davis, & Milliken, 2015), and handwriting (Geller, Still, Dark, & Carpenter, 2018). The predominate theoritical explanaton for thsi is that the experience of disfluency serves as a metacognitive (subjective) cue, and this engenders deeper processing of material (but see Geller et al., 2018 for an alterantive account). This desirable effect of perceptual disfluency on memory is called the disfluency effect (Bjork & Yue, 2016).

Given the desribale effects on memory and the relative ease of implementation, it is clear why perceptual disfluency is such an appealing strategy. However, 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 (i.e., metamemory judgements that assess future memory) (d = -0.043). In the laboratory, Geller et al. (2018) and Geller & Still (2018) manipulated several boundary conditions (e.g., level of degradation (easy-to-read vs. hard-to-read), type of judgement of learning (list-wise vs item-by-item), retention interval (3 minites vs. 24-hours), and testing expectancy (incidental vs. intentional)). Geller et al. (2018) found that you can get positive memory effects from perceptual disfluent materials, but the perceptual manipulation must be sufficnely disfluent in order to have a positive effect on memory. They found an inverted u shaped pattern showing better memory for easy to read handwritten cursive than hard to read hadnwritten cursive.

Despite the weak evidence, perceptual disfluency is still being touted as a viable learning tool. Most recently, a font called Sans Forgetica has recieved a lot of attention for its purported positive effects on memory. Sans Forgteica is a variation of a sans-serif typeface that consists of intermittent gaps in letters that are back slanted (see fig. 1). These perceptual characteristics are thought to make it desirable for learning.Indeed, Major news sources like the Washington Post (**???**) and National Public Radio (NPR; https://www.npr.org/2018/10/06/655121384/sans-forgetica-a-font-to-remember) have reported on the positive effects of Sans Forgetica on memory. The Sans Forgetica font is even available on multiple operating platforms.

What do we know about Sans Forgetica?

Figure 1. Example of Sans Forgetica font.

With all this positive attention Sans Forgetica is receiving, there must be strong evidence for it? In an inteview coducted Earrp () we are provided with some details on two unpublished studies conducted by the creators of Sans Forgetica showing postive effects. In a lab experiment (N=96), participants read 20 word highly associated word pairs (e.g., girl - guy) in one of them being Sans Forgetica) and a typical or common font. The font pairs were presented were counterbalanced across participants. What this means is that all fonts were shown, but the same pairs were never presneted in more than one type of font. Each word pair was presnted on the screen for 100 ms (that is super fast...). For a final test, they were given the cue (e.g., *girl*) and had to respond with the target (*guy*). What did they find? According to the interview, targets were recalled 68% of time when presented in a common font. For cue-target pairs in SF, targets were recalled 69% of the time–a negeliable difference.

In an online experiment, participants were presented passages (250 words in total) where one of the paragraphs was presented in SF. Each participant saw five different texts in total. For each text they were asked one question about the part written in SF and another question about the part written in standard Arial. Participants remembered 57% of the text when a section was written in Sans Forgetica, compared to 50% of the surrounding text that was written in a plain Arial font.

talk about failure to replicate studt by (**???**)

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 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 has only been two empirical studies published examining the effectviness of Sans Forgetica in generating a desirable difficulty (Eskenazi & Nix, 2020). In a successful replication, 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 (learning orthographic and semantic features of a word), it is not clear if the benefits of Sans Forgetica font extends to other memory processes. (**???**) provide some evidence that it does not. In conduceted one of the first set of empirical studies examining the Sans Forgetica memory

benfit across several experiments showing no effect for Sans Forgetica in paired-assoicate cued recall and prose recall. In three high-powered pre-registered studies, we conducted conceptual repliactions and extensions of oftheir findings. Expeiments 1 and 2 served to conceptual replicate their findings. In Experiment 3, we extend these findings to recognition memory. In addition, we aimed to compared the Sans forgetica effect with other notable learning technqies–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 While (**???**) examined the effect of Sans Forgetica font on cued-recall memory using highly associated words, we used wealkly assoicated word pairs. It is possible that (**???**) failed to find a Sans Forgetica effect because of the high cue strength. With highly associated paris, successful retreial is due to the highly associated nature of the pairs, and not recollection processes per se. To this end we use weakly associated pairs to examine the effect of Sans forgetica and generation on memory. We predict that 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. Further, if it is similar to the generation effect, the magnitude of the memory benefit between the two should be similar.

Method

Sample size, experimental design, hypotheses, outcome measures, and analysis plan for each experiment were pre-registered and can be found on the Open Science Framework (https://osf.io/mjcn9). Raw data and R scripts for analysis and plots can be found at https://osf.io/m42wq/.

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 sized 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 24 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. This left us with 22 weakly realted cue-target pairs. 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 the 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. For all the experiments described, an alpha level of .05 is maintained. Cohen's d and generalized eta-squared (; Olejnik & Algina, 2003) are used as effect size measures. Alongside traditional analyses that utilize null hypothesis significance testing (NHST), I also report the Bayes' factors for null findings when conducting my planned comparisons. One shortcoming of NHST is that it does not allow one to measure support that there is a true null difference between conditions. Within the NHST framework, a null result could occur if there truly is no effect of a manipulation, but also if the experiment did not have enough power. Additional Bayes' factor analyses is particularly important in the following set of experiments, as prior studies have failed to find a Sans Forgetica effect. All prior probabilities are Cauchy distributions centered at zero, and effect sizes are specified through the r-scale, which is the interquartile range (i.e., how spread out the middle 50% of the distribution is). For the null model, the prior is set to zero. All data were analyzed in R (vers. 3.5.0; R Core Team, 2019), with models fit using the afex (vers. 0.27-2; Singmann, Bolker, Westfall, Aust, & Ben-Shachar, 2020) and BayesFactor packages (vers. 0.9.12-4.2; Morey & Rouder, 2018).

Results and Discussion

Per our pregreistation, cued-recall accuracy was analyzed with a A 2 (DT) x 2 () Mixed ANOVA. There was no difference in cued recall between the Generation and Sans Forgetica groups, $F(1, 230) = 0.19$, ges=<.001, p = .661. b = -0.09, SE = 0.11, 95% CI [-0.30, 0.13], $p = 0.431$, $d = 0.05$). Individuals recalled more disfluent target words than fluency target words, $F(1, 230) = 25.31$, ges=.017, $p < .001$. This was qualified by an interaction between difficulty type and disfluency, $F(1, 230) = 25.06$, ges = .017, $p <$ .001. A Bayesian ANOVA indicated strong evidence for the interaction model over the main effects model, $BF_{10} > 100$. 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]
```

Univariate Type III Repeated-Measures ANOVA Assuming Sphericity

```
     Sum Sq num Df Error SS den Df  F value    Pr(>F)
```

(Intercept) 54.725 1 23.3314 230 539.4760 < 2.2e-16 ***condition 0.020 1 23.3314 230 0.1933 0.6606***

dis 0.477 1 4.3304 230 25.3118 9.827e-07 condition:dis 0.472 1 4.3304 230 25.0599 1.105e-06 *** — Signif. codes: 0 “***’ 0.001 ’’ 0.01 ’*” 0.05 “.” 0.1 ’ ’ 1

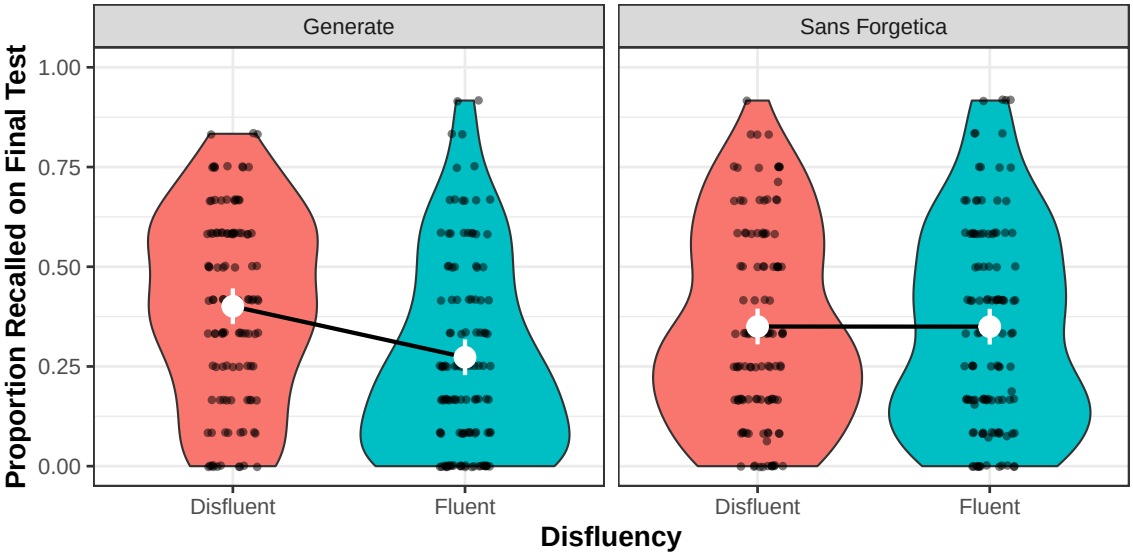

Figure 2. Accuracy on cued recall test. Violin plots represent the kernal density of aveargе accuracy(black dots) with the fixed effect mean (white dot) and 95% CIs derived from the glmer model.

The results for Experiment 1 are clear-cut. Cued recall for Sans Forgetica font was similar to normal, Arial, font. Thus, there was no Sans Forgetica benefit. We did, however, observe greater recall for generated items, which replicates decades of literature on the generation effect (Bertsch et al., 2007). Taken together, these results suggest 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

In one of the few studies to assess the Sans Forgetica effect ((**???**)), a within-subjects design was used wherein a participnat sees two levels of memoranda—a fluent level and disfluent (Sans Forgetica) level. (**???**) has argued that the utilization of a within-subjects design might have the undesriable consequence of masking a disfluency effect. That is, deeper processing evoked by disfluent items might carry-over to the fluent items (**???**). This could be a potential reason (**???**) failed to find a Sans Forgetica effect. To remedy this, Experiment 2 tested the mnemonic effects of Sans Forgetica with a between-subjects manipulation. Instead of using simple cue-target paris, we examined memory for sentences presented in Sans Forgetica which is more represenative of what students do while studying. In addition to examining the effects of Sans Forgetica, we also looked at the effects of pre-highlighing. One of the main functions of the Sans Forgetica font is to highlight information one needs to remember. This is similar to pre-highlighting, wherein important study information is highlghted prior to studying. It has been shown that 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 prose 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 moderated by task design (within vs. between) 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 draw attention to the material.

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. We report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study. 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 the 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 [^2]: 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] 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 manip-

ulated 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 after reading the passage. The two questions were: "Do you feel that the presentation of the material helped you remember it better" 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

Per our pregreistation, cued-recall accuracy was analyzed with a one-way ANOVA (Passage Type: Pre-highlighting vs. Sans Forgetica vs. Unmodified). 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, $t(525) = 2.45$, $SE = 0.028$, $p = .039$, $d = 0.26$. There was weak evidence for no effect between sentences presented in Sans Forgetcia and pre-highlighted, $t(525) = 0.049$, $SE = 0.028$, $p = .202$, $d = 0.18$, $\text{BF}_{01} = 2.36$. Critically, there was no difference between sentences presented normally or in Sans Forgetcia, $t(525) = 0.02$, $SE = 0.028$, $p = .734$, $d = 0.079$. A Bayes factor indicated strong evidence of no effect between the two conditions ($\text{BF}_{01} = 6.47$).

Exploratory Analysis

In Experiment 2 we also asked students about their metacognitive awareness of the manipulations. 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, $t(525) = 3.56$, $p = .406$.

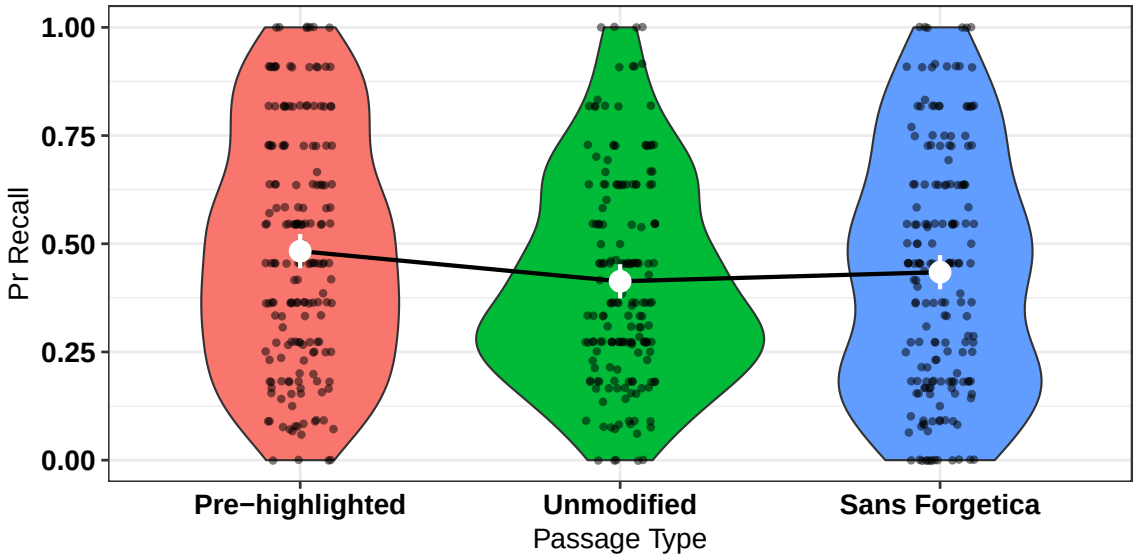

Figure 3. Probablity of recall as a function of passage type. Violin plots represent the kernal density of avearge accuracy(black dots) with the fixed effect mean (white dot) and 95% CIs derived from the ANOVA model.

Passage	emmean	SE	df	lower.CL	upper.CL
Pre-highlighted	50.31250	1.965191	525	46.45190	54.17310
Unmodified	57.39205	1.965191	525	53.53144	61.25265
Sans Forgetica	53.82955	1.965191	525	49.96894	57.69015

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. With a between-subjects design, it is not uncommon to observe no JOL differences between fluent and disfluent materials (cf. Geller et al., 2018). Indeed, (**???**) showed JOL differences between passages presented in normal (Arial) font and Sans Forgetica using a within-subject design. We did, however, find a JOL effect for pre-highlighted information. 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

In (**???**) Sans Forgetica was tested in using cued recall. In previous studes, perceptual disfluency has been shown to enhance performance on yes/no recognition tests, even when recall does not show the same benefits (Geller et al., 2018; Rosner et al., 2015). The proposed reason for this discrepancy is that during the initial perceptual identification process, the learner is focusing on surface-level aspects. Doing so would aid later recognition, but not recall, for fluent items, given that recall relies more on item elaboration than on

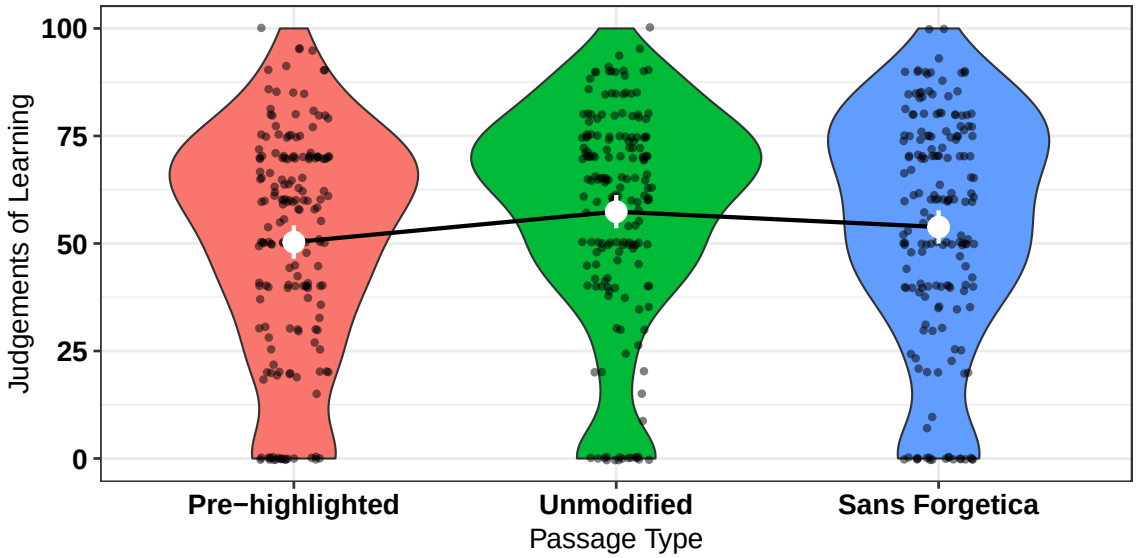

Figure 4. Judgements of learning as a function of passage type.

perceptually distinctive features (Nairne, 1988). In Experiment 3, we tested whether Sans Forgetica would lead to similar benefits in recognition memory. It is possible then that Sans Forgetica serves to increase surface-level familiarity of a word, while recollection is unchanged. This is tested in Experiment 3 by employing an old-new recognition test.

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. Sample size was determined by a similar procedure to the above experiments. No participants had to be thrown out for failing to meet the exclusion criteria noted above.

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

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 tasks are available to preview and copy from Gorilla Open Materials at https://gorilla.sc/open materials/72765.

After reading and signing a consent form, participants first completed the 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 chose 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 discussion

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 (DeCarlo, 1998). In its simplest from, SDT models are probit regressions. To estimate the SDT parameter of interest (d'), we fit a logistical mixed model (with a probit link) to participant responses (sayold; whether participants said "old" or "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 model, 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. We hypothesized that there would be no difference in d' between Sans Forgetica and Arial font.

Hit rates and false alarm rates can be seen in Fig. 3. The results are straight-forward. Individuals were more biased to say Sans Forgetica stimuli were old, $b = 0.26$, $SE = 0.026$,

95% CI [0.217, 0.319], $p < .005$. Consistent with our hypothesis, there was no difference in d' between Sans Forgetica and Arial fonts, $b = 0.033$, $SE = 0.05$, 95% CI [-0.138, 0.065], $p = .519$. There was strong evidence for no effect ($BF_{01} = 13.68$).

Similar to experiments 1 and 2, we did not find an effect of Sans Forgetica font on recognition memory.

Discussion

The purpose of the three experiments was to determine whether Sans Forgetica served as a desirable difficulty. With over 800 participants the main finding from all three experiments is that Sans Forgetica is not a desirable difficulty. While it has been claimed in unpublished and published studies (Eskenazi & Nix, 2020) that Sans Forgetica has a positive effect on memory, we report results from three high-powered memory experiments arguing against this claim. Specifically, we demonstrated that Sans Forgetica does not enhance recall for cue-target pairs (Experiment 1), words embedded in sentences from a passage (Experiment 2), or recognition memory (Experiment 3). 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 & Castel, 2008, 2009; Rummer et al., 2016; Xie et al., 2018; Yue et al., 2013). Nonsurprisingly, we did observe a memory advantage for items that had to be generated (Experiment 1) and that were pre-highlighted (Experiment 2).

Limitations

In order for perceptual disfluency to have an effect on memory it must be sufficently disfluent (Geller et al., 2018; Rosner et al., 2015). In many studies perceptual disfluency is assumed, but never explicitly tested. Thus, it could be that the failure to observe an effect in the current set of studies is because Sans Forgetica font is not perceptually disfluent. Although we did not pre-register explicit hypotheses about the the perceptual disfluncy of the Sans Forgetica font, there is some preliminary evidence that Sans Forgetica is not disfluent. In general, perceptual disfluency is thought to lower JOLs and produce longer latencies (Geller et al., 2018; Xie et al., 2018). In Experiment 2, Sans Forgetica font did not produce lower JOLs. In Experiment 3, we collected self-paced reading times for each stimulus. Self-paced reading times have been used as an objective proxy for disfluency (see Carpenter & Geller, 2020). Looking at the difference in self-paced reading times, we did not observe a significant difference between Sans Forgetica ($M = 1481$ ms, $SD = 1750$ ms) and Arial ($M = 1500$ ms, $SD = 2344$ ms) fonts, $t(59) = 0.469$, $p = 0.641$. This could explain why we did not observe an effect of Sans Forgetica on memory across three experiments.

There are a number of boundary conditions or moderating factors that determine whether perceptual disflunecy is desirable or not (Eskenazi & Nix, 2020, @Geller2018, Geller & Still, 2018). This makes it impossible to test ever single moderating factor in a single paper. We concede that it is possible that the Sans Forgetica effect does have positive effects, but is limited to a certain set of conditions. For instance, (Eskenazi & Nix, 2020) showed that Sans Forgetica can have a desirable effect, but only if you are are a good

speller. Better spellers are thought to have a more precise mental lexicon which allows for more efficient processing at multiple levels of representation (i.e,, orthographic, phonological, and semantic; Perfetti, 2007). When confronted with perceptual degradation, better spellers would be able to process a stimulus at a deeper level, which could give rise to better memory. It is unclear how this could explain some of the results obtained herein. For instance, in Experiment 3, we used high-frequency words that are well known. Nonetheless, future studies should examine spelling ability as a potential moderating factor into the perceptual disfluency effect.

Lastly, it is possible that the effect size of the Sans Forgetica effect is smaller than we could detect across our three studies. We powered our studies to detect a medium-sized effect (d=.35). If the Sans Forgetica effect is small, it is not clear what the educational use for it would be. If more research is published, a meta-analysis can be conducted to determine the true effect size and any moderating factors of the Sans Forgetica effect.

Conclusion

The three experiments herein present evidence against claims put forth by its creators and the media (also see Eskenazi & Nix, 2020). We ultimately recommend caution in using Sans Forgetica font. Sans Forgetica does not enhance memory. Students looking to remember more and forget less should use other "power tools" shown to enhance learning. Sans Forgetica font is really forgettable.

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

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

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., & Geller, J. (2020). Is a picture really worth a thousand words? Evaluating contributions of fluency and analytic processing in metacognitive judgements for pictures in foreign language vocabulary learning. *Quarterly Journal of Experimental Psychology*, *73*(2), 211–224. https://doi.org/10.1177/1747021819879416

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

DeCarlo, L. T. (1998). Signal Detection Theory and Generalized Linear Models. *Psychological Methods*, *3*(2), 186–205. https://doi.org/10.1037/1082-989X.3.2.186

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

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

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

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

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

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of bayes factors for common designs*. Retrieved from https://CRAN.R-project.org/package=BayesFactor

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

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2020). *Afex: Analysis of factorial experiments*. Retrieved from https://CRAN.R-project.org/package=afex

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

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