Sans Forgetica is not desirable for learning

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

[1] University of Iowa

[2] Skidmore College

Author Note

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

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 other similar perceptual disfluency effects, this led us to examine the mnemonic effects of Sans Forgetica more closely in comparison to other learning strategies. In three preregistered experiments, we tested if Sans Forgetica is really unforgettable. In Experiment 1 (N = 233), participants studied weakly related cue-target pairs with targets presented in either Sans Forgetcia or with missing letters (e.g., CUE - G_RL, the generation effect). Cued recall performance showed a robust effect of generation, 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 Forgetica font, yellow pre-highlighting, or unmodified. Cued recall for select words was better for pre-highlighted information than a unmodifed pure reading condition. Critically, presenting sentences in Sans Forgetica did not elevate cued recall compared to a unmodified pure reading condition or a prehighlighed condition. In Experiment 3 (N = 60), indiviudals did not have better discriminability for Sans Forgetica relative to a fluent condition in an old-new recognition test. Our findings suggest that Sans Forgetica really is forgettable.

Keywords: Disfluency, Recall, Desirable Difficulty, Learning and Memory

Word count: 4458

Sans Forgetica is not desirable for learning

Students want to remember more and forget less. Being able to recall and apply previously learned information is key for successful acadmeic performance at all levels. Many students are attracted to learning interventions that require little effort, but these are not always the best was to achieve durable learning. Importantly, 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 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 startegy that may impart a desirable difficulty involves changing the physical characteritics of stimuli to make it harder to read or more disfluent. Whereas making stimuli easier to proccess or more fluent produces overconfidence (predicted memory > actual memory; Rhodes and Castel (2008), Rhodes and Castel (2009)), disflueny (i.e., the subjective experience of effort during processing) serves to better calibrate learners and produces better memory. In a seminal article by Diemand-Yauman, Oppenheimer, and Vaughan (2011) demonstrated in the laboratroy and classroom that placing to be studied materials in atypical fonts (e.g., Comic Sans, Montype Corsiva, Bodoni, and Haettenshweiler) 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 (presenting a row of hashmarks before the presentation of a stimulus; Mulligan, 1996), inversion (Sungkhasettee, Friedman, & Castel, 2011), blurring (Rosner, Davis, & Milliken, 2015), and handwriting (Geller et al., 2018). The predominate theoritical explanaton for this finding is that the experience of disfluency serves as a metacognitive (subjective) cue, and this engenders deeper, more semantic, processing of material (but see Geller et al., 2018 for an alternative account). The benefit associated with processing difficulty brought forth by changing perceptual characteristics of stimuli is called the disfluency effect.

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).

Although evidence for perceptual disflueny is weak, it important to point out that there are important boundary conditions of the disfluency effect (see Geller & Still, 2018). One important boundary condition to consider is the type of disfluency manipulation used (Geller et al., 2018). Not all perceptual disfluency manipulations are created equal. Using handwritten cursive that was either easy to read or hard to read, Geller et al. (2018) observed an inverted u-shaped pattern–easy to read and hard to read cursive were better remembered than print, but easy to read cursive was better remembered than hard to read cursive. This pattern suggests that manipulations that are disfluent can produce better memory, but they need to be optimally disfluent.

Recently, a team of psychologists, graphic designers, and marketers set out to find a font providing an optimal level of disfluency desirable for learning. According to Earp

(2018), to find the optimal font, the team conducted a cued recall experiment (N=96) werein participants read 20 highly associated word pairs (e.g., girl - guy) each for 100 ms in a moderate disfluent font, an extremely disfluent font, a slighly disfluent font, and a normal, fluent, font. Results showed that the pairs in the moderate disfluent font were recalled slighly better (69%) than the normal font (68%). The team named this font Sans Forgetica. The font itself is a varaint of sans serif font with intermittent gaps in letters that are back slanted (see figure 1). The intermittent gaps of Sans Forgetica are thought to require readers to generate or fill in the missing pieces thereby producing a memory advantage. This mechanism of action is thought to be similar to that of the generation effect, wherein information is better remembered when generated or filled-in compared to if it is simply read.

Since the release of the Sans Forgetica font, there has been a lot of attention and press. Sans Forgetica was covered by major news sources like the Washington Post (**???**) National Public Radio (NPR; https://www.npr.org/2018/10/06/655121384/sans-forgetica-a-font-to-remember), and The Guardian. In 2019, Sans Forgetica won the GoodDesign, Best in Class award (Good Desigm, 2019). Sans Forgetica font can even be downloaded and used on your computer.

Despite all the attention and marketing, evidence for Sans Forgetica font is mixed. Initial evidence for the Sans Forgetica font comes from an unpublished study by the Sans Forgetica team (Earp, 2018). In an online experiment (N = 303), participants were presented passages (250 words in total) where one of the paragraphs was presented in Sans Forgetica. Each participant saw five different texts in total. For each text participants were asked one question about the part written in Sans Forgetica and another question about the part written in normal font. Placing text passages in Sans Forgetica font resulted in better memeory (57%) than if matearis were presented in normal font (50%). Additonal evidence for the Sans Forgetica font comes from a study conducted by Eskenazi and Nix (2020). They 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. Despite positive evidence for Sans Forgetica, Taylor, Sanson, Burnell, Wade, and Garry (2020) recently conducted a conceptual replication of the studies by the Sans Forgetica team and found no evidence for Sans Forgetica font enhancing memory. In fact, Sans Forgetica seemed to harm memory (Expeirment 2).

The Present Studies

The question of whether Sans Forgetica produces a mnnmenomic benefit has clear practical and theoritical implications. In the educational domain, it would be relatively quick and easy to place materials in Sans Forgetica font. However, in order for Sans Forgetica font to be useful to researchers and educators, we need to better understand the condtions under which Sans Forgetica font is and is not beneficial for learning To this end, the current set of experiments focused on the method used by Taylor et al. (2020) and The Sans Forgetica team (Earp, 2018) and aimed to identify why this procedure failed to produce a pattern of results consistent with other work [e.g., Earp (2018); Eskenazi and Nix (2020)). As the Taylor et al. (2020) study is the only published replication attempt, we thought it pertinent to follow up on their claims. In particular, we focused on two possibilities. First, that Sans Forgetica font can have a desirable difficulty effect, but that the effect may hinge on a number of specific study parameters. Second, that Sans Forgetica font simply is not a desirable difficulty and cannot be demonstrated. To this end, the current research aims to identify some of the moderating or boundary conditions of the Sans Fogetica effect, if any (Oppenheimer & Alter, 2014).

In Experiment 1, we examined the impact of cue strength and study duration by looking at weakly related cue-target pairs presented for 2 seconds. In Experiment 2, we examined design type by examining prose passage recall using a between-subjects manipulation. In Experiment 3, we examined if the type of test moderated the Sans

Forgetica effect by using a yes/no recognition test. 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.

Experiment 1

In Experiment 1 we were interested in answering two questions. Can we get a benefit for Sans Forgetica when using weakly associated pairs? If so, is the Sans Forgetica effect on memory similar in magnitude to another desriable difficulty–generation? Taylor et al. (2020) (Experiment 2) used cue-target pairs that were highly associated. Carpenter (2009) has argued that weakly related cue-target pairs produce more elaborative processing and leads to better memory especially when the targets to be remembered require generation or retreival (the elaborative retreival hypothesis). It is possible, then, that the use of highly associated pairs weakened or dampened the Sans Forgetica effect. In Experiment 1 we examined the mnemonic benefit of Sans Forgetica and generation by looking at cued recall performance. In Experiment 1 participants studied weakly associated pairs for 2 seconds to examine the effect of Sans Forgetica and generation on memory. We opted to present pairs for 2 seconds rather than the 100 ms duration used by Taylor et al. (2020) and Sans Forgetica team (Earp, 2018). With a 100 ms duration, participants might have struggled to read the word pairs properly, or to process the word pairs deeply enough, for any benefits of Sans Forgetica to take effect. 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 (Exeriments 1 and 2: https://osf.io/d2vy8/; Experiment 3: https://osf.io/dsxrc/). All raw and summary data along with R scripts for pre-processing, analysis, and plotting can be found at https://github.com/jgeller112/SF_Expt.

Participants. Two-hundred and thirty-two people from Amazon's Mechanical Turk Service participated for monetary compensation. 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 aninteraction effect size of .35 differs from zero. No participants met our pre-registered exclusion criteria (i.e., did not complete the experiment, started the experiment multiple times, experienced technical problems, or reported familiarity with the stimuli), yielding 116 participants in each between-subjects condition.

Design. Fluency (fluent vs. disfluent) was manipulated within-subjects and Disfluency Type (Generation vs. Sans Forgetica) was manipulated between participants. For the Sans Forgetica condition, disfluent targets were presented in Sans Forgetica while the fluent targets were presented in Arial font. In the Generation condition, disfluent targets were presented with missing letters (vowels were replaced by underscores) and the other half were intact (Arial font).

Materials, and Procedure. Participants were presented with 24 weakly related cue-target pairs taken from Carpenter, Pashler, and Vul (2006). 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 22 cue-target pairs were all nouns, 5–7 letters and 1–3 syllables in length, high in concreteness (400–700), high-frequency (at least 30 per million), and had similar forward and backward association strengths. Two counterbalanced lists were created for each Disfluency Type (Generation and Sans Forgetica) so that each item could be presented in each fluency condition without repeating any items for an individual participant.

Participants completed the experiment on-line via the Qualtrics survey platform hosted on Amazon Mechanical Turk and were paid XXX per hour. Participants were randomly assigned to one of two conditions: The Generation condition or the Sans Forgetica condition. Prior to studying the pairs, participants were instructed to mentally "fill in" the targets to come up with the correct target. Participants were also 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 5 seconds each. After a short 2-minute distraction task (anagram generation), participants completed a self-paced cued recall test. During cued recall, participants were presented 22 cues one at a time and asked to provide the target word. A short demographics survey followed this final test, after which participants were debriefed.

Scoring. Spell checking was automated with the hunspell package in R (Ooms, 2018). 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.

Results and Discussion

Analytical Strategy

For all the experiments described, an alpha level of .05 is maintained. Cohen's d and generalized eta-squared (η_g^2; Olejnik & Algina, 2003) are used as effect size measures. Alongside traditional analyses that utilize null hypothesis significance testing (NHST), we also report the Bayes' factors for pre-registered analyses. 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).

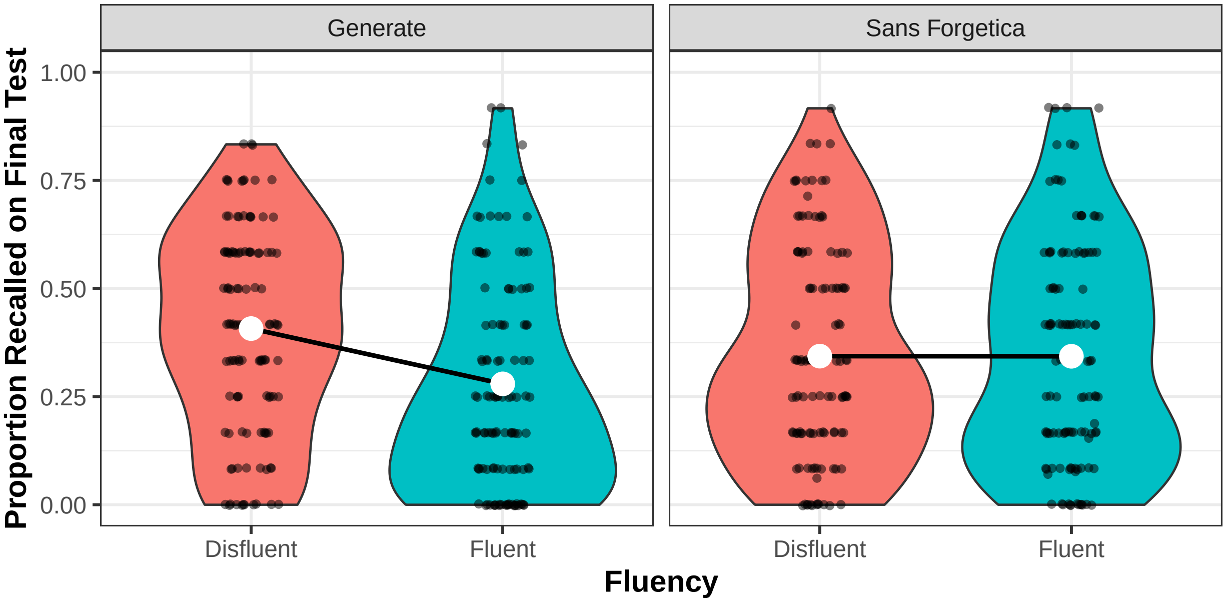
Per our pregreistation, cued recall accuracy was analyzed with a 2 (Fluency: Fluent vs. Disfluent) x 2 (Disfluency Type: Generation vs. Sans Forgetica) Mixed ANOVA. There was no difference in cued recall between the Generation and Sans Forgetica groups, $F(1, 230) = 0.19$, η_g^2 <.001, p = .752. Individuals recalled more disfluent target words than fluent target words, $F(1, 230) = 25.31$, η_g^2 =.017, $p < .001$. This was qualified by an interaction between difficulty type and disfluency, $F(1, 230) = 25.06$, $\eta_g^2 = .017$, $p < .001$. A Bayesian ANOVA indicated strong evidence for the interaction model over the main effects model, $\text{BF}_{10} > 100$. As seen in Fig. 2, the magnitude of the generation effect was larger than the Sans Forgetica effect.

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



The results for Experiment 1 are clear-cut. While the benefit of generating information was clear, there was no benefit of studying items in the Sans Forgetica font. Thus, presenting weakly associated targets in Sans Forgetica for 2 seconds produced no memory benefit. This was confirmed by a Bayesian analysis denoting strong eveidence for the presence of the interaction compared to a main effects model. Although overall accuracy was low in the current experiment (38%), it is important to note that the level of performance was comparable to what was observed by Carpenter et al. (2006) using the same materials (30% overall for study trial). In additon, accuracy was comparable to Experiment 2 from Taylor et al. (2020), with highly associated cue target pairs (~45%). 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 like generation is.

Experiment 2

Taylor et al. (2020) examined recall performance for Sans Forgetica using a within-subjects manipulation wherein a participant sees two levels of memoranda—a fluent level and disfluent (Sans Forgetica) level. Doing this, they did not see a memory benefit for Sans Forgetica font. In our Experiment 1, we also did not observe a benefit. (**???**) 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 (**???**) and our (Experiment 1) 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-highlighting. One of the main functions of the Sans Forgetica font is to highlight information one needs to remember. This is similar to pre-highlighting, whereby important study information is highlghted prior to studying. This is used by instructors and textbook creators to enhance learning. Indeed, it has been shown that when students read pre-highlighted passages, there is some evience that they may 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

Participants. Five hundred and twenty-eight undergraduates ($N = 528$) participated for partial completion of course credit. After excluding participants based on our preregistered exclusion criteria, we were left with unequal group sizes. After excluding participants based on our preregistered exclusion criteria ($n = 111$), we were left with unequal group sizes. Because of this, we ran several more participants per group, yielding 176 participants in each of the three conditions.

Materials. 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 yellow (pre-highlighted), Sans Forgetica, 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 attention check question at the start of the final test: “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. 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 11 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). The one-way ANOVA was significant, $F(2, 525) = 3.16$, $\eta_g^2 = .012$, **p* = .043. 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?” To examine differences we conducted seperate independent t-tests. Looking at JOLs, the unmodified passage was 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.

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 a unmodified or pre-highlighted passage. 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 (@ Magreehan et al., 2016; Yue et al., 2013). Indeed, using a within-subjects manipulation of fluency, (**???**) showed JOL differences between passages presented in normal (Arial) font and Sans Forgetica. 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. This suggests that pre-highlighted information might serve as a more powerful metacognitive cue than Sans Forgetica.

Experiment 3

In Experiments 1 and 2 we tested the Sans Forgetica effect using cued recall. In previous studes, perceptual disfluency has been shown to enhance performance on yes/no recognition tests, even when there is no recall benefit (Nairne, 1988). 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 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.

Method

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. Disfluency (Sans Forgetica vs. Fluent) was the single factor, manipulated within-subjects. We used 188 words, 94 at study (47 in each script condition) and 188 at test (94 old and 94 new). Words were counterbalanced across the disfluency and study/test conditions, such that each word served equally often as a target and a foil in both fonts. The experiment was created and conducted using the Gorilla Experiment Builder [Anwyl-Irvine, Massonnié, Flitton, Kirkham, and 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. 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, with old words always presented in the same script at test as it was at study.

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.

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

D' values along with hit rates and false alarm rates can be seen in Fig. 3. The results are straight-forward. Consistent with our hypothesis, there was no difference in d' between Sans Forgetica and Arial fonts, t (59) = 0.281 , SE = 0.05, 95% CI [-0.096, 0.127], p = .780. There was strong evidence for no effect (BF_{01} = 13.68).

We did not find an effect of Sans Forgetica font on recognition memory. Given that we did not observe a significant effect of Sans Forgetica in Experiment 1 and 2 using cued recall, it does not seem like Sans Forgetica is moderated by type of test.

Bayes factor analysis

[1] Alt., r=0.707 : 0.1466792 ±0%

Against denominator: Null, mu = 0 — Bayes factor type: BFoneSample, JZS

General Discussion

The creators of the Sans Forgetica font as well as the media have made strong claims regarding the mnemonic benefit of Sans Forgetica font. The aim of the current experiments was to replicate and extend the findings of Taylor et al. (2020). In Experiment 1, we did not show a mnemonic benefit for Sans Forgetica font in a cued recall task with weakly realted cues and presented for a longer duration (2 seconds). In Experiment 2, using a between-subjects design, we did not show a mneminic benefit for Sans Forgetica font for a prose passage on ground water. In Experiment 3, we did not find a mnemonic benefit for Sans Forgetica using a yes/no recognition test. Similar to Taylor et al. (2020), we did not find evidence for a mnemonic benefit of Sans Forgetica font. While it has been claimed, in unpublished and published studies (Earp, 2018; Eskenazi & Nix, 2020), that Sans Forgetica has a positive effect on memory, our high-powered studies with over 800 participants argue against this claim. The main conclusion drawn from all three experiments is that Sans Forgetica is not a desirable difficulty. Theortically, these findings add 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 find a memory advantage for other learning techniques such as generation (Experiment 1) and pre-highlighting (Experiment 2).

What might account for the null effect of Sans Forgetica font? In many studies, perceptual disfluency is assumed but never objectively 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 preregister explicit hypotheses about objective disfluncy of the Sans Forgetica font, we have some preliminary evidence that Sans Forgetica is not objectively disfluent. In Experiment 3, we collected self-paced study times for each stimulus. (We did not collect readings times in Experiments 1 and 2 because it is not clear how prescise timing is within Qualtircs whereas Gorilla boosts millisecond accuray Anwyl-Irvine et al., 2020). Self-paced study 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 font (M = 1481 ms, SD = 1750 ms) and Arial font (M = 1500 ms, SD = 2344 ms) fonts, $t(59) = 0.469$, $p = 0.641$. Given this lack of difference, this could be a potentional explanation for why we did not observe an effect of Sans Forgetica font. It is worth noting, however, that self-paced study times might reflect things other than, or in addition to, processing disfluency. Although self-paced study provides one way of measuring processing fluency, more precsise measures of processing fluency should be considered as well (but see Eskenazi & Nix, 2020 for eye-tracking evidence for Sans Forgetica font in good spellers).

A number of boundary conditions that determine when perceptual disfluency will and

and will not be a desirable difficulty have been established over the past several years (Geller et al., 2018, Geller & Still, 2018).In the current set of experiments, we examined whether things such as cue strenth, study time, design type, and type of test influenced whether or not we could find a mnemonic effect of Sans Forgetica on memory. We did not find any evidence the Sans Forgetica effect is moderated by these factors. We cannot rule out that the postive benefits of Sans Forgetica might arise under different conditions, however. For instance, (Eskenazi & Nix, 2020) showed that Sans Forgetica can lead to better orthographic distinctiveness and semantic acquistion, but only if you are are a good speller. This is because 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. Furture research should examine the role of individual difference measures such as spelling ability or working memory capacity along with other design factors not tested (e.g., test delay).

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, particularly given that the generation and pre-highlighting manipulations were more effective.

Conclusion

Students are attracted to learning interventions that are easy to implement (Geller, Toftness, et al., 2018). It is not surprising why Sans Forgetia font has gained so much media attention. However, in our current age of uncertainty, it is important to properly evaluate scientific claims made by the media, even if that information comes from widely trusted news soruces like the the Washington Post, NPR, or the Guardian. As scientists, our job is to properly evaulate the evidence and correct errorenous information. From a practical standpoint, we we have to argue against the claims made by the Sans Forgetica team and various news outlets and conclude that Sans Forgetica should not be used as a learning technque to bolster learning. Our results suggest that placing material in Sans Forgetica font does not lead to more durable learning. It is our reccomendation to students looking to remember more and forget less, that they use learning tools such as testing or spacing that have stood the test of time.

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.

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. (2009). Cue Strength as a Moderator of the Testing Effect: The Benefits of Elaborative Retrieval. *Journal of Experimental Psychology: Learning Memory and Cognition*, *35*(6), 1563–1569. https://doi.org/10.1037/a0017021

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

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

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

Geller, J., Toftness, A. R., Armstrong, P. I., Carpenter, S. K., Manz, C. L., Coffman, C. R., & Lamm, M. H. (2018). Study strategies and beliefs about learning as a function of academic achievement and achievement goals. *Memory*, *26*(5), 683–690. https://doi.org/10.1080/09658211.2017.1397175

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

Nairne, J. S. (1988). The Mnemonic Value of Perceptual Identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(2), 248–255. https://doi.org/10.1037/0278-7393.14.2.248

Olejnik, S., & Algina, J. (2003). Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. https://doi.org/10.1037/1082-989X.8.4.434

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

Oppenheimer, D. M., & Alter, A. L. (2014). The Search for Moderators in Disfluency Research. *Applied Cognitive Psychology*, *28*(4), 502–504. https://doi.org/10.1002/acp.3023

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357–383. https://doi.org/10.1080/10888430701530730

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*

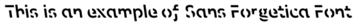
Bulletin and Review, *18*(5), 973–978. https://doi.org/10.3758/s13423-011-0114-9

Taylor, A., Sanson, M., Burnell, R., Wade, K. A., & Garry, M. (2020). Disfluent difficulties are not desirable difficulties: the (lack of) effect of Sans Forgetica on memory. *Memory*, 1–8. https://doi.org/10.1080/09658211.2020.1758726

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



This is an example of Sans Forgetica Font

Figure 1. Example of Sans Forgetica font.

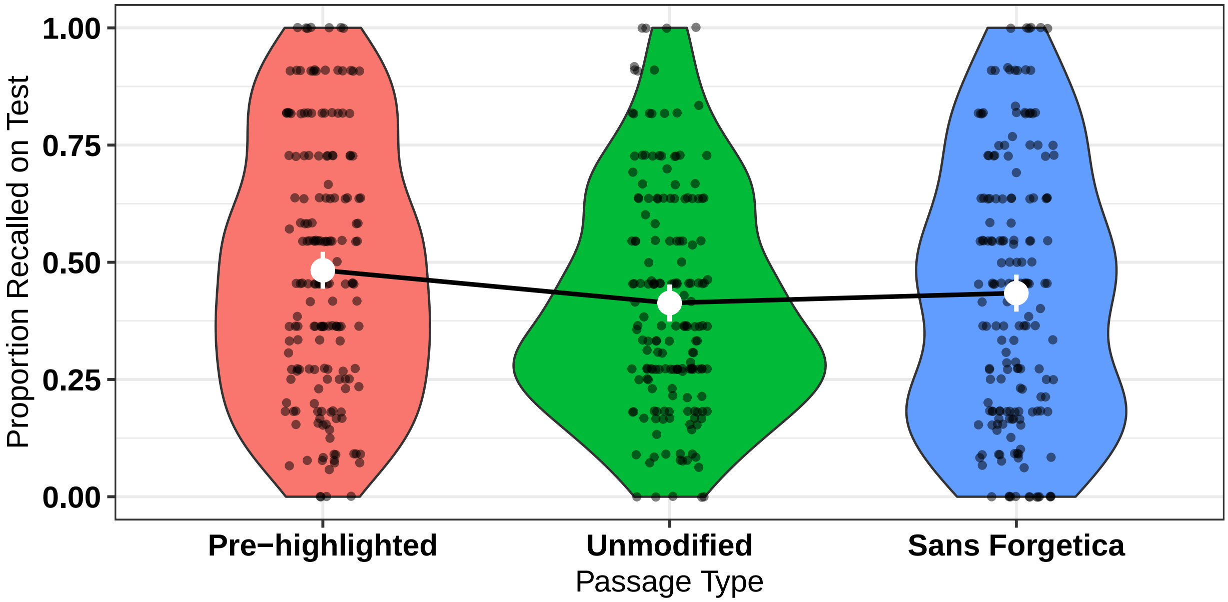


Figure 2. Proportion recalled as a function of passage type. Violin plots represent the kernal density of average accuracy (black dots) with the fixed effect mean (white dot) and 95% CIs derived from the ANOVA model.

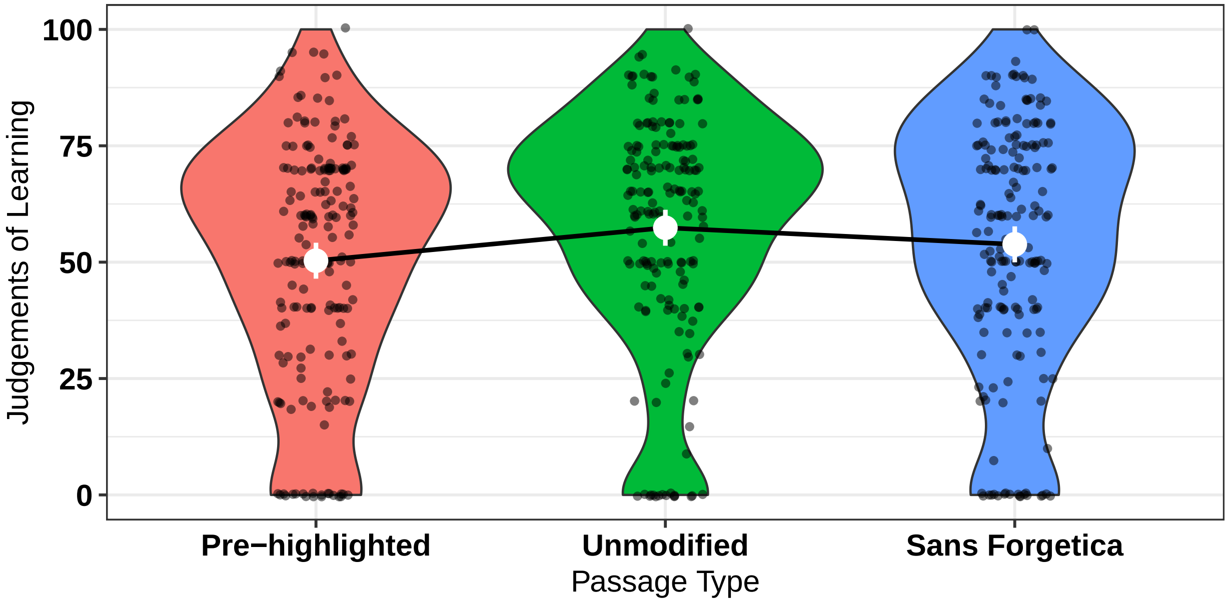


Figure 3. Judgements of learning as a function of passage type.

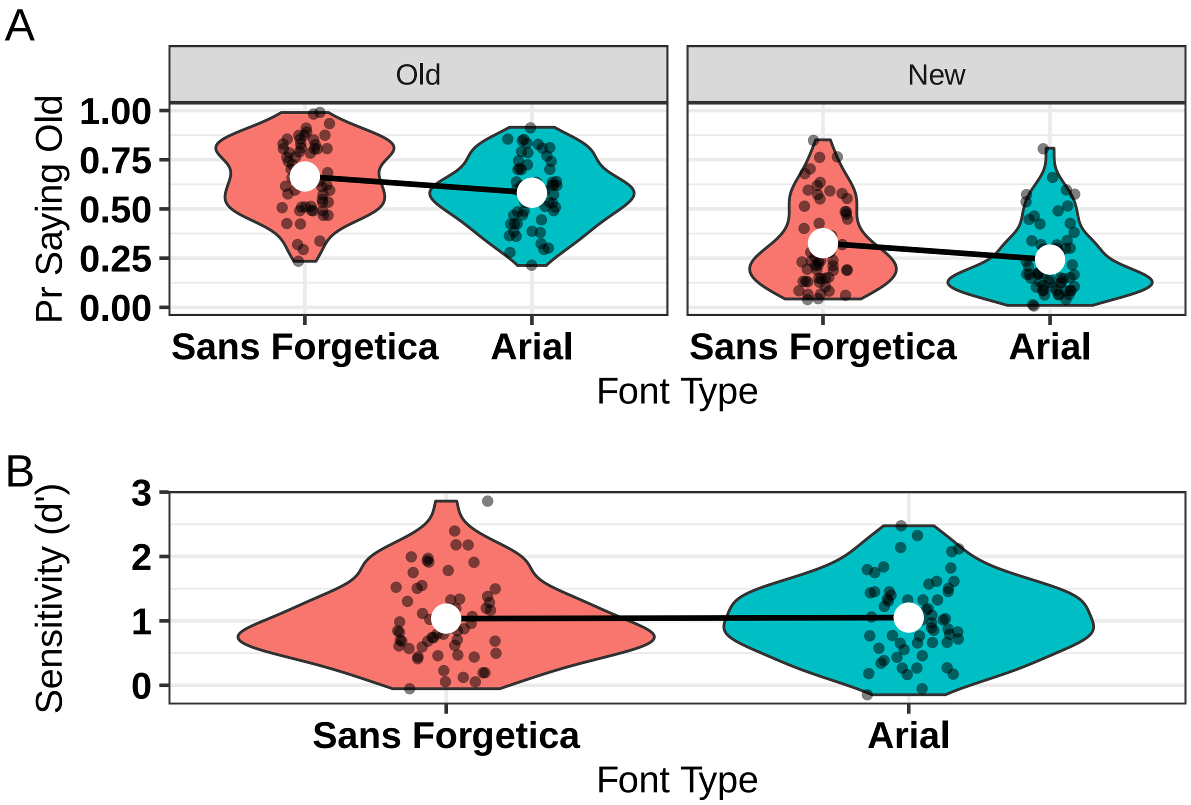


Figure 4. A. Mean proportions of "old" responses. Violin plots represent the kernal density of average probability (black dots) with the mean (white dot) and within-subject 95% CIs. B. Memory sensitivity (d'). Violin plots represent the kernal density of average sensitivity (black dots) with the mean (white dot) and within-subject 95% CIs