Sans Forgetica is Really Forgettable

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

[1] University of Iowa
[2] Skidmore College

Do students learn better with material that is perceptually harder-to-process? One way to to make material harder-to-process is by placing it in an atypical font. One font claimed to enhance memory is Sans Forgetcia, despite little to no empirical evidence that Sans Forgetica . In two preregistered experiments, we tested if Sans Forgetica is really unforgetable. In Experiment 1 (*N* = 215), participants studied weakly realted cue-target word 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 on ground water with select sentences presented in either Sans Forgetcia, yellow highlighting, or unchanged. Cued recall for selelct words were better for pre-highlighted information than when no changes to the passage were made. Critically, presenting sentences in Sans Forgetica did not produce better cued recall than pre-highlighted sentences or sentences presented unchanged. Our findings suggests that Sans Forgetica is really forgeticable.

Keywords: keywords
Word count: X

Students want to remember more and forget less. Decades of research 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 diffuclty principle (Bjork, 1994). Notable examples of desirable difficulties include having participants generate information from word fragments instead of passively reading intact words (e.g., Slamecka & Graf, 1978), spacing out study sessions instead of massing them (e.g., Carpenter, 2017), and having participants engage in retrieval practice after studying instead of simply restudying the information (Kornell & Vaughn, 2016). Another simple strategy that has gained some attention is to make material more perceptually disfluent. This can be done by changing the material's perceptual characteristics (Diemand-Yaumen, Oppenheimer, & Vaughan, 2011; French et al., 2013). Visual material that is masked (Mulligan, 1996), inverted (Sungkhasette, Friedman, & Castel, 2011), presented in an atypical font (Diemand Yaumen et al., 2011), blurred (Rosner, Davis, & Milliken, 2015), or even in handwritten cursive (Geller, Still, Dark, Carpenter, 2018) have all been shown to produce memory benefits. The desirable effect of perceptual disfluency on memory is called the disfluency effect (Bjork, 2016)

Although appealing as a pedagogical strategy due to the relative ease of implementation, there have been several experiments that failed to find memorial benefits for perceptually disfluent materials (e.g., Magreehan, Serra, Schwartz & Narciss, 2016; Rhodes & Castel, 2008, 2009; Rummer, Scheweppe, & Schewede, 2016; Yue, Castel, & Bjork, 2013), casting doubt upon the robustness of the disfluency effect. Corrobroating this, A recent meta-analysis by Xie, Zhou, and Liu (2018) pooled across 25 studies and over 3,000 particapnts (Xie, Zhou, & Liu, 20018) and found a small, nonsignificant, effect of perceptual disfluency on recall and (*d* = -0.01) and transfer (*d* = 0.03). Despite having no mnnmemonic effect,perceptual disfluency can produce longer reading times (*d* = 0.52) and produce lower judgments of learning (*d* = -0.043). Experimentally, Geller et al.(2018) and Geller & Still (2018) manpiulated several boundary conditions (e.g., level of degradation, type of judgement of learning, retentional interval, and testing expectany) and found you can get mnnmeonic benefits from perceptual disflunet mateirals, but it is rather fickle and not robust. Taken together, the evidence suggests that utility of perceptual disfluency is rather limited.

Despite the weak evidence, perceptual disfluency is still being touted as a viable learning tool, especially in the popular press. Recently, reputable

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

Correspondence concerning this article should be addressed to Jason Geller, Postal address. E-mail: jason-geller@uiowa.edu

news soruces like Washington Post (https://www.washingtonpost.com/business/2018/10/05/introducing-sans-forgetica-font-designed-boost-your-memory/) and NPR (https://www.npr.org/2018/10/06/655121384/sans-forgetica-a-font-to-remember claimed that a new font called Sans Forgetica can enhance memory. Since those articles, the SF font is available on all operating systems (all you have to do is downlaod the font file), some browsers (e.g., Chrome), and as a phone application. With this much attention and marketing, there has to be solid empirical evidence backing it up, right? Not quite.

What do we know about SF?

There is not a lot information on SF. The typyface itself is a variation of a sans-serif typeface. SF is a typeface that consists of intermitten gaps in letters that are back slanted (see below picture). The design features of this typeface require readers of it to “fill-in” the missing pieces like a puzzle. As it pertains to the empirical validation of the claims made, the website does offer some information about SF and how the original results were obtained, but not enough information to replicate the studies.

Accorind to an interview conducted by Earp (2018), In the first experiment (*N*=96), they had participants read 20 word pairs (e.g., girl - guy) in three new fonts (one of them being SF) and a typical or common font. The font pairs were presented in was counterbalanced participants. What this means is that all fonts were showns, 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 the second experiment ((*N* = 300) 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.

At the time of this writing, these studies have not been published nor is there a preprint available. I reached out to the creators of SF, but they refused to share the materials with me. Instead of waiting, I elicited the help of Sara Davis and Daniel Peterson at Skidmore university to test the mnenmomic benefits of Sans Forgetica.

Experiment 1

In the first study we compared the mnenmonic benefits of SF against a robust technique known to enhance memory—generation. The generation effect is a phenomenon where information is better remembered when retrieved than if it is simply read. In a prototypical experiment,participants are asked to generate words from word fragments DOLL - DR__ or read intact cue-target pairs (*DOLL-DRESS*). Compared to the intact condition, individuals recall the generated target words at a higher rate. The nature of generation is where the supposed mnnmeoic benefit of SF comes from. We examined this in the current experiment.

Participants

We recruited 230 people from Amazon’s Mechanical Turk Service. 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 effect size (*d* = .35). We choose this effect size as our SESOI due in part to the small effect sizes seen in actaul classroom studies (Bulter et al., 2014). Therefore, assuming an alpha of .05 and a desired power of 90%, a sample size of 270 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 (aspredicted.org) for Experiment 1 can be found here. All materials, data, and analysis scirpts can be found here (https://osf.io/d2vy8/). The results contained herein are computationally reproducible by going to the primary author’s github and clicking on the binder button (https://github.com/jgeller112/SF_Expt1; https://github.com/jgeller112/SF_Expt2).

Participants were presented with 22 weakly related cue-target pairs taken from Carpenter et al., 2012)[^1]: Two cue-target pairs () had to be thrown out as they were not preseted due to a coding error. The cue-target pairs were all nouns, 5–7 letters and 1–3 syllables in length, and high in concreteness (400–700) and frequency (at least 30 per million).

Procedure and Design

The experiment began with the presentation of 22 word pairs, shown one at a time, for 2 secconds 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 distractor task (anagram generation). Finally participants completed a cued recall test. During cued recall, particpants were presented 24 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. After they were asked several demographic questions.

We used a 2 x 2 mixed design. The within-subjects factor (Disfluency: fluent vs. disfluency) was manipulated across items and participants. The between-subjects factor (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 distractor task (anagram generation), they completed a cued recall test. During cued recall, particpants were presented 24 cues one at a time and asked to provide the target word. After they were thanked and debriefed.

Spell checking was automated with the hunspell package in R (Ooms, 2018) using spellCheck.R. At the next step we manually examined the output to catch incorrect suggestions and to add their own corrections. Becasuse participants were recruited in the United States, we used the American English dictionary. A nice walkthrough on how to use the package can be found in Buchcamam, De Deyne, and Montefinese (2019). Using the package, each response was corrected for misspelings. Corrected spellings are provided in the most probable order, therefore, the first suggestion is selected as the correct answer. Answers were marked correct if they provided the exact response. In order for a response to be judged correctly, the response had to match the correct answer.

Results

Scoring

Accuracy was automated with the hunspell package in R (Ooms, 2018) using spellCheck.R.A A nice walkthrough on how to use the package can be found in Buchcamam, De Deyne, and Montefinese (2019). Becasuse participants were recruited in the United States, we used the American English dictionary. Eeach participant response was corrected for misspelings. In the package, corrected spellings are provided in the most probable order, therefore, the first suggestion is selected as the correct answer. As a second pass, we went throigh and made sure the program slected the correct spelling. If the response was close to the correct response, it was marked as correct.

Results

Accuracy on the cued recall test was examined using a logistical mixed model (logit link) in R (R studio, 2019) using the lme4 package (Bates, Machler, Bolker, and Walker, 2015) with disfluency and diffilcuty type as a fixed effect and random intercepts for subjects (N=233) and target type (N =22) and random slopes for the factor of disfluency by participant and target: full_model=glmer(acc~condition*dis + (1+ dis|ResponseID) + (1+dis|target), data=sfgen1, contrasts = list(dis="contr.sum", condition="contr.sum"), family="binomial", control=glmerControl(optimizer="bobyqa",optCtrl=list(maxfun=100000))). This was the most complext model we could get to converge (Barr, Levy, Scheepers, & Tily, 2013). condition and disflunecy were sum coded (1, -1).

In Experiment 1 there was no effect of difficulty type, *Estimate* = -0.043, *exp(b)* = .961, *SE* .102, *Z* = -.430, *p* = .667, **d* =. There was an effect of disfluency, *Estimate* = 0.224, *exp(b)* = 1.251, *SE* = .062, *Z* = 3.622, *p* < .001, *d* = .654. Crucially, there was a significant interacion between difficulty type and disfluency, *Estimate* = 0.249, *exp(b)* = 1.28, *SE* = .041, *Z* = 6.098, *p* < .001, *d* = .67. This reflected a sizeable generation effect, but no SF effect (See figure below). Although not specified in the preregistration, a Bayes factor (BF) using weakly informative default priors for the estimates (Gelamn, Jakulin, Grazia,Pittaum & Sung Su, 2008) derived from the the full model using brms () and bayestestR indicated more support for a model with the interaction over a model without the interaction (BF = 9.19).

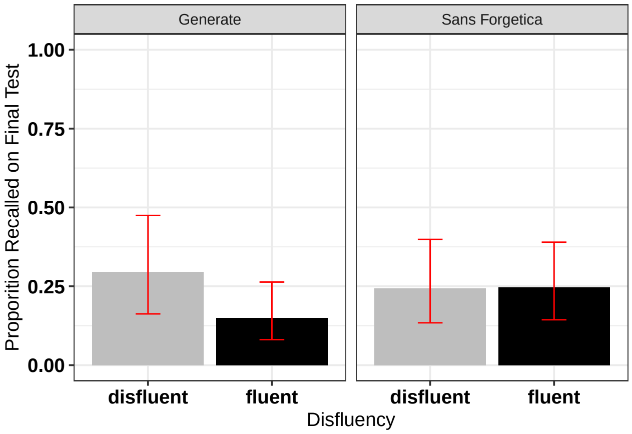
```
library(qualtRics)
library(tidyverse)
library(effects)
library(here)
library(lme4)
library(ggpol)
library(knitr)
library(here)
library(report)

sfgen=read_csv(here("Expt1_data", "sfgenerate_fina
```

```
## Warning: Missing column names filled in: 'X1' [
full_model=glmer(acc~condition*dis + (1+ di                                   data=sfge


paste(report(full_model))
```

```
## [1] "We fitted a logistic mixed model (e                                   to predict
```



```
ef1 <- effect("condition:dis", full_model) #take final glmer model
summary(ef1)
```

```
##
##  condition*dis effect
##                dis
## condition         disfluent    fluent
##   Generate        0.2952348 0.1507106
##   Sans Forgetica 0.2429061 0.2469149
##
##  Lower 95 Percent Confidence Limits
##                dis
## condition         disfluent     fluent
##   Generate        0.1625725 0.08088777
##   Sans Forgetica 0.1343397 0.14399034
##
##  Upper 95 Percent Confidence Limits
##                dis
## condition         disfluent    fluent
##   Generate        0.4747772 0.2635241
##   Sans Forgetica 0.3987912 0.3898998
```

```
x1 <- as.data.frame(ef1)

bold <- element_text(face = "bold", color = "black", size = 14) #axis bold

p<- ggplot(x1, aes(dis, fit, fill=dis))+ facet_grid(~condition)+
  geom_bar(stat="identity", position="dodge") +
  geom_errorbar(aes(ymin=lower, ymax=upper), width=0.2, position=position_dodge(width=0.9),co
  theme(legend.position = "none") +
  scale_fill_manual(values=c("grey", "black")) + ggplot2::coord_cartesian(ylim = c(0, 1)) + th

p
```

Experiment 2

The procedure in Experiment 1 could be argued to lack educational realsim. To test the effects of sans forgetica in a more relaistic situation, Experiment 2 presented participants a passage on ground water where some of the material was either: pre-highlighted, presented in SF, or presneted normally. This was a between-subjects manipulation.

Participants

Participants were 528 undergraduates who participated for partial completion of course credit. Sample size was calculated based on the samllest effect of interest (Lakens & Evers, 2014). In this case, 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 is required to detect whether an effect size of .35 differs from zero. After excluding participants based on our preregistered exclusion critera, we were left with unequal group sizes. Becasue of this, we ran six more pariticpants per group, giving us 176 participants in each of the three conditions.

Materials

The preregistatiron (aspredicted.org) for Experiment 2 can be found here. All materials, data, and analysis sciprts can be found here (https://osf.io/d2vy8/). The results contained herein are computationally reproducible by going to the primary author's github and clicking on the binder button (https://github.com/jgeller112/SF_Expt1; https://github.com/jgeller112/SF_Expt2).

Participants read a passage on ground water (856 words) taken from from the U.S. Geological Survey (see Yue et al.)

Eleven critical phrases[1] each containing a different keyword, were selected from the passage (e.g., the term *recharge* was the keyword in the phrase: Water seeping down from the land surface adds to the ground water and is called recharge water.) and were either presented in SF, highlighted, or unchanged. 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).

Design and Procedure

Participants were randomly assigned to either the pre-highlighted codnition, sans forgetica condition, or normal condition. Our design employed three between-subject variables: pre-highlighting, sans forgetica, and normal.

Participants completed the experiment on-line via the qualtrics survey platform. Participant read the passage on ground water in its entirety. Participants were given 10 minutes to read the passage. Participants in the pre-highlighted condition received some of the passages in yellow highlighting. Participants in the sans forgetcia codnition were presnetd some of the sentences in the sans forgetica font. Participants in the normal passage condition were presented sentences with no changes. All particiapnts were instructed to read the passage as though they were studying material for a class.

After 10 minutes, all participants were given a brief questionnaire (2 questions) asking them to indicate their metacognitive beliefs afte reading the passage. The two questions were: "Do you feel that the presentation fo the material helped you remember" and "How likely is it that you will be able to recall material from the passage you just read on a scale of 0 (not likely to recall) to 100 (likely to recall) in 5 minutes?" Participants were then given a short distractor task (anagrams) for 3 minutes. Finally, all participants were given 11 fill-in-the-blank test questions, one at a time. There was 1 manipulation multiple choice questions: What was the passage you read on?."

Results

Accuracy on the fill-in-the-blank test was examined using a logistical mixed model (logit link) in R (R studio, 2019) using the lme4 package (Bates, Machler, Bolker, and Walker, 2015) with passage type as a fixed effect and random intercepts for subjects (*n*=528) and questions (*n*=11): acc=glmer(auto_acc~passage_type+(1|ResponseId) + (1|Question), data=data, family="binomial"). Passage type was treatment coded thus estimates represent simple effects. 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 Figure 2). Results indicated that pre-highlighted sentences were better remembered than sentences presented normally, *Estimate* = .381, *exp(B)* = 1.46, *SE* = .167, *z* = 2.281, *p* = .023 *d* = .81 [^3: odds ratios were converted to d by dividing the ln(OR) by 1.81 (Chinn, 2000)] and were marginally better remmebered than sentences presented in sans forgetcia,*Estimate* = -.317, *exp(B)* = 1.37, *SE* = .168, *z* = -1.89, *p* = .059, *d* = .76. Critically, there was no difference between sentences presented normally and in sans forgetcia, *Estimate* = .065, *exp(B)* = 1.07, *SE* = .167, *z* = 0.386, *p* = 0.700, *d* =.04. A Bayes factor using the brms package (Burkner, 2015) was computed for no difference found that probability of this effect being zero was 12.72 to 1.

```
library(qualtRics)
library(tidyverse)
library(afex)
library(emmeans)
library(here)
library(ggpol)
library(knitr)

ground <- qualtRics::read_survey(here("Expt2_Data"

ground_change <- ground %>%
  mutate(Passage=ifelse(FL_149_DO=="Highlight", "P

#data was collected until the last day of the fall
# loading needed libraries
full_model=glmer(auto_acc~Passage+(1|ResponseId) +
#fit full model

paste(report(full_model))
```

```
## [1] "We fitted a logistic mixed model (estimate
```

```
ef1 <- effect("Passage", full_model) #take final g
summary(ef1)
```

```
##
##  Passage effect
## Passage
##          Normal Pre-highlighted  Sans Forgetica
```

[1] orginally 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

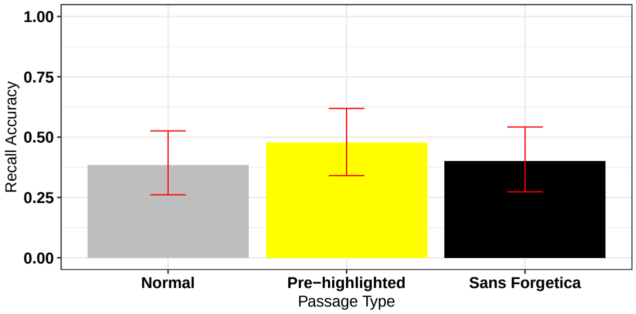
```
##        0.3847685        0.4779490        0.4001575
##
##  Lower 95 Percent Confidence Limits
## Passage
##            Normal Pre-highlighted  Sans Forgetica
##        0.2608280        0.3405702        0.2733269
##
##  Upper 95 Percent Confidence Limits
## Passage
##            Normal Pre-highlighted  Sans Forgetica
##        0.5257163        0.6187467        0.5419488
```

```
x1 <- as.data.frame(ef1)

bold <- element_text(face = "bold", color = "black", size = 14) #axis bold
p<- ggplot(x1, aes(Passage, fit, fill=Passage))+
  geom_bar(stat="identity", position="dodge") +
  geom_errorbar(aes(ymin=lower, ymax=upper), width=0.2, position=position_dodge(width=0.9),col
  scale_fill_manual(values=c("grey", "yellow", "black"))+
  theme(axis.text=bold, legend.position = "none") + ggplot2::coord_cartesian(ylim = c(0, 1))

p
```

\begin{figure}



{

}

\caption{Accuracy on Cued Recall Test. Error bars are 95% CI dervied from the GLMER model.} \end{figure}

Exploratory Analysis

In Experiment 2 we also asked students about their metacognitive awarness. Specifically we asked them: "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?" Initials analyses suggest that the normal 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 presneted normally, *t*(525) = 3.56, *p* = .406.

Table 1

contrast	estimate	SE	df	t.ratio	p.value
Highlight - Normal	-7.08	2.78	525.00	-2.55	0.03
Highlight - Passage	-3.52	2.78	525.00	-1.27	0.42
Normal - Passage	3.56	2.78	525.00	1.28	0.41

One potential reason for pre-highlighted information recieving lower JOLs than the normal passage is that pre-highlighted information served to focus participants attention specific parts of the passage. Given the question, pariticpants might thought this would hinder them if tested over the passage as a whole. Future research should

contrast	estimate	SE	df	t.ratio	p.va
Highlight - Normal	-7.079546	2.7792	525	-2.547332	0.02991
Highlight - Passage	-3.517046	2.7792	525	-1.265488	0.41539
Normal - Passage	3.562500	2.7792	525	1.281844	0.40605

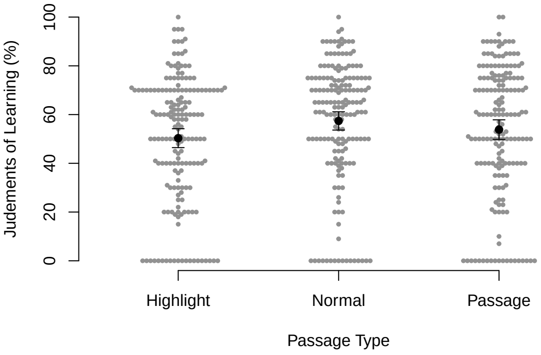


Figure 1. Judgements of learning as a function of passage type.

NULL

We hypothezied that sentences pre-highlighted or presented in sans forgetica would be better remembered than sentences presented normally. Further, we predicted that there would be no recall differences between the pre-highligted and the sans forgetica conditions. Our hypothese were only partially confirmed. We found that infromation that was pre-hightlighted had better recall than passages presentened normally, *Estimate* = -.328, *SE* = .166, *z* = -1.97, *p* = .048. Sentences that were pre-highlighted were also remembered marginally better than senetnces presented in sans forgetica,*Estimate* = -.307, *SE* = .167, *z* = -1.84, *p* = .066.

Looking at Bayes Factor for this comparison suggests that evidence for a difference between the two conditions is faily weak. Critically, sentences presented in sans forgetcia were not better remembered than sentences presented normally, *Estimate* = -.328, *SE* = .166, *z* = -1.97, *p* = .048, *BF*=).

Dicussion

Across two experiment The evidence contained herein suggests that SF does not have the mnemonic effects pruported by its creators. Now it is possible that there is an effect of SF, but the effect size might be smaller than we could detect acorss our two studies. Our SESOI was d = .35. If so, it probably does not have any real educational benefit. It is our conclsuion that SF is really forgetable and you should not be using it as a way to boost leanring.

References