Is Sans Forgetica Really Unforgettable?

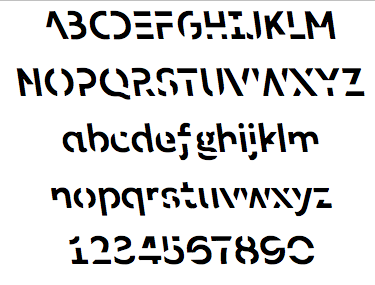
The primary goal of most humans is to remember the information they lean. Decades of research have put forth the paradoxical idea that making learning harder (not easier) should have the desirable effect of improving long-term retention of material–called the desirable diffuclty principle (Bjork, 1994). 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 called the disfluency effect.

Although appealing as a pedagogical strategy, 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 veracity of the disfluency effect. Recent studies by Geller et al.(2018) and Geller & Still (2018) found that perceptual disfluency can have a beneficial effect on memory, but seems to be rather fickle, thus delimiting its educational usefuleness.

Given the weak evidence, it came as a surprise to me when a little over a year ago, I saw a font by the name of Sans Forgetica (SF) getting a ton of press coverage. The mnnmenomic benefits of this font, *based on cognitive psychology*, were being touted in reputable news sources 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>, amongst others. The creators even made the SF font available for mac and pc operating systems–all you have to do is downlaod the font file and you to can remember everything you read :). There is even a Chrome browser extension and cellphone application that allows users to place material in Sans Forgetica. 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 information about SF. The typyface itself is a variation of a sans-serif typeface. It 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.



image

Earp (2018) conducted an interview with the creators of SF and here is what we know. Apparently two studies were conducted. In a lab experiment (*N*=96), they had participants read 20 word pairs (e.g., girl - guy; called a paried associates task in cognitive parlance) in three new fonts (one of them being SF) 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.

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. We preregistered two studies (<https://osf.io/d2vy8/>). All materials, data, and analysis scripts can be found at that website. On my github, you can reproduce all analyses for the two reported experiments in an online container (<https://github.com/jgeller112/SF_Expt1>; <https://github.com/jgeller112/SF_Expt2>).

# 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 natural of generation is where the power of SF comes from, according to the authors. We put that to the test in the first experiment.

We preregistered a sample size of 230 participants (115 per group) to detect a small-to medium size interaction effect (d = .35) with 90% power. We choose this effect size as our SESOI due in part due the small effect sizes seen in actaul classroom studies (Bulter et al., 2014). We recruited 230 people from Amazon’s Mechanical Turk Service. We initially collected data from 120 people, and then excluded 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, or 5) reported seeing our stimuli before.

In Experiment 1 participants were presented with 40 weakly related cue-target pairs. The 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). For half of the participants, half of the targets were presented in SF while the other half were presented in Arial font; for the other half of participants, the 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.

In order for a response to be judged correctly, the response had to match the correct answer. Scoring was automated with the hunspell package in R (Ooms, 2018). Using the package, each response was corrected for misspelings. 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, & Montefinese (2019). The most probably response (the first answer) was chosen as the correct spelling. What did we find?

In Experiment 1 we found a sizeable generation effect, which replicated past work. However, we did not find a SF effect (See figure below)

library(qualtRics)  
library(tidyverse)

## ── Attaching packages ──────────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.2.1 ✓ purrr 0.3.3  
## ✓ tibble 2.1.3 ✓ dplyr 0.8.3  
## ✓ tidyr 1.0.0 ✓ stringr 1.4.0  
## ✓ readr 1.3.1 ✓ forcats 0.4.0

## ── Conflicts ─────────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(afex)

## Loading required package: lme4

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

## Registered S3 methods overwritten by 'car':  
## method from  
## influence.merMod lme4  
## cooks.distance.influence.merMod lme4  
## dfbeta.influence.merMod lme4  
## dfbetas.influence.merMod lme4

## \*\*\*\*\*\*\*\*\*\*\*\*  
## Welcome to afex. For support visit: http://afex.singmann.science/

## - Functions for ANOVAs: aov\_car(), aov\_ez(), and aov\_4()  
## - Methods for calculating p-values with mixed(): 'KR', 'S', 'LRT', and 'PB'  
## - 'afex\_aov' and 'mixed' objects can be passed to emmeans() for follow-up tests  
## - NEWS: library('emmeans') now needs to be called explicitly!  
## - Get and set global package options with: afex\_options()  
## - Set orthogonal sum-to-zero contrasts globally: set\_sum\_contrasts()  
## - For example analyses see: browseVignettes("afex")  
## \*\*\*\*\*\*\*\*\*\*\*\*

##   
## Attaching package: 'afex'

## The following object is masked from 'package:lme4':  
##   
## lmer

library(emmeans)

## Welcome to emmeans.  
## NOTE -- Important change from versions <= 1.41:  
## Indicator predictors are now treated as 2-level factors by default.  
## To revert to old behavior, use emm\_options(cov.keep = character(0))

library(here)

## here() starts at /Users/gellr/SF\_Expt2

library(ggpol)  
library(knitr)  
sfgen1<-read\_csv("expt1agg1.csv")

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

## Parsed with column specification:  
## cols(  
## X1 = col\_double(),  
## ResponseID = col\_character(),  
## condition = col\_character(),  
## dis = col\_character(),  
## accuracy = col\_double()  
## )

## get aggreagte recall per subject, condition, and dis  
  
sfgenagg <- sfgen1 %>%   
 dplyr::group\_by(ResponseID, condition, dis) %>%  
 dplyr::summarise(accuracy=mean(accuracy))  
  
#perform mixed ANOVA  
a1 <- aov\_ez("ResponseID", "accuracy", sfgenagg,   
 within = c("dis"), between = c("condition")) # 2 X 2 Mixed ANOVA

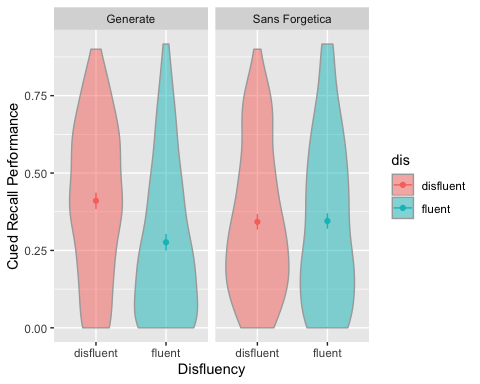
## Converting to factor: condition

## Contrasts set to contr.sum for the following variables: condition

#plot the results  
  
kable(nice(a1))

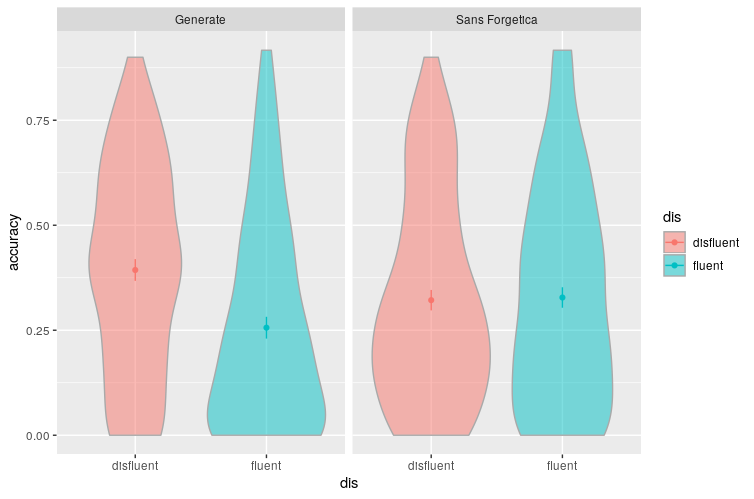
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Effect | df | MSE | F | ges | p.value |
| condition | 1, 230 | 0.10 | 0.00 | <.0001 | .99 |
| dis | 1, 230 | 0.02 | 25.44 \*\*\* | .02 | <.0001 |
| condition:dis | 1, 230 | 0.02 | 27.48 \*\*\* | .02 | <.0001 |

af1=afex\_plot(a1, x = "dis", panel = "condition",   
 error = "within", data\_geom=ggplot2::geom\_violin, mapping = c("color", "fill"))  
  
af2=af1+  
 theme\_set(theme\_light(base\_size = 14)) + labs(x="Disfluency", y="Cued Recall Performance")  
  
print(af2)



ls1 <- emmeans(a1, c("dis"), by="condition") # get the simple effects test for signifcant interaction.   
  
#pairwise comparison  
flex1=pairs(ls1)  
  
kable(flex1)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| contrast | condition | estimate | SE | df | t.ratio | p.value |
| disfluent - fluent | Generate | 0.1339080 | 0.0184121 | 230 | 7.2728251 | 0.0000000 |
| disfluent - fluent | Sans Forgetica | -0.0025797 | 0.0184121 | 230 | -0.1401076 | 0.8886976 |



plot

# Experiment 2

Experiment 2 examined the SF in a more educatioanlly realistic scenairo. We presented participants a passage on ground water where some of the material was either: pre-highlighted, presented in SF, or presneted with no changes. This was a between-subjects manipulation. Specifically participants read a passage about ground water (856 words) from the U.S. Geological Survey website (Yue, Storm, Kornell, Bjork, 2014). Eleven critical phrases [[1]](#footnote-31) 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).

Particiapnst were recruited from the SONA pool. Data was collected data until the end of the Fall semester 2019 and we ended up with 545 Ps (181 in the Highlight condition, 182 in the Normal condition, and 182 in the Passage condition). We preregistered a sample size of 510 participants (170 per group) to detect a small-to medium effect (d = .35) with 90% power.

## [1] "/Users/gellr/SF\_Expt2"

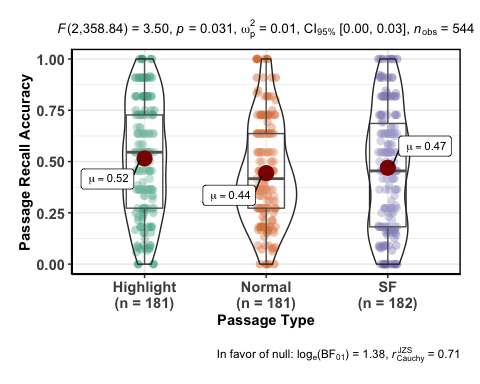
## Warning: Missing column names filled in: 'X1' [1]

## Warning: Duplicated column names deduplicated: 'X1' => 'X1\_1' [7]

Table 2.Pairwise Comparisons

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Effect | df | MSE | F | ges | p.value |
| Passage | 2, 541 | 0.07 | 3.27 \* | .01 | .04 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| contrast | estimate | SE | df | t.ratio | p.value | d |
| Highlight - Normal | 0.0714965 | 0.0282951 | 541 | 2.5268165 | 0.0315955 | 0.2656131 |
| Highlight - SF | 0.0451158 | 0.0282562 | 541 | 1.5966697 | 0.2479288 | 0.1676075 |
| Normal - SF | -0.0263807 | 0.0282562 | 541 | -0.9336249 | 0.6192373 | -0.0980056 |

 We found that infromation that was pre-hightlighted had better recall than passages presentened normally. We did not find that passages in SF were better remembered than normal or highligheted passages.

# Conclusions

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.

1. orginally we had 12 critical phrases but a pilot test after the pregreistation 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 [↑](#footnote-ref-31)