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

Funding Source: This research was supported by grant number 220020429 from the James S. McDonnell Foundation awarded to the third author.

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: drjasongeller@gmail.com

Abstract

Do students learn better with material that is perceptually hard to process? While evidence is equivocal on the matter, recent claims suggest that placing materials in Sans Forgetica, a perceptually difficult-to-process typeface, has positive impacts on student learning. Given the weak evidence for other similar perceptual disfluency effects, we examined the mnemonic effects of Sans Forgetica more closely in comparison to other learning strategies across three preregistered experiments. In Experiment 1 ($N = 233$), participants studied weakly related cue-target pairs with targets presented in either Sans Forgetica 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 an educational passage about ground water with select sentences presented in either Sans Forgetica typeface, yellow pre-highlighting, or unmodified. Cued recall for select words was better for pre-highlighted information than a unmodified pure reading condition. Critically, presenting sentences in Sans Forgetica did not elevate cued recall compared to an unmodified pure reading condition or a pre-highlighted condition. In Experiment 3 ($N = 60$), individuals 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, Recognition

Word count: 5708

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 academic performance at all levels. Many students are attracted to learning interventions that require minimal effort, but such approaches are rarely the best ways to achieve durable learning (Geller et al., 2018). Research in both the laboratory and classroom supports the paradoxical idea that making the encoding or retrieval of information more difficult, not easier, has the desirable effect of improving long-term retention (Bjork & Bjork, 2011). Notable examples of desirable difficulties include the generation effect (improved memory for information that has been actively generated rather than passively read; Bertsch, Pesta, Wiscott, & McDaniel, 2007; Slamecka & Graf, 1978), and the testing effect (improved memory for information that was actively recalled rather than passively re-read; see Kornell & Vaughn, 2016).

Research has conclusively established that when encoding is sufficiently effortful, memory outcomes improve. More recently, attention has focused on just how effortful that processing needs to be to observe mnemonic benefits. Specifically, researchers have examined whether subtle perceptual manipulations that change the physical characteristics of to-be-learned stimuli (rendering them harder to read or more disfluent) can similarly improve retention. Some research suggests it can. Diemand-Yauman, Oppenheimer, and Vaughan (2011), for instance, demonstrated that placing words in atypical typefaces (e.g., Comic Sans, Montype Corsiva) resulted in better memory than if the material was in a common typeface. A similar perceptual disfluency effect has been found with other perceptual manipulations including masking [e.g., presenting a stimulus very briefly (~100 ms) and forward or backing masking it; Mulligan (1996), inversion (Sungkhasettee, Friedman, & Castel, 2011), blurring (Rosner, Davis, & Milliken, 2015), and handwriting (Geller et al., 2018). The predominant theoretical explanation unifying these effects is that disfluency triggers metacognitive monitoring ("This is difficult to process"), which subsequently cues the learner to engage in deeper (i.e., more semantic) processing of material [but see Geller et al. (2018) for an alternative account).

Though this makes for a compelling story, other research casts serious doubt on disfluency as an effective pedagogical tool (e.g., Magreehan, Serra, Schwartz, & Narciss, 2016; Rhodes & Castel, 2008, p. @Rhodes2009; Rummer, Schweppe, & Schwede, 2016; Yue, Castel, & Bjork, 2013). A recent meta-analysis by Xie, Zhou, and Liu (2018) which included 25 studies and 3,135 participants found a small non-significant effect of perceptual disfluency on recall (d = -0.01) and transfer (d = 0.03). Most damning, despite having no mnemonic effect, perceptual disfluency manipulations generally produced longer reading times (d = 0.52) and lower judgments of learning (JOLs) (d = -0.43).

In trying to make sense of these disparate results, it is important to consider boundary conditions of the disfluency effect (see Geller & Still, 2018; Geller et al., 2018). As Geller et al. (2018) argue, not all perceptual disfluency manipulations are created equal. In Geller et al. (2018), they compared memory outcomes for information that was presented in print or handwritten cursive that was either easy or hard-to-read. Results showed both easy 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 disfluency manipulations can enhance memory, but such manipulations need to be optimally disfluent to exert a positive effect on memory.

Recently, a team of psychologists, graphic designers, and marketers set to create a new typeface specifically optimized to improve retention through perceptual disfluency. According to unpublished data from an interview taken from Earp (2018), to create the optimal typeface, the team conducted a cued recall experiment (N = 96) wherein participants read 20 related word pairs (e.g., *girl - guy*) each for 100 ms in a normal typeface (Albion) or one of three different disfluent typefaces (slightly disfluent, moderately disfluent, or extremely disfluent). They found an inverted U-shaped pattern wherein the moderately disfluent typeface was better remembered than the slightly and extremely disfluent typefaces. Further, they found that pairs in the moderately disfluent typeface were recalled slightly better than the normal typeface, although whether this meets the criteria for statistical significance is unclear. As a result of the memory boost, the team coined this moderately disfluent typeface Sans Forgetica. Sans Forgetica is a variant of sans serif typeface 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 (Slamecka & Graf, 1978).

To examine the effects of Sans Forgetica under more educationally realistic conditions, the Sans Forgetica team presented participants (N = 303) with 5 passages (~250 words in total) where one paragraph of three was presented in either the Sans Forgetica typeface or left in an unmodified (Arial) typeface (manipulated between subjects; Earp, 2018). For the Sans Forgetica condition, after each passage was read, participants were asked one question about the information written in the Sans Forgetica typeface and another question about the information written in Arial typeface. Performance was compared to the group that was presented with passages presneteed in a normal typeface. Placing text passages in the Sans Forgetica typeface resulted in better memory than if materials were presented in a normal typeface.

Since the release of the Sans Forgetica typeface, there has been a great deal of attention from the press. Sans Forgetica received coverage from major news sources like the Washington Post, National Public Radio, and The Guardian. In 2019, Sans Forgetica won the GoodDesign, Best in Class Award (Good Design, 2019). Commercially, Sans Forgetica typeface is freely available to users, and is marketed as a study tool Despite all the attention and marketing, empirical evidence for Sans Forgetica typeface is lacking.

Initial evidence for the Sans Forgetica typeface comes from the aforementioned unpublished studies by the Sans Forgetica development team (Earp, 2018). Note, however, that the modest group differences observed were not accompanied by any inferential hypothesis testing, nor were any of these claims subjected to the peer-review process. However, some evidence for the effectiveness of the Sans Forgetica typeface comes from a study by Eskenazi and Nix (2020), who found that words and definitions in Sans Forgetica typeface led to better orthographic discriminabity (i.e., choosing the correct spelling of a word) and semantic acquisition (i.e., retrieving the definition of a word), but only if participants were good spellers. This suggests that the utility of the Sans Forgetica typeface as a study tool may be quite limited. Recently, Taylor, Sanson, Burnell, Wade, and Garry (2020) found that although participants reported experiencing Sans Forgetica as disfluent (Experiment 1), there was no evidence that Sans Forgetica yielded a boost relative to Arial in cued recall (Expt 2) or for prose recall that tested factual (Experiment 3) and conceptual information (Experiment 4).

The Present Studies

The question of Sans Forgetica's effectiveness at producing a mnemonic benefit has clear educational implications. Demonstrating support for a purported study aid as quick and easy as swapping to-be-learned text from one typeface to another would be a boon for students. However, recent research calls into question the legitimacy of this typeface as a study aid at all (Taylor et al., 2020). Given the mixed evidence, the current set of studies aimed to further investigate the mnemonic benefit of the Sans Forgetica effect on memory.

In Experiment 1, we examined the impact of cue strength and study duration on cue-target pairs presented in a normal typeface or Sans Forgetica. In Experiment 2, we focused on more complex, educationally relevant prose passages. In Experiment 3, we examined if the type of test moderated the Sans Forgetica effect by using a yes/no recognition test. Importantly, we also compared the Sans Forgetica effect with other, more empirically supported learning techniques: generation (Experiment 1) and pre-highlighting (Experiment 2).

Experiment 1

In Experiment 1 we were interested in answering two questions. Does Sans Forgetica facilitate the retention of weakly associated cue-target pairs? If so, is this facilitation similar in magnitude to another desirable difficulty phenomenon—the generation effect? Taylor et al. (Experiment 2) used cue-target pairs that were highly associated and failed to find a memory benefit for Sans Forgetica typeface. It has been argued (e.g. Carpenter, 2009) that weakly related cue-target pairs produce more elaborative processing and lead to better memory, especially when the targets to be remembered require generation or retrieval (called the elaborative retrieval hypothesis). It is possible, then, that the use of highly associated pairs in Taylor et al. (2020) served to dampen the Sans Forgetica typeface effect. In Experiment 1 we examined the mnemonic benefit of Sans Forgetica typeface and generation by looking at cued recall performance with weakly associated cue-target pairs. In addition, we opted to present pairs for two 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.

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 (Experiments 1 and 2: https://osf.io/d2vy8/; Experiment 3: https://osf.io/dsxrc/). All raw and summary data, materials, and R scripts for preprocessing, analysis, and plotting can be found at https://osf.io/d2vy8/.

Participants. We recruited subjects on Amazon's Mechanical Turk (MTurk) platform, all of whom completed the study using Qualtrics survey software. A total of 232 people completed the experiment in return for U.S.$1.00. Sample size was based on a priori power analyses conducted using PANGEA v0.2 (Westfall (2016)). 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 between the generation effect and the Sans Forgettica effect (d = 0.35). 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 interaction effect size of 0.35 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 subjects. For the Sans Forgetica group, disfluent targets were presented in Sans Forgetica typeface while fluent targets were presented in Arial typeface. In the generation group, disfluent targets were presented in Arial typeface with missing letters (vowels replaced by underscores) and fluent targets were intact. See Figure 1 for an example of the stimuli used.

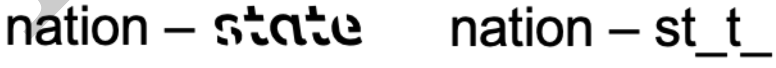

Figure 1. Example of cue-target pairs. Left: Sans Forgetica condition. Right: Generation condition.Sans Forgetica is licensed under the Creative Commons Attribution-NonCommercial License (CC BY-NC; https:// creativecommons.org/licenses/by-nc/3.0/)

Materials and Procedure. Participants were presented with 22 weakly related cue-target pairs taken from Carpenter et al. (2006). The pairs were all nouns, 5–7 letters and 1–3 syllables in length, high in concreteness (400–700), high in frequency (at least 30 per million), and had similar forward ($M = 0.031$) and backward ($M = 0.033$) 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 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 target when presented with the cue. The experiment began with the presentation of the word pairs, presented one at a time, for five seconds each. After a short two-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 type in 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). 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.

Analytic Strategy. For all the experiments, 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 each analysis. 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 et al., 2020) and BayesFactor packages (vers. 0.9.12-4.2; Morey & Rouder, 2018). All figures were generated using ggplot2 (vers. 3.3.0; Wickham, 2006).

Results and Discussion

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 fluency target words, $F(1, 230) = 25.31$, η_g^2 =.017, $p < .001$. This was qualified by an interaction between Fluency and Difficulty Type, $F(1, 230) = 25.06$, $\eta_g^2 = .017$, $p < .001$. A Bayesian ANOVA indicated strong evidence for the interaction model over the main effects model, $\text{BF}_{10} > 100$. As seen in Figure 2, the magnitude of the generation effect was larger than the Sans Forgetica effect, which was, in fact, negligible.

While the benefit of generating information was clear, there was no benefit of studying items in the Sans Forgetica typeface. Thus, presenting weakly associated targets in Sans Forgetica for two seconds produced no memory benefit. This null result was confirmed by a Bayesian analysis denoting strong evidence for the presence of the interaction compared to a main effects model. Although participants' 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, Pashler, and Vul (2006) using the same materials (30% overall for restudy compared to tested trials). In addition, accuracy was comparable to Experiment 2 from Taylor et al. (2020) who used highly associated cue target pairs (~45%). Finally, overall performance was strong enough to reveal a significant generation effect, minimizing concerns that a difficult task might be suppressing an

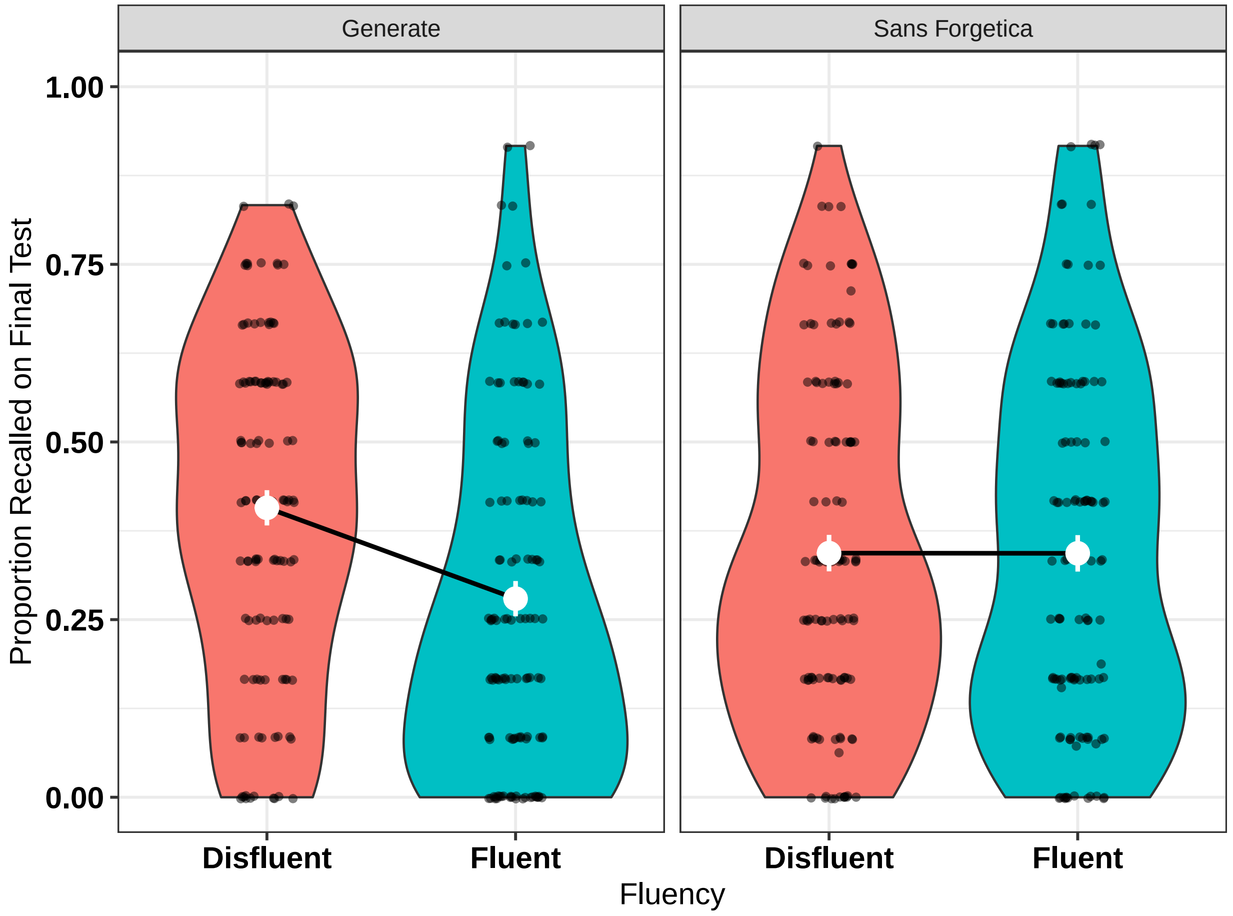

Figure 2. Accuracy on cued recall test. Violin plots represent the kernal density of avearge accuracy (black dots) with the mean (white dot) and Cousineau-Morey within-subject 95% CIs.

otherwise robust effect of Sans Forgettica . Taken together, these results suggest that (1) presenting materials in Sans Forgetica does not lead to better memory and (2) the effect of Sans Forgetica on memory is most likely not a desirable difficulty.

Experiment 2

Experiment 1 failed to reveal a memory benefit for the Sans Forgetica typeface. One potential limitation is that the atomistic stimuli employed in Experiment 1 (cue-target word pairs) may not provide an ecologically valid lens under which to study real classroom learning. To address this, in Experiment 2, we examined the mnemonic effects of Sans Forgetica using more complex prose materials. Like before, we wanted to compare the (potential) benefits of Sans Forgetica to something with empirical scrutiny. Accordingly, in Experiment 2, we examined how Sans Forgetica stacked up against pre-highlighting.

One of the purported functions of the Sans Forgetica typeface is to call attention to information one needs to remember. This is functionally similar to pre-highlighting, whereby important study information is highlighted prior to studying. Pre-highlighting is often used by instructors and textbook creators to enhance learning. Indeed, when students read pre-highlighted passages, there is some evidence that they recall more of the highlighted information and less of the non-highlighted information when 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 text. We hypothesized that both Sans Forgetica and pre-highlighting should enhance memory for selected passages compared to an unmodified passage.

Method

Participants. We preregistered a sample size of 510 (170 per group). When initial data collection finished, 683 participants participated for partial completion of course credit. After excluding participants based on our reregistered exclusion criteria (see above), we were left with unequal group sizes. Because of this, we ran six more participants per group, giving us 528 participants—176 participants in each of the three conditions.

Materials. Participants read a passage on ground water (856 words) taken from the U.S. Geological Survey (see Yue et al., 2014). Eleven critical phrases, 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 presented in yellow (pre-highlighted), Sans Forgetica typeface, or unmodified typeface. 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 one attention check question at the beginning of the final test.

Design and Procedure. Participants completed the experiment online via the Qualtrics survey platform. Participants were randomly assigned to one of three conditions: pre-highlighting, Sans Forgetica, or unmodified text. 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 question asking them to provide a judgement of learning after reading the passage: "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 three 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 preregistration, 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 pre-highlighted and Sans Forgetica sentences would be better remembered than normal sentences and that there would be no recall differences in recall between the highlighted and Sans Forgetica sentences. Our hypotheses were partially supported (see Figure 3). Examining our planned comparisons, we found that pre-highlighted sentences were better remembered than sentences presented in unformatted text, $t(525) = 2.45$, $SE = 0.028$, p_{tuk}

$= .039$, $d = 0.26$. There was weak evidence for no effect between sentences presented in Sans Forgetica and pre-highlighted, $t(525) = 0.049$, $SE = 0.028$, $p_{\text{tuk}} = .202$, $d = 0.18$, $\text{BF}_{01} = 2.36$. Critically, there was no difference between sentences presented normally or in Sans Forgetica, $t(525) = 0.02$, $SE = 0.028$, $p_{\text{tuk}} = .734$, $d = 0.079$, $\text{BF}_{01} = 6.47$.

In short, we did not find that select information presented in Sans Forgetica produced better memory than select information left unmodified or pre-highlighted. The finding that Sans Forgetica typeface does not enhance memory for prose passages replicates the findings of Taylor et al. (2020) (Experiments 3 and 4). We did, however, observe better memory for pre-highlighted information compared to words presented unmodified or in a Sans Forgetica typeface.

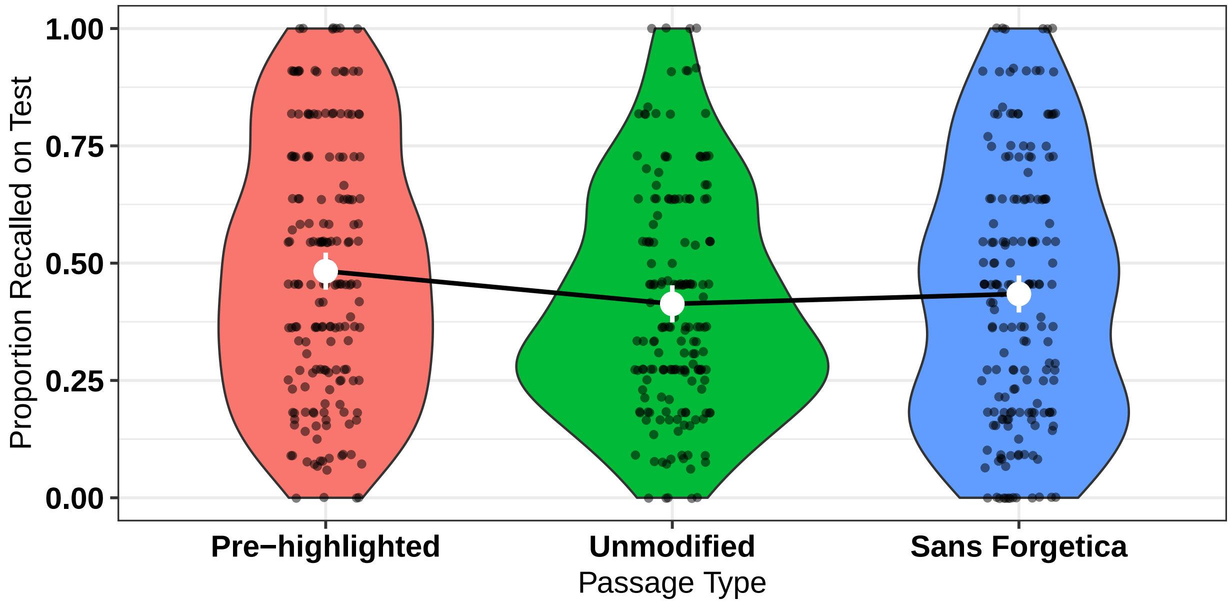

Figure 3. Proportion recalled as a function of passage type. Violin plots represent the kernal density of avearge accuracy (black dots) with the mean (white dot) and 95% CIs.

Exploratory Analysis. In Experiment 2 we also asked students about their metacognitive awareness of the manipulations known as JOLs. To examine differences, we conducted separate independent t-tests. Looking at JOLs (see Figure 4), the unmodified passage was given higher JOLs ($M = 57.4$, $SD = 25.2$) than the pre-highlighted passage

(M = 50.3, SD = 26.0), $t(525)$ = -2.55, SE = 2.78, p_{tuk} = .030. There were no reliable differences between the pre-highlighted and Sans Forgetica (M = 53.8, SD = 27.0) passages, $t(525)$ = -1.26,SE = 2.78, p_{tuk} = .415, or between the passage in Sans Forgetica and the passage presented normally, $t(525)$ = 1.28, SE = 2.78, p_{tuk} = .406. That is, passages in Sans Forgetica typeface did not produce lower judgement of learning compared to an unmodified or pre-highlighted passage. Interestingly, individuals gave lower JOLs to pre-highlighted information compared to materials presented in a normal typeface. 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). Despite this, we did find lower JOLs for pre-highlighted information compared to unmodified 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 to specific parts of the passage. Given the question (i.e., "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 might have thought pre-highlighting would hinder their ability to answer questions on the whole passage.

Experiment 3

Both Experiments 1 and 2 utilized cued recall for the final criterion test. In previous studies, perceptual disfluency has been shown to enhance performance on yes/no recognition tests, even when there is no recall benefit (Nairne, 1988). This is thought to be because the learner is focusing on surface-level aspects during the initial perceptual identification process. This strategy should aid later recognition, but not recall, for fluent items, given that recall relies more on item elaboration than on perceptually distinctive features. In Experiment 3, we tested whether Sans Forgetica would lead to similar benefits in recognition memory. It is possible that Sans Forgetica serves to increase surface-level familiarity of a word and thus recognition, while recall is unchanged. We hypothesized no

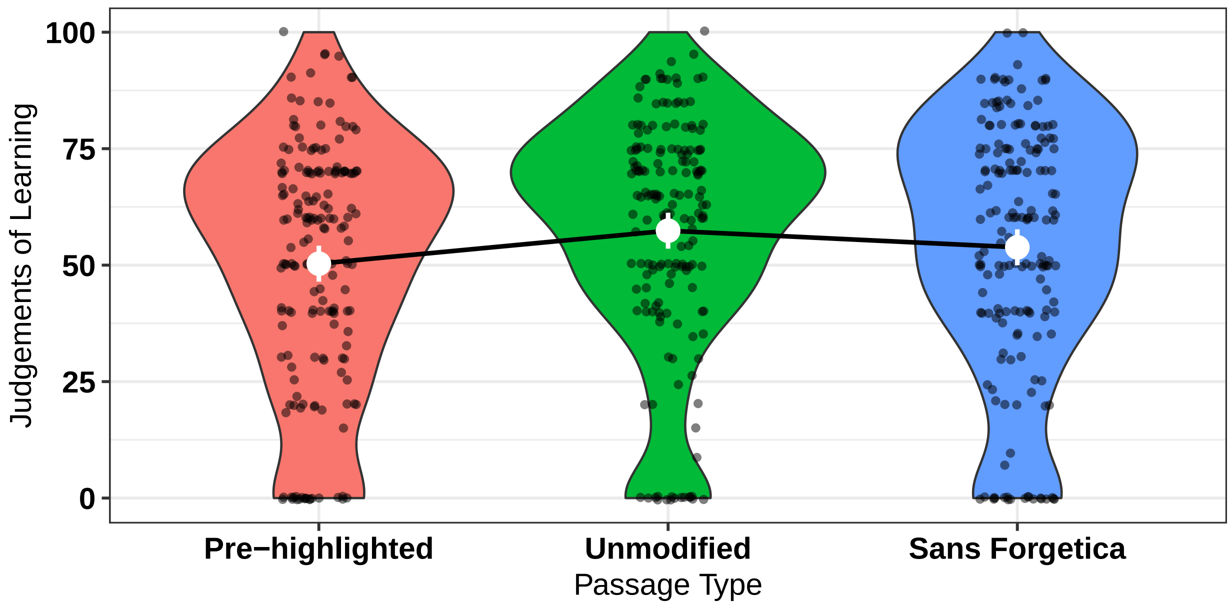

Figure 4. Judgements of learning as a function of passage type. Violin plots represent the kernal density of avearge accuracy (black dots) with the mean (white dot) and 95% CIs.

recognition memory benefit for Sans Forgetica.

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 were removed for failing to meet the exclusion criteria noted above.

Materials. Stimuli were 188 single-word nouns taken from Geller et al. (2018). All words were from the English Lexicon Project database (Balota et al., 2007). Both word frequency (all words were high frequency; mean log HAL frequency = 9.2) and length (all words were four letters) 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 variable, manipulated within-subjects. We presented participants with 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 typefaces. 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/openmaterials/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 they were at study.

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 (see the General Discussion for the study time data). After the study phase, participants completed a short three-minute distractor task wherein they wrote down as many U.S. state 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 typeface or Sans Forgetica typeface. 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.

Results and Discussion

Performance was examined with d', a memory sensitivity measure derived from signal detection theory (Macmillan & Creelman, 2005). Hits or false alarms at ceiling or floor were changed to .99 or .01. Hits and false alarms along with sensitivity (d') can be seen in Figure 5.

Consistent with our preregistered hypothesis, there was no difference in d' between Sans Forgetica and Arial typefaces, $t(59) = 0.281$, $SE = 0.05$, $p = .780$. There was strong evidence for no effect ($\mathrm{BF}_{01} = 13.68$).

Overall, we did not find an effect of Sans Forgetica typeface on recognition memory. This study provides further evidence that Sans Forgetica typeface is not desirable for memory, regardless of the final test format.

General Discussion

The creators of the Sans Forgetica typeface as well as the media have made strong claims regarding the mnemonic benefits of Sans Forgetica. The aim of the current experiments was to test those claims empirically. In Experiment 1, Sans Forgetica typeface did not enhance memory in a cued recall task with weakly related cues. In Experiment 2, Sans Forgetica typeface did not enhance memory for a complex prose passage. In Experiment 3, Sans Forgetica typeface did not enhance recognition memory. Even with every opportunity to reveal itself, we did not find any evidence for a mnemonic benefit of Sans Forgetica typeface.

While it has been posited both 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. This, along with Taylor et al. (2020) provides converging evidence that that Sans Forgetica typeface is not a desirable difficulty. Theoretically, these findings add to the literature showing that perceptual disfluency has little impact on actual memory performance (e.g., Magreehan et al., 2016; Rhodes & Castel, 2008, p. @Rhodes2009; Rummer et al., 2016; Xie et al., 2018; Yue et al., 2013). Importantly, we did find a memory advantage for other learning

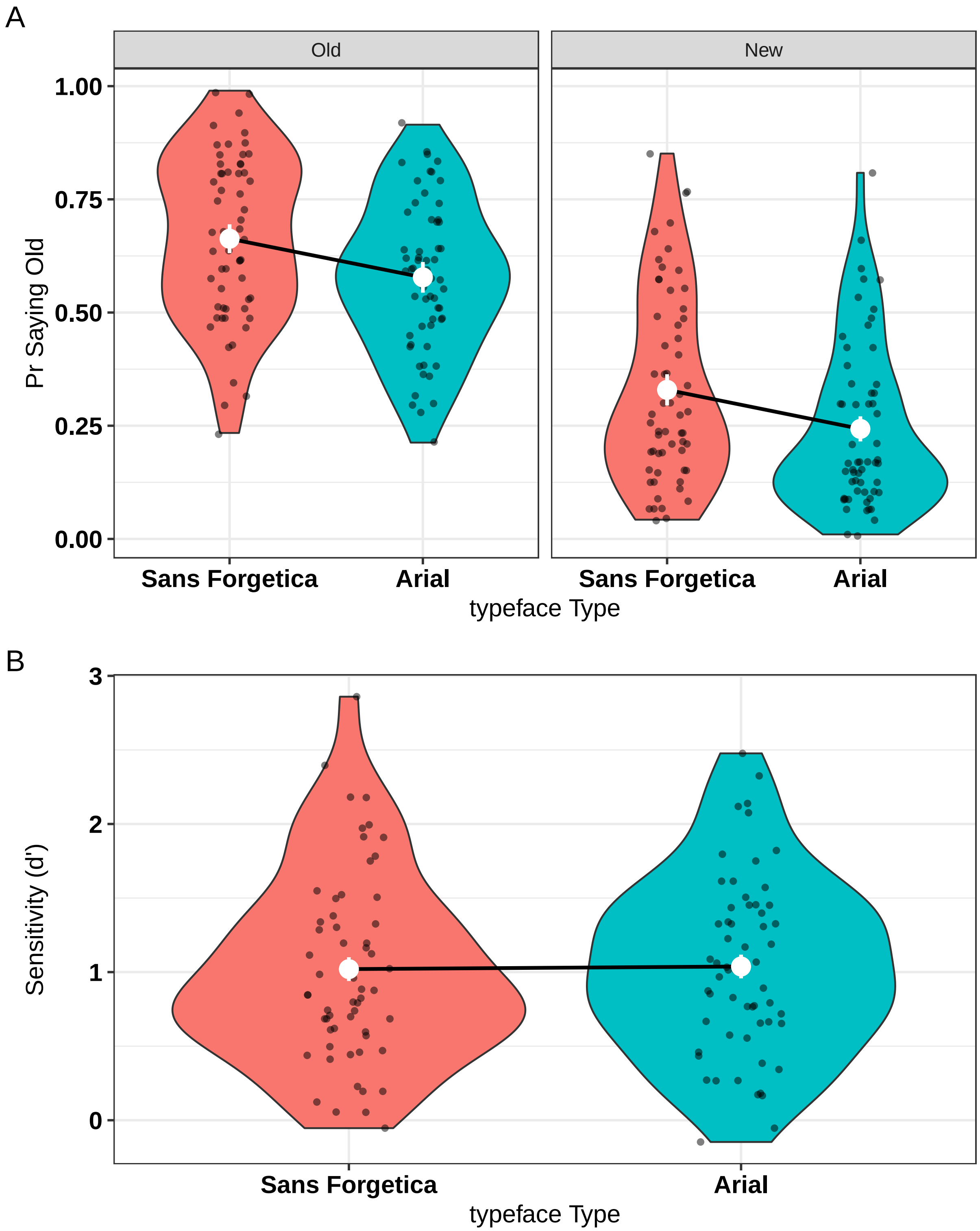

Figure 5. A. Mean proportions of "old" responses. Violin plots represent the kernal density of average probability (black dots) with the mean (white dot) and within-subject 95% CIs. B. Memory sensitivity (d'). Violin plots represent the kernal density of avearge accuracy (black dots) with the mean (white dot) and Cousineau-Morey within-subject 95% CIs.

techniques such as generation (Experiment 1) and pre-highlighting (Experiment 2) whose efficacy is robustly supported in the literature. That is, the conditions were ripe for a Sans Forgetica effect to emerge if it were an actual mnemonic effect.

What might account for the null effect of Sans Forgetica typeface on memory? Drawing valid conclusions about disfluency requires the use of objective disfluency measures. In many studies, perceptual disfluency is using subjective measures (e.g., JOLs or difficulty ratings), 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 typeface is simply not perceptually disfluent. Although we did not preregister explicit tests of objective disfluency, we have some preliminary evidence that Sans Forgetica is not disfluent. In Experiment 3, we collected self-paced study times for each stimulus. 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 typeface (M = 1481 ms, SD = 1750 ms) and Arial typeface (M = 1500 ms, SD = 2344 ms), $t(59) = 0.469$, $p = .641$, $\mathrm{BF}_{01} = 6.67$. The absence of a study time disparity suggests the typeface may not be considered disfluent and could therefore explain why we did not observe an effect of Sans Forgetica typeface. It is worth noting, however, that self-paced study times might reflect variables other than, or in addition to, processing disfluency. Although self-paced study provides one way of measuring processing fluency, more precise measures of processing disfluency should be considered as well (but see Eskenazi & Nix, 2020 for eye-tracking evidence for Sans Forgetica typeface in good spellers).

While the current set of experiments (see also Taylor et al., 2020) did not find a memory benefit for Sans Forgetica typeface, we cannot rule out that the effect might arise under different conditions. A number of boundary conditions that determine when perceptual disfluency will and will not be a desirable difficulty have been established over the past several years (see Geller et al., 2018, Geller & Still, 2018). In the current set of experiments, we examined whether cue strength, study time, 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 that the memory benefit from Sans Forgetica is moderated by these factors. Despite this, future research should examine the role of individual difference measures (e.g., working memory capacity; Lehmann, Goussios, & Seufert, 2016) along with other design features not tested (e.g. test delay, testing expectancy).

Conclusion

Students are attracted to learning interventions that are easy to implement (Geller et al., 2018). It is no surprise, then, that Sans Forgetica has garnered so much media attention. However, in our current age of uncertainty about the quality of information, it is important to properly evaluate scientific claims made by the media, even if that information comes from widely trusted news sources. As scientists, our job is to properly evaluate the evidence and correct erroneous information. Accordingly, we are compelled 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 technique to bolster learning. Our results suggest that placing material in Sans Forgetica typeface does not lead to more durable learning. It is our recommendation that students looking to remember more and forget less use learning tools such as testing or spacing that have stood the test of time.

Disclosures

Conflicts of Interest. The authors declare that they have no conflicts of interest with respect to the authorship or the publication of this article.

Author Contributions. JG wrote the first draft of the manuscript, collected data, and conducted all statistical analyses. JG, SD, and DP conceptualized the studies, reviewed, and edited the manuscript.

R and R package acknowledgements. The results were created using R (Version 4.0.0; R Core Team, 2019) and the R-packages *afex* (Version 0.27.2; Singmann, Bolker, Westfall, Aust, & Ben-Shachar, 2019), *BayesFactor* (Version 0.9.12.4.2; Morey & Rouder, 2018), *bit* (Version 1.1.15.2; Oehlschlägel, 2020, 2017), *bit64* (Version 0.9.7; Oehlschlägel, 2017), *carData* (Version 3.0.4; Fox, Weisberg, & Price, 2019), *coda* (Version 0.19.3; Plummer, Best, Cowles, & Vines, 2006), *data.table* (Version 1.12.8; Dowle & Srinivasan, 2019), *dplyr* (Version 0.8.5; Wickham et al., 2019), *effects* (Version 4.1.4; Fox & Weisberg, 2018; Fox, 2003; Fox & Hong, 2009), *emmeans* (Version 1.4.7; Lenth, 2020), *forcats* (Version 0.5.0; Wickham, 2019a), *ggplot2* (Version 3.3.0; Wickham, 2016), *ggpol* (Version 0.0.6; Tiedemann, 2019), *here* (Version 0.1; Müller, 2017), *janitor* (Version 2.0.1; Firke, 2020), *knitr* (Version 1.28; Xie, 2015), *lattice* (Version 0.20.41; Sarkar, 2008), *lme4* (Version 1.1.23; Bates, Mächler, Bolker, & Walker, 2015), *lubridate* (Version 1.7.8; Grolemund & Wickham, 2011), *Matrix* (Version 1.2.18; Bates & Maechler, 2019), *modelbased* (Version 0.1.2; Makowski, Lüdecke, & Ben-Shachar, 2020), *papaja* (Version 0.1.0.9942; Aust & Barth, 2020), *patchwork* (Version 1.0.0; Pedersen, 2019), *plyr* (Version 1.8.6; Wickham et al., 2019; Wickham, 2011), *purrr* (Version 0.3.4; Henry & Wickham, 2019), *qualtRics* (Version 3.1.3; Ginn & Silge, 2020), *readr* (Version 1.3.1; Wickham, Hester, & Francois, 2018), *Rmisc* (Version 1.5; Hope, 2013), *see* (Version 0.5.0; Lüdecke, Makowski, Waggoner, & Ben-Shachar, 2020), *stringr* (Version 1.4.0; Wickham, 2019b), *tibble* (Version 3.0.1; Müller & Wickham, 2019), *tidyr* (Version 1.1.0; Wickham & Henry, 2019), and *tidyverse* (Version 1.3.0; Wickham, 2017).

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

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown.* Retrieved from https://github.com/crsh/papaja

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The english lexicon project. Springer New York LLC. https://doi.org/10.3758/BF03193014

Bates, D., & Maechler, M. (2019). *Matrix: Sparse and dense matrix classes and methods.* Retrieved from https://CRAN.R-project.org/package=Matrix

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

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

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

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., 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

Dowle, M., & Srinivasan, A. (2019). *Data.table: Extension of ‘data.frame‘*. Retrieved from https://CRAN.R-project.org/package=data.table

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

Firke, S. (2020). *Janitor: Simple tools for examining and cleaning dirty data*. Retrieved from https://CRAN.R-project.org/package=janitor

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

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software*, *8*(15), 1–27. Retrieved from http://www.jstatsoft.org/v08/i15/

Fox, J., & Hong, J. (2009). Effect displays in R for multinomial and proportional-odds logit models: Extensions to the effects package. *Journal of Statistical Software*, *32*(1), 1–24. Retrieved from http://www.jstatsoft.org/v32/i01/

Fox, J., & Weisberg, S. (2018). Visualizing fit and lack of fit in complex regression models with predictor effect plots and partial residuals. *Journal of Statistical Software*, *87*(9), 1–27. https://doi.org/10.18637/jss.v087.i09

Fox, J., Weisberg, S., & Price, B. (2019). *CarData: Companion to applied regression data sets*. Retrieved from https://CRAN.R-project.org/package=carData

Geller, J., & Still, M. L. (2018). Testing expectancy, but not judgements of learning, moderate the disfluency effect. In J. Z. Chuck Kalish Martina Rau & T. Rogers (Eds.), *CogSci 2018* (pp. 1705–1710).

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

Ginn, J., & Silge, J. (2020). *QualtRics: Download 'qualtrics' survey data*. Retrieved from https://CRAN.R-project.org/package=qualtRics

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25. Retrieved from http://www.jstatsoft.org/v40/i03/

Henry, L., & Wickham, H. (2019). *Purrr: Functional programming tools*. Retrieved from https://CRAN.R-project.org/package=purrr

Hope, R. M. (2013). *Rmisc: Rmisc: Ryan miscellaneous*. Retrieved from https://CRAN.R-project.org/package=Rmisc

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

Lehmann, J., Goussios, C., & Seufert, T. (2016). Working memory capacity and disfluency effect: an aptitude-treatment-interaction study. *Metacognition and Learning*, *11*(1), 89–105. https://doi.org/10.1007/s11409-015-9149-z

Lenth, R. (2020). *Emmeans: Estimated marginal means, aka least-squares means.* Retrieved from https://github.com/rvlenth/emmeans

Lüdecke, D., Makowski, D., Waggoner, P., & Ben-Shachar, M. S. (2020). *See: Visualisation toolbox for 'easystats' and extra geoms, themes and color palettes for 'ggplot2'.* Retrieved from https://CRAN.R-project.org/package=see

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide, 2nd ed.* (pp. xix, 492–xix, 492). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

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

Makowski, D., Lüdecke, D., & Ben-Shachar, M. S. (2020). *Modelbased: Estimation of model-based predictions, contrasts and means.* Retrieved from https://CRAN.R-project.org/package=modelbased

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

Müller, K. (2017). *Here: A simpler way to find your files.* Retrieved from https://CRAN.R-project.org/package=here

Müller, K., & Wickham, H. (2019). *Tibble: Simple data frames.* Retrieved from https://CRAN.R-project.org/package=tibble

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

Oehlschlägel, J. (2017). *Bit64: A s3 class for vectors of 64bit integers.* Retrieved from https://CRAN.R-project.org/package=bit64

Oehlschlägel, J. (2020). *Bit: A class for vectors of 1-bit booleans.* Retrieved from https://CRAN.R-project.org/package=bit

Pedersen, T. L. (2019). *Patchwork: The composer of plots.* Retrieved from https://CRAN.R-project.org/package=patchwork

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for mcmc. *R News*, *6*(1), 7–11. Retrieved from https://journal.r-project.org/archive/

R Core Team. (2019). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

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

Sarkar, D. (2008). *Lattice: Multivariate data visualization with r*. New York: Springer. Retrieved from http://lmdvr.r-forge.r-project.org

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. (2019). *Afex: Analysis of factorial experiments*. Retrieved from https://CRAN.R-project.org/package=afex

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning & Memory*, *4*(6), 592–604. https://doi.org/10.1037/0278-7393.4.6.592

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

Tiedemann, F. (2019). *Ggpol: Visualizing social science data with 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=ggpol

Westfall, J. (2016). *PANGEA: Power ANalysis for GEneral Anova designs*. Retrieved from http://jakewestfall.org/pangea/

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*(1), 1–29. Retrieved from http://www.jstatsoft.org/v40/i01/

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

Wickham, H. (2017). *Tidyverse: Easily install and load the 'tidyverse'*. Retrieved from https://CRAN.R-project.org/package=tidyverse

Wickham, H. (2019a). *Forcats: Tools for working with categorical variables (factors)*. Retrieved from https://CRAN.R-project.org/package=forcats

Wickham, H. (2019b). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H., François, R., Henry, L., & Müller, K. (2019). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2019). *Tidyr: Tidy messy data*. Retrieved from https://CRAN.R-project.org/package=tidyr

Wickham, H., Hester, J., & Francois, R. (2018). *Readr: Read rectangular text data*. Retrieved from https://CRAN.R-project.org/package=readr

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

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from https://yihui.name/knitr/

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