

Surprise! Low Testing Expectancy Moderates the Sans Forgetica Effect

Jason Geller^{1,2} & Kelly A. Kane³

¹ University of Iowa

² Rutgers University Center for Cognitive Science

³ Glenville State College

Abstract

Recent work examining the mnemonic effects of Sans Forgetica has yielded discrepant findings. To clarify this discrepancy, the present experiments examined a boundary condition that determines when Sans Forgetica is and is not beneficial to learning. This boundary condition is knowledge about an upcoming test (high test expectancy) versus not (low test expectancy). This boundary condition was tested across two experiments. In Experiment 1 (pre-registered, $N = 231$), Sans Forgetica elicited lower judgements of learning and longer study times, but only improved memory on a yes/no recognition test when there was low test expectancy (compared to a high test expectancy group). In Experiment 2 ($N = 116$) using a cued recall test with low test expectancy, we found a similar pattern of results to Experiment 1. Taken together, Sans Forgetica can be a desirable difficulty, but only when testing expectancy is low. However, caution should be taken in interpreting these results. Not only was the effect size small, but low testing expectancy is not educationally realistic. Echoing previous sentiments, students wanting to remember more and forget less should stick to other desirable difficulties shown to enhance memory.

Keywords: Disfluency

Word count: 3500

Successful remembering is impacted by innumerable factors. Based off the influential desirable difficulty principle (Bjork & Bjork, 2011), one factor that has been purported

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, Rutgers University Center for Cognitive Science (RuCCS), 152 Frelinghuysen Road, Busch Campus, Piscataway, New Jersey 08854. E-mail: jason.geller@ruccs.rutgers.edu

to enhance remembering is the degree to which material to be learned is perceptual disfluent or not. Perceptual manipulations such as placing material in atypical fonts(???), blurring(Rosner et al., 2015), or in handwritten cursive (???) have been shown to yield better memory—called the disfluency effect. One such perceptual manipulation garnering increased attention is Sans Forgetica. Sans Forgetica is a typeface developed by a team of psychologists, graphic designers, and marketers, consisting of intermittent gaps and black-slanted letters (???). The disfluent perceptual characteristics of the typeface are purported to stave off forgetting and enhance learning. The claims surrounding Sans Forgetica have lead to extensive press coverage from major news outlets (NPR, Washington Post), and have lead to browser extensions and OS applications that allows users to place content in Sans Forgetica. As the famous astronomer Carl Sagan once said, "Extraordinary claims require extraordinary evidence (Sagan, 1980).

There is a growing body of evidence suggesting perceptual disfluency manipulations are simply not desirable for learning (see Xie et al., 2018 for a meta-analysis). Does the same hold true for Sans Forgetica? In two independent studies, Taylor et al. (2020) and Geller et al. (2020) set out to examine whether Sans Forgetica is *really* desirable for learning. In the first conceptual replication of the Sans Forgetica effect, Taylor et al. (2020) found (in a sample of 882 people across 4 experiments) that while Sans Forgetica was perceived as more disfluent by participants (Experiment 1) there was no evidence that Sans Forgetica yielded a mnemonic boost in cued recall with highly related word pairs (Experiment 2) compared to a fluent typeface (Arial) or when learning simple prose passages (Experiments 3-4). Extending these findings, Geller et al. (2020) conducted three pre-registered experiments (with over 800 participants), and found, similar to Taylor et al. (2020), Sans Forgetica does not enhance learning for weakly related word pairs (Experiment 1), a complex prose passage on ground water (Experiment 2), or when the type of test was changed to a recognition memory test (Experiment 3). Taken together, across two independent replication attempts, and over a 1000 participants, there is weak evidence for Sans Forgetica as a desirable difficulty.

Despite these findings, some evidence for the effectiveness of the Sans Forgetica typeface does exist. For instance, Eskenazi and Nix (2020) found that Sans Forgetica can enhance learning. In their study, they had participants learn the spelling and meaning for 15 low-frequency words each presented in the context of two sentences while their eye movements were monitored. During the test phase, orthographic discriminability (i.e., choosing the correct spelling of a word) and semantic acquisition (i.e., retrieving the definition of a word) were assessed. The authors reported a memory benefit for both orthographic discriminability and semantics for words presented in Sans Forgetica compared to a normal (Courier) typeface, but only for participants that were good spellers.

The mixed findings reported above suggest mnemonic benefit of Sans Forgetica may be fickle, with positive effects potentially bounded by specific conditions. Probing into the design features of Eskenazi and Nix (2020), a critical difference between their study and (???) and Geller et al. (2020) is testing expectancy. Eskenazi and Nix (2020) did not tell participants about the upcoming orthographic and semantic tests. Thus, one common design feature that may moderate whether we see a Sans Forgetica effect is high testing expectancy.

It is well known that testing expectancy can positively influence memory. Expecting a test of any kind can lead to enhanced processing of studied material, by either reducing learners' mind-wandering during studying (Szpunar et al., 2007) or by reducing interference from previously studied information (Weinstein et al., 2014). In the context of perceptual disfluency effects, Eitel and Köhl (2016) reasoned that if the disfluency effect arises because of deeper, more effortful, processing, telling participants about a memory test should eliminate the effect. This occurs because testing expectancy counteracts the effects of perceptual disfluency by eliciting enhanced processing for both fluent and disfluent stimuli. In contrast, low testing expectancy is less likely to impact processing of individual items, leaving effects of processing difficulty intact. While Eitel and Köhl (2016) found evidence for a general testing expectancy effect (better memory for high vs. low testing expectancy) they not find evidence for a moderated disfluency effect. However, Geller and Still (2018), following up on this, demonstrated in a yes/no recognition memory test that the disfluency effect only occurred under low testing expectancy. Given this, it is possible, then, that Sans Forgetica (a disfluent font) might arise when participants have low test expectancy.

Experiment 1

In Experiment 1 we examined whether the positive effects of Sans Forgetica are moderated by testing expectancy. Using a yes/no recognition memory test, we manipulated testing expectancy by telling half the participants about the upcoming memory test while for the other half being surreptitious about the upcoming memory test. In addition, we collected list-wide judgments of learning (i.e., a subjective memory prediction about future memory performance taken after all items are studied) and study times as a manipulation check to ensure Sans Forgetica is perceptually disfluent. We preregistered that we would observe an interaction between typeface (Arial vs. Sans Forgetica) and Test Expectancy. Specifically, if participants were not told about a memory test (low test expectancy) we would see a memory boost for Sans Forgetica stimuli, but not if they were told about a memory test. For JOLs, we predicted that we would not see JOL differences as function of typeface or testing expectancy. In terms of reading times, we predicted we would see longer study times for Sans Forgetica, but only in the low testing expectancy condition. These predictions are based on Geller et al. (2020) (Experiments 2 and 3). # Method

The preregistered analysis plan for Experiment 1 can be found here: <https://osf.io/wgp9d>. All raw and summary data, materials, and R scripts for pre-processing, analysis, and plotting can be found at <https://osf.io/d2vy8/>.

Participants

We preregistered a sample size of 230. All participants were recruited through prolific (prolific.co), and completed the study on the Gorilla platform [www.gorilla.sc; Anwyl-Irvine 2020]. The sample size was based off a previous experiment (Geller et al. (2020), Experiment 1), wherein they calculated power to detect a medium sized interaction effect ($d = 0.35$) using a similar design to the current study. After data collection had ended

94 we had a total of 231 participants. Participants completed the experiment in return for
95 U.S.\$8.00 an hour.

96 *Materials*

97 Stimuli were 188 single-word nouns taken from Geller et al. (2018). All words were
98 from the English Lexicon Project database (Balota et al., 2007). Both word frequency (all
99 words were high frequency; mean log HAL frequency = 9.2) and length (all words were four
100 letters) were controlled. The full set of stimuli can be found at <https://osf.io/dsxrc/>.

101 *Design*

102 Per our pre-registration, d' , JOLs, and study times were analyzed with a 2 (Type-
103 face: Arial vs. Sans Forgetica) \times 2 (Testing Expectancy: High vs. Low) mixed analysis of
104 variance (ANOVA).

105 *Procedure*

106 Similar to Geller et al. (2020) (Experiment 3), four lists (94 words each; 47 in each
107 typeface condition) were used to create the stimuli for a total of 188 words. Ninety-four
108 words from the two of the lists were presented in both the study and test phases and were
109 consider “old”, while the 94 words from the other two lists were presented only in the
110 test phase and were considered “new.” Words were counterbalanced across the typeface and
111 study/test conditions, such that each word served equally often as a target and a foil in both
112 typefaces across participants. The four word lists were counterbalanced across participants,
113 so that each list was assigned to each role (old/new, Arial/Sans Forgetica) an equal number
114 of times. Word order was completely randomized, such that Arial and Sans Forgetica words
115 were randomly intermixed in the study phase, and Arial and Sans Forgetica old and new
116 words were randomly intermixed in the test phase, with old words always presented in the
117 same typeface at test as they were at study.

118 The main difference between the current experiment and Geller et al. (2020) (Ex-
119 periment 3) is that participants were randomly assigned to one of two conditions: the high
120 expectancy test condition or the low expectancy test condition. Interested readers can
121 view the entire task including instructions for each condition by following these links (High
122 Test Expectancy experiment <https://gorilla.sc/openmaterials/72765>; Low test expectancy
123 experiment: <https://gorilla.sc/openmaterials/116227>).

124 The experiment proper consisted of four phases: a study phase, JOL phase, distractor
125 phase, and test phase. During the study phase, a fixation cross appeared at the center of
126 the screen for 500 ms. The fixation cross was immediately replaced by a word in the same
127 location. To continue to the next trial, participants pressed the continue button at the
128 bottom of the screen. Each trial was self-paced. In the JOLs phase, participants provided
129 list-wide JOLs which required them to denote on a scale of 0-100 how likely it will be that
130 they will recall the words studied in Arial and Sans Forgetica on a final test. In the distractor
131 phase, participants completed a short three-minute distractor task wherein they wrote down

as many U.S. state capitals as they could. In the test phase, participants took a yes/no recognition memory test. During the test phase, 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 typeface, 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 studied the word during study, and a box labeled “new” to indicate they did not remember studying the word. Sans Forgetica 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

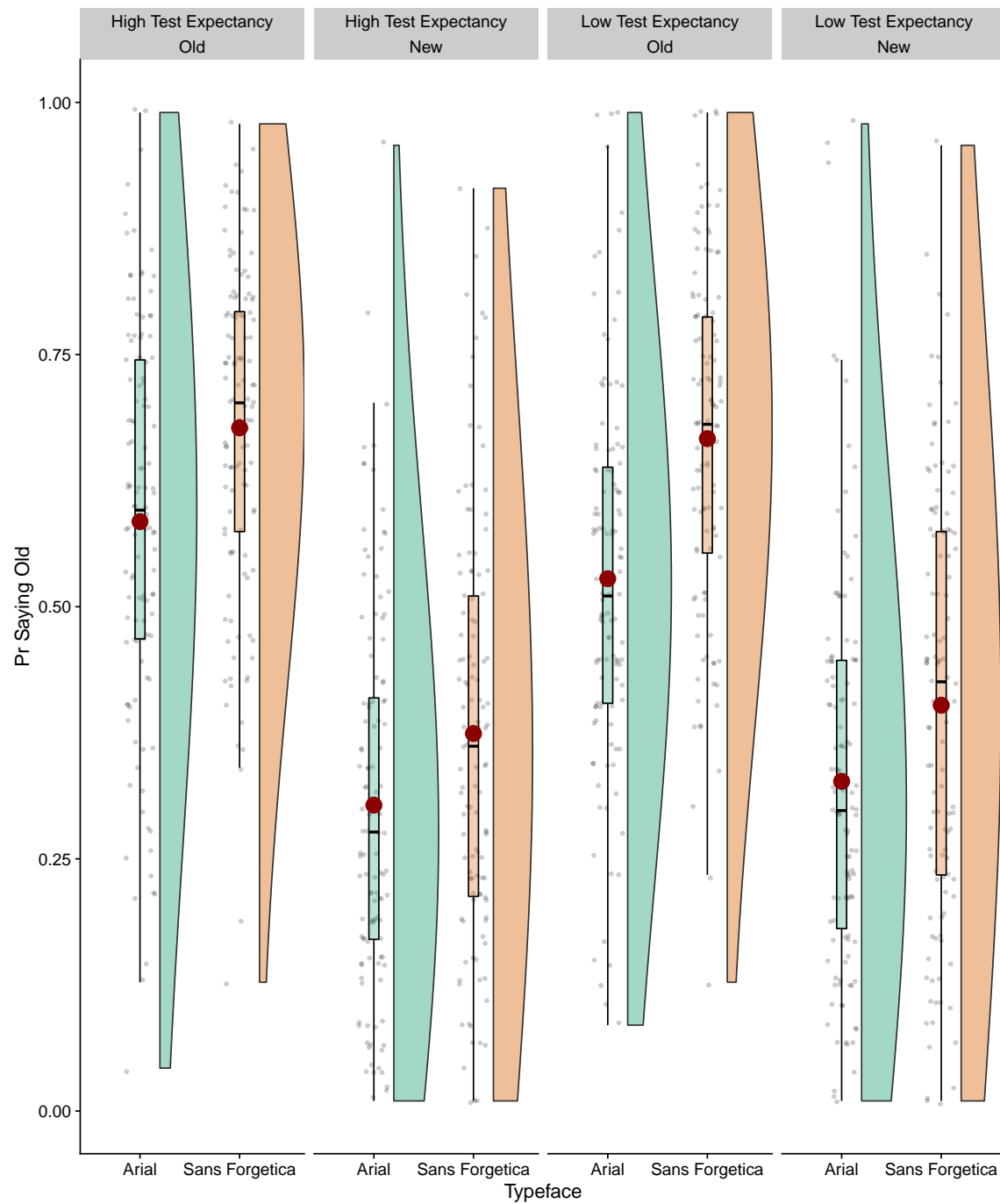
A variation of Cohen’s d (d_{avg}) and generalized eta-squared (η_g^2 ; ???) are used as effect size measures. Alongside traditional analyses that utilize null hypothesis significance testing (NHST), we also report the Bayes factors (BFs) for reported null effects. A Bayes Factor ≥ 3 will be deemed as moderate evidence for null; $\text{BF} \geq 10$ strong evidence for the null. All data were analyzed in R (vers. 4.0.2; R Core Team, 2020), with models fit using the afex (vers. 0.27-2; Singmann et al. (2020)) and BayesFactor packages (vers. 0.9.12-4.2; Morey and Rouder (2018)). All figures were generated using ggplot2 (vers. 3.3.0; Wickham, 2006).

Recognition Memory.

Performance was examined with d' , a memory sensitivity measure derived from signal detection theory (Macmillan & Creelman, 2005). The proportions of “old” responses for old/new items are displayed in Fig. 1. Hits or false alarms at ceiling or floor were changed to .99 or .01. Sensitivity (d') values be seen in Figure 2a. The analysis revealed that when told about a memory test, participants had better discriminatory ability than those not told about a memory test (0.88 vs. 0.72), $M_{\text{diff}} = 0.16$, $F(1, 229) = 4.11$, $\eta_g^2 = .014$, $p = .044$. Individuals were better at discriminating target words presented in Sans Forgetica than Arial (0.86 vs. 0.74), $M_{\text{diff}} = 0.12$, $F(1, 229) = 10.73$, $\eta_g^2 = .010$, $p = .001$. This was qualified by an interaction between Test Expectancy and Typeface, $F(1, 229) = 4.34$, $\eta_g^2 = .004$, $p = .038$. Simple effects showed that individuals in the low expectancy group showed better recognition memory for words presented in Sans Forgetica font compared to Arial, $F(1, 229) = 14.297$, $p < .001$, $d_{\text{avg}} = 0.31$. In the high test expectancy group, there were no differences between the two typefaces, $F(1, 229) = 0.716$, $p = .398$, $d_{\text{avg}} = 0.07$, $\text{BF}_{01} = 5.83$.

JOLs

Seven participants did not provide JOLs to each typeface. We did not analyze the data for those participants. Using the same model as above, participants in the high testing expectancy group had higher JOLs than those in the low testing group, $F(1, 221) = 16.01$, $\eta_g^2 = .065$, $p < .001$. Arial elicited higher JOLs than Sans Forgetica (61.5 vs. 57.5), M_{diff}

**Figure 1**

Raincloud plots (Allen et al., 2019) depicting raw data (dots), box plots, and half violin kernel density plots, with mean (red dot). Proportion of "old" responses as a function of Test Expectancy for Experiment 1.

172 = 4.0, $F(1,221) = 27.05$, $\eta_g^2 = .004$, $p < .001$. There was no interaction between Testing
 173 Expectancy and Typeface, $F(1,221) = 0.13$, $\eta_g^2 < .001$, $p = .715$. Compared to a main
 174 effects-only model, there was strong evidence for no interaction, $BF_{01} = 7.28$.

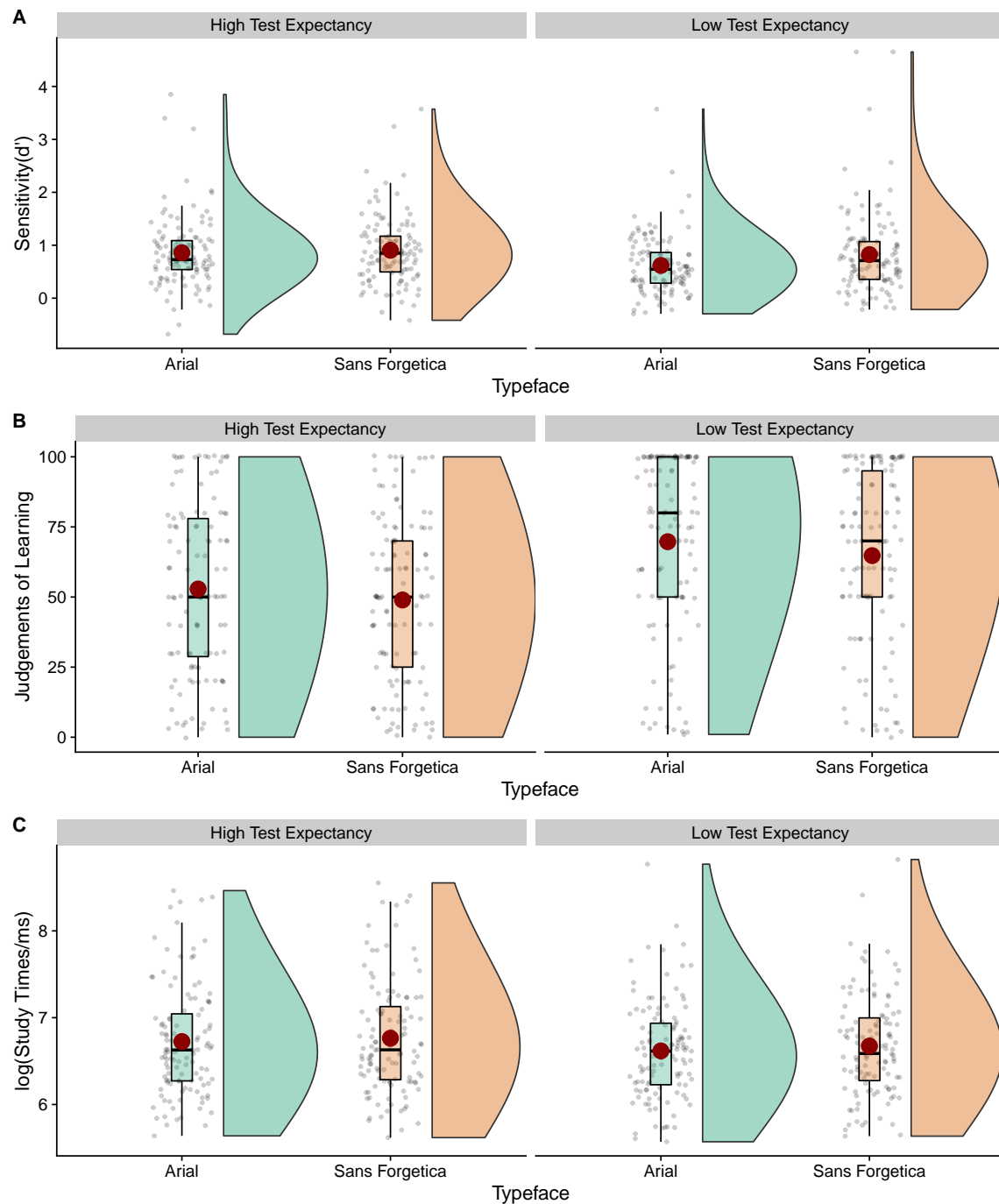
175 *Study Times*

176 Although not pre-registered, study times less than 200 ms and reaction times greater
 177 than 2.5 SD above the mean per condition for each participant were removed. This outlier
 178 procedure removed $\sim 3\%$ of the data. Given the heavy positive skew of the data, we log
 179 transformed study times to better approximate a normal distribution(see Fig.1C). Evidence
 180 for testing expectancy effects on log-transformed study times were inconclusive, $F(1,229)$
 181 $= 1.97$, $\eta_g^2 = .008$, $p = .162$, $BF = 1.822$. Typeface did influence study times: study times
 182 were slower for Sans Forgetica than Arial, $F(1,229) = 30.91$, $\eta_g^2 = .001$, $p < .001$. There
 183 was no interaction between Testing Expectancy and Typeface, $F(1,229) = 1.10$, $\eta_g^2 < .001$,
 184 $p = .296$. Compared to a main effects-only model, there was strong evidence that there was
 185 no interaction between Testing Expectancy and Typeface, $BF_{01} = 5.25$.

186 *Dicussion*

187 As predicted, memory sensitivity for Sans Forgetica was higher when testing ex-
 188 pectancy was low, but not when testing expectancy was high. This suggests that one
 189 potential reason for Taylor et al. (2020) and Geller et al. (2020) failing to find a Sans
 190 Forgetica effect was high test expectancy. This finding replicates what Geller and Still
 191 (2018) found with a masking perceptual disfluency manipulation. We also found that par-
 192 ticipants gave lower JOLs to stimuli studied in the Sans Forgetica typeface. These findings
 193 are inconsistent with the predictions pre-registered, and contradict the findings of Geller
 194 et al. (2020) (Experiment 2) and Taylor et al. (2020) (Experiment 1). One reason for
 195 this is that in the current experiment, we used a within-subject manipulation of typeface
 196 whereas Geller et al. (2020) (Experiment 2) and Taylor et al. (2020) (Experiment 1) used
 197 a between-subjects typeface manipulation. The finding of lower JOLs to disfluent stimuli
 198 compared to more fluent stimuli is inline with other studies using a within-participant ma-
 199 nipulation of fluency (Besken and Mulligan (2013); Geller et al. (2018); Rhodes and Castel
 200 (2008); Rhodes and Castel (2009) Besken and Mulligan (2013)). In relation to study times,
 201 we found that participants studied Sans Forgetica stimuli longer than Arial, regardless of
 202 test expectancy. This contradicts the null finding of Geller et al. (2020) (Experiment 3). It
 203 is important to note, however, that the examination of study times in Geller et al. (2020)
 204 were unplanned, and purely exploratory, making it hard to draw firm conclusions about the
 205 effect fo Sans Forgetica on study times.

206 In Experiment 2, we attempt to replicate these findings using a different criterion
 207 test: cued recall. Taylor et al. (2020) failed to find a Sans Forgetica effect using highly
 208 related cue-target pairs. However, participants were told about the upcoming test. Us-
 209 ing Taylor et al. (2020)'s stimuli we we examined cued recall accuracy with low testing
 210 expectany, along with JOLs and RTs.

**Figure 2**

Raincloud plots (Allen et al., 2019) depicting raw data (dots), box plots, and half violin kernel density plots. A. Memory sensitivity (d') as a function of Typeface and Testing Expectancy. B. Judgements of Learning as a function of Typeface and Test Expectancy. C. Study times (log transformed) as a function of Typeface and Test Expectancy. Raincloud plots (Allen et al., 2019) depicting raw data (dots), box plots, and half violin kernel density plots represent the kernel density of average accuracy (black dots) with the mean (white dot)

Experiment 2

Methods

Participants

One hundred and sixteen participants ($N = 116$) participated through Prolific for U.S. \$2.43. All participants were native English speakers with normal or corrected-to-normal vision. A sensitivity analysis conducted with the R package pwr(Champely, 2020) indicated that our sample size provided 90% power to detect a small effect size ($d = 0.16$) or larger.

Design

Cued recall accuracy, JOLs, and reading times to Typefaces (Sans Forgetica vs. Arial) were analyzed with a paired t -test.

Materials and Procedure

The materials were adopted from Taylor et al. (2020, Experiment 2). Twenty highly associated word pairs were used (see osf page for stimuli characteristics).

Similar to Experiment 1, Experiment 2 consisted of four phases, and was administered online through the gorilla.sc platform. The entire experiment can be run by following the following link: <https://gorilla.sc/openmaterials/116224>. During the study phase, participants were presented with a series of 20 word pairs, presented one at time. They were not told about the upcoming memory test and were told to simply read the cue-target pairs. Participants were told to press the continue button after they had read each word. Half of the word pairs were presented in Sans Forgetica and half in Arial. We created two versions of the word pair list, so that each cue-target pair was presented in each typeface across participants. All counterbalanced lists contained the same word pairs. In the JOL phase, participants made list-wide JOLs. In the distractor Phase, participants took part in the same distractor task as Experiment 1. Finally, in the test phase of the experiment, participants' memory for the word pairs was tested by presenting the first word of the pair they studied during phase 1 and asking them to type the second word of that pair into a box. We presented the memory test in a font not tied to the study phase so as not to reinstate context at test. The cued words presented during Phase 1 were presented one-by-one, in a random order.

Scoring

To score typed responses during the cued recall phase, we used the lrd package in R (Nicholas P. Maxwell, 2020). The lrd package provides an automated way to score word responses. A partial match of 80% was used to determine whether a typed response was correct or not.

Results and Discussion

Cued Recall

With low testing expectancy, performance was better when words were presented in Sans Forgetica (47% vs. 42%), $M_{\text{diff}} = 5\%$, $t(115) = 2.363$, $SE = 0.046$, $p = .020$, 95 CI% [0.008, 0.090], $d_{\text{avg}} = 0.18$. See fig 2a.

JOLs

The analysis of JOLs revealed that Participants' JOLs were lower for Sans Forgetica than Arial (65.83 vs. 70.84), $M_{\text{diff}} = -5.02$, $t(108) = -3.12$, $SE = 1.61$, 95 CI% [0.030, 0.114], $p = .002$, $d_{\text{avg}} = 0.15$. See fig 2a.

Reaction Times

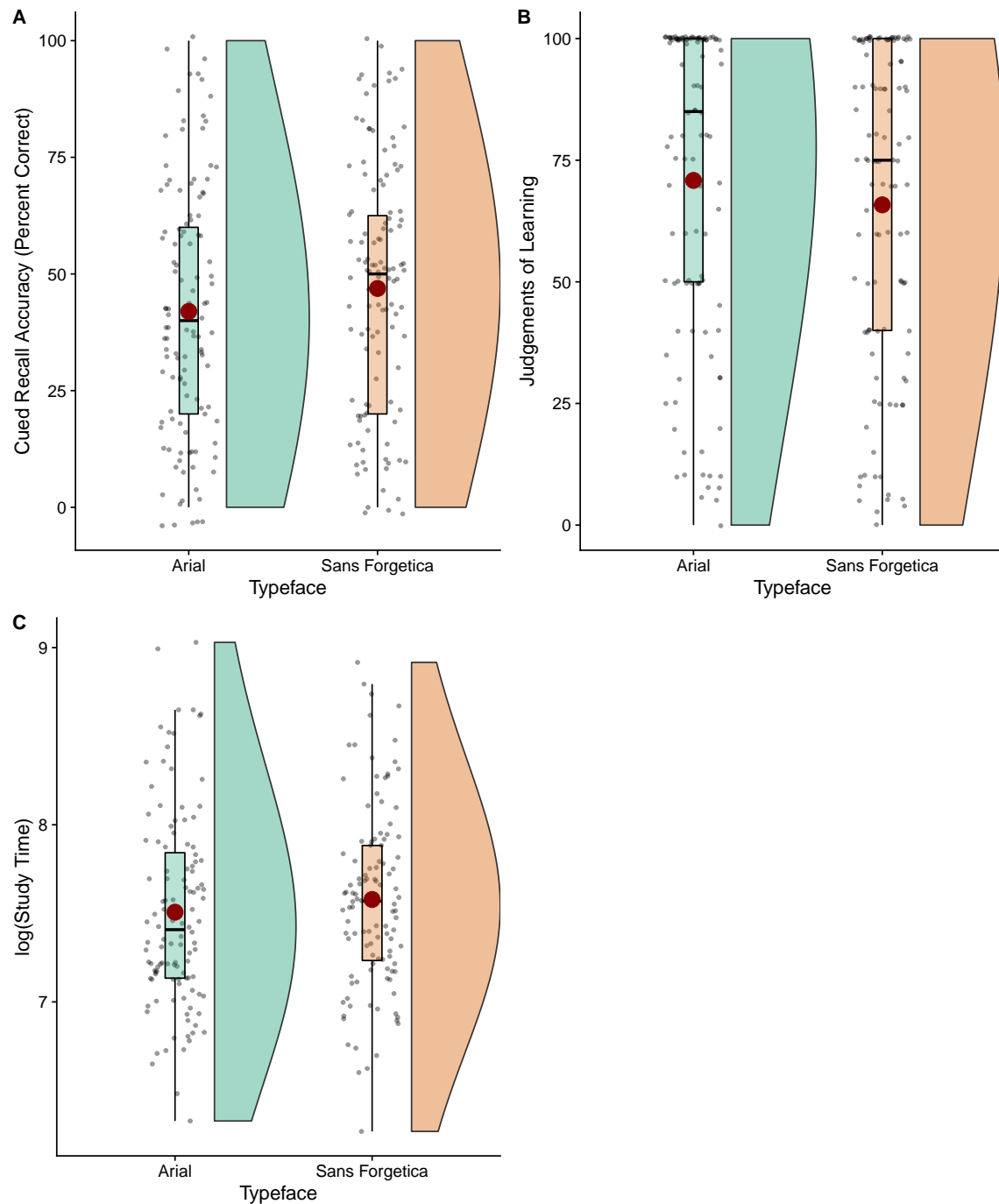
Similar to Experiment 1, we excluded reaction times less than 200 ms and reaction times greater than 2.5 SD above the mean per condition for each participant. The outlier procedure removed $\sim 3\%$ of the data. We also log transformed the data (see Fig.1C for reaction time data). An analysis of study time using a paired t -test on mean log RTs revealed that study times were longer for Sans Forgetica than Arial (7.58 vs. 7.51), $M_{\text{diff}} = 0.072$, $t = 3.40$, $SE = 236$, $p < .001$, 95 CI% [0.030, 0.114], $d_{\text{avg}} = 0.13$.

Using a cued recall test, we have again showed that if test expectancy is low, Sans Forgetica can constitute a desirable difficulty. We observed a 5% increase when participants studied cue-target pairs in Sans Forgetica. Further, we also showed that again Sans Forgetica produced lower JOLs and leads to longer study times.

General Discussion

The present experiments focused on examining whether testing expectancy serves as boundary condition to the Sans Forgetica effect. Specifically, it was assumed that if Sans Forgetica is a desirable difficulty, it fosters learning by increasing mental effort and by stimulating deeper processing - but only when students are endangered to process materials superficially. When students study in preparation for an upcoming test (high test expectancy), they invest mental effort and take their time to elaborate on all context, regardless of whether the to-be-learned information is fluent or disfluent. However, when students do not expect a test (low test expectancy), they might choose to study the text they deem more difficult (e.g., see the discrepancy-reduction model, (???)). This would lead to a desirable effect of Sans Forgetica on memory.

In line with this prediction, Experiment 1, using a yes/no recognition memory test, revealed a desirable effect of Sans Forgetica only when participants were not told about an upcoming memory test. Further, In Experiment 2, using a low testing expectancy design, cued recall performance was significantly higher for Sans Forgetica than Arial. Furthermore, in both experiments Sans Forgetica produced lower JOLs and longer study times overall

**Figure 3**

Raincloud plots (Allen et al., 2019) depicting raw data (dots), box plots, and half violin kernel density plots. A. Memory sensitivity (d') as a function of Typeface and Testing Expectancy. B. Judgements of Learning as a function of Typeface and Test Expectancy. C. Study times (log transformed) as a function of Typeface and Test Expectancy. Raincloud plots (Allen et al., 2019) depicting raw data (dots), box plots, and half violin kernel density plots represent the kernel density of average accuracy (black dots) with the mean (white dot)

thereby suggesting that Sans Forgetica is perceptually disfluent (see @ further evidence for this with eye-tracking evidence). Taken together, it appears that testing expectancy is a powerful moderating factor of the Sans Forgetica effect.

While the finding that Sans Forgetica can enhance memory is in accordance with with other perceptual disfluency manipulations shown to enhance memory (e.g., masking, handwritten cursive; (???); Geller et al. (2018)), these results need to be interpreted with caution. First, and most importantly, the finding that Sans Forgetica is only beneficial to memory under low test expectancy makes it educationally unrealistic. Students always know about upcoming tests. Second, looking at the mnemonic effect sizes of the Sans Forgetica effect (Experiment 1: $d = 0.30$; Experiment 2: $d = .25$), the effects are quite small in nature. It is unclear if these effects would replicate in an educational setting where effect sizes are known to be a lot smaller [1]. Second, the finding that Sans Forgetica is only beneficial to memory under low test expectancy makes it educationally unrealistic. Students always know about upcoming tests. More generally, these findings have broader implications for the disfluency effect. Given that testing expectancy can temper the positive effects of perceptual disfluency, it calls into question the overall usefulness of disfluency as a study tool. Recent research showing disfluency effects (e.g., Geller et al. (2018); Rosner 2015;) have all used designs that utilize low testing expectancy. It is unclear if some perceptual disfluency manipulations are more robust than others to testing expectancy. Future research should examine this. Taken together, while we show positive effects of Sans Forgetica on memory, it is not advised that students utilize it as a study tool.

Conclusion

Recent new reports have recommended that teachers and students use perceptual disfluency to enhance learning. Although we have shown that Sans Forgetica can enhance learning in a very simplified context (i.e., list learning), its effectiveness as a potential learning technique is tempered by the finding that testing expectancy can eradicate the effect. In an educational setting, students are always told about upcoming tests. Thus, Sans Forgetica, and perceptual disfluency in general, might not be an effective manipulation to enhance memory in a more ecologically valid setting. What is clear from the current findings is that the impact of perceptual disfluency manipulations, such as Sans Forgetica, on memory is straightforward. Future research should continue to explore the boundary conditions of the disfluency.

References

- Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). *The english lexicon project* (Nos. 3; Vol. 39, pp. 445–459). Springer New York LLC. <https://doi.org/10.3758/BF03193014>
- Besken, M., & Mulligan, N. W. (2013). Easily perceived, easily remembered? Perceptual interference produces a double dissociation between metamemory and memory performance. *Memory and Cognition*, 41(6), 897–903. <https://doi.org/10.3758/s13421-013-0307-8>
- 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). Worth Publishers.
- Champely, S. (2020). *Pwr: Basic functions for power analysis*. <https://CRAN.R-project.org/package=pwr>
- Eitel, A., & Köhl, T. (2016). Effects of disfluency and test expectancy on learning with text. *Metacognition and Learning*, 11(1), 107–121. <https://doi.org/10.1007/s11409-015-9145-3>
- Eskenazi, M. A., & Nix, B. (2020). Individual Differences in the Desirable Difficulty Effect During Lexical Acquisition. *Journal of Experimental Psychology: Learning Memory and Cognition*. <https://doi.org/10.1037/xlm0000809>
- Geller, J., Davis, S. D., & Peterson, D. J. (2020). Sans Forgetica is not desirable for learning. *Memory*. <https://doi.org/10.1080/09658211.2020.1797096>
- 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>
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*, 2nd ed. (pp. xix, 492–xix, 492). Lawrence Erlbaum Associates Publishers.
- Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of bayes factors for common designs*. <https://CRAN.R-project.org/package=BayesFactor>
- Nicholas P. Maxwell, E. M. B., Mark J. Huff. (2020). *Lrd: A package for processing lexical response data*.
- 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>

- 353 Rhodes, M. G., & Castel, A. D. (2008). Memory Predictions Are Influenced by Percep-
354 tual Information: Evidence for Metacognitive Illusions. *Journal of Experimental*
355 *Psychology: General*, 137(4), 615–625. <https://doi.org/10.1037/a0013684>
- 356 Rosner, T. M., Davis, H., & Milliken, B. (2015). Perceptual blurring and recognition
357 memory: A desirable difficulty effect revealed. *Acta Psychologica*, 160, 11–22. <https://doi.org/10.1016/j.actpsy.2015.06.006>
- 359 Sagan, C. (1980). *Broca's brain: Reflections on the romance of science*. [https://books.](https://books.google.com/books?hl=en%7B/&%7Dlr=%7B/&%7Ddid=GlXPqexwO28C%7B/&%7Ddoi=fnd%7B/&%7Dpg=PR4%7B/&%7Ddots=65nePfKWk5%7B/&%7Dsig=CTTgqKJLaozsFvFqBYjBd%7B/_%7DEOkxE)
360 [google.com/books?hl=en%7B/&%7Dlr=%7B/&%7Ddid=GlXPqexwO28C%7B/](https://books.google.com/books?hl=en%7B/&%7Dlr=%7B/&%7Ddid=GlXPqexwO28C%7B/&%7Ddoi=fnd%7B/&%7Dpg=PR4%7B/&%7Ddots=65nePfKWk5%7B/&%7Dsig=CTTgqKJLaozsFvFqBYjBd%7B/_%7DEOkxE)
361 [&%7Ddoi=fnd%7B/&%7Dpg=PR4%7B/&%7Ddots=65nePfKWk5%7B/&%7Dsig=](https://books.google.com/books?hl=en%7B/&%7Dlr=%7B/&%7Ddid=GlXPqexwO28C%7B/&%7Ddoi=fnd%7B/&%7Dpg=PR4%7B/&%7Ddots=65nePfKWk5%7B/&%7Dsig=CTTgqKJLaozsFvFqBYjBd%7B/_%7DEOkxE)
362 [CTTgqKJLaozsFvFqBYjBd%7B/_%7DEOkxE](https://books.google.com/books?hl=en%7B/&%7Dlr=%7B/&%7Ddid=GlXPqexwO28C%7B/&%7Ddoi=fnd%7B/&%7Dpg=PR4%7B/&%7Ddots=65nePfKWk5%7B/&%7Dsig=CTTgqKJLaozsFvFqBYjBd%7B/_%7DEOkxE)
- 363 Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2020). *Afex:*
364 *Analysis of factorial experiments*. <https://CRAN.R-project.org/package=afex>
- 365 Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2007). Expectation of a final
366 cumulative test enhances long-term retention. *Memory and Cognition*, 35(5), 1007–
367 1013. <https://doi.org/10.3758/BF03193473>
- 368 Taylor, A., Sanson, M., Burnell, R., Wade, K. A., & Garry, M. (2020). Disfluent difficul-
369 ties are not desirable difficulties: the (lack of) effect of Sans Forgetica on memory.
370 *Memory*, 1–8. <https://doi.org/10.1080/09658211.2020.1758726>
- 371 Weinstein, Y., Gilmore, A. W., Szpunar, K. K., & McDermott, K. B. (2014). The role of test
372 expectancy in the build-up of proactive interference in long-term memory. *Journal*
373 *of Experimental Psychology: Learning Memory and Cognition*, 40(4), 1039–1048.
374 <https://doi.org/10.1037/a0036164>
- 375 Xie, H., Zhou, Z., & Liu, Q. (2018). Null Effects of Perceptual Disfluency on Learning
376 Outcomes in a Text-Based Educational Context: a Meta-analysis. *Educational Psy-*
377 *chology Review*, 30(3), 745–771. <https://doi.org/10.1007/s10648-018-9442-x>