Surprise! Low Testing Expectancy Moderates the Sans Forgetica Effect

Jason Geller[1,2] & Kelly A. Kane[3]

[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 yiedled discrepant findings. To clarify this discrepancy, the present experiments examined a boudnary condition that determines when Sans Forgetica is and 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 eliciated lower judgements of learning and longer study times, but only improved memory on a an old/new recognition test when there was low test expectancy (compared to a high test expectancy group). In Experiment 2 ($N = 116$) using a low testing expectancy cued recall test, 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 intreprting these results. Not only was were effect sizes small, but low testing expectancy is not practical. Echocing previous sentiments, students wanting to remember more and forget less should stick to other desirable difficultues shown to enhance memory.

Keywords: Disfluency, Desirable Difficuties, Recognition, Recall

Word count: 3700

Surprise! Low Testing Expectancy Moderates the Sans Forgetica Effect

Successful remembering is impacted by innumerable factors. One factor that has been purported to enhance remembering is perceptual disfluency. Interfering with word perception during encoding by blurring (Rosner et al., 2015), inversion (Sungkhasettee et al., 2011), or placing the word in a atypical font (Diemand-Yauman et al., 2011) can enhance explicit memory, a phenomenon dubbed the perceptual interference effect (Nairne, 1988), or more recently, the disfluency effect (Geller et al., 2018). 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 (Earp, 2018). The Sans Forgetica typeface is purported to be a desirable difficulty [Bjork and Bjork (2011)[that staves off forgetting and enhances learning due to the disfluent perceptual characteristics of the typeface (e.g., the letters are blank slanted and have intermittent gaps). The claims surrounding Sans Forgetica have lead to extensive press coverage from major news outlets (e.g., NPR, Washington Post), and have lead to the development of 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). 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 discriminabity (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 discrimnability 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 Taylor et al. (2020) and Geller et al. (2020) is testing expectancy. Eskenazi and Nix (2020) did 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 know 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 perecptual disfluency effects, Eitel and Kühl (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 countervails 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 Kühl (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 occured 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 old/new 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 Fagoretica is perceptual 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-Irvine2020]. 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 we had a total of 231 participants. Participants completed the experiment in return for U.S.$8.00 an hour.

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

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

Procedure

Similar to Geller et al. (2020) (Experiment 3), four lists (94 words each; 47 in each typeface condition) were used to create the stimuli for a total of 188 words. Ninety-four words from the two of the lists were presented in both the study and test phases and were consider "old", while the 94 words from the other two lists were presented only in the test phase and were considered "new." Words were counterbalanced across the typeface and study/test conditions, such that each word served equally often as a target and a foil in both typefaces across participants. The four word lists were counterbalanced across participants, so that each list was assigned to each role (old/new, Arial/Sans Forgetica) an equal number of times. 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 typeface at test as they were at study.

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

The experiment proper consisted of four phases: study, JOLs, distractor, and test. 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 teh same location. To continue to the next trial, participants pressed the continue button at the bottom of the screen. Each trial was self-paced. In the JOLs phase, participants provided list-wide JOls which required them to denote on a scale of 0-100 how likely it will be that they will recall the words studied in Arial and Sans Forgetica on a final test. The distractor task between encoding and test lasted approximately 3 minutes during which participants wrote down as many U.S. state capitals as they could. In the test phase, participants took an old/new 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. Stimuli stayed on the screen until participants cliked on either the "old" or "new" box. 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}; 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 (BFs) for reported null effects. A Bayes Factor $> = 3$ will be deemed as moderate evidence for null; BF $> =$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 (2018a)). 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). Hits or false alarms at ceiling or floor were changed to .99 or .01. Sensitivity (d') values can 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_{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),$M_{\text{diff}} = 0.12$, $F(1, 229) = 10.73$, η_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, $F(1, 229) = 0.716$, $p = .398$, $d_{\text{avg}} = 0.07$, $\text{BF}_{\text{O1}} = 5.83$.

JOLs

JOLs are presented in Figure 1B. 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), $M_{\text{diff}} = 4.0$, $F(1{,}221) = 27.05$, $\eta_g^2 = .004$, $p < .001$. There was no interaction between Testing Expectancy and Typeface, $F(1{,}221) = 0.13$, $\eta_g^2 < .001$, $p = .715$. Compared to a main effects-only model, there was strong evidence for no interaction, $\text{BF}_{01} = 7.28$.

Study Times

Although not pre-registered, study times less than 200 ms and reaction times greater than 2.5 SD above the mean per condition for each participant were removed. This outlier procedure removed ~3 % of the data. Given the heavy positive skew of the data, we log transformed study times to better approximate a normal distribution (see Fig.1C).

Evidence for testing expectancy effects on log-transformed study times were inconclusive, $F(1{,}229) = 1.97$, $\eta_g^2 = .008$, $p = .162$, BF = 1.822. Typeface did influence study times: study times were slower for Sans Forgetica than Arial, $F(1{,}229) = 30.91$, $\eta_g^2 = .001$, $p < .001$. There was no interaction between Testing Expectancy and Typeface, $F(1{,}229) = 1.10$, $\eta_g^2 < .001$, $p = .296$. Compared to a main effects-only model, there was strong evidence that there was no interaction between Testing Expectancy and Typeface, $\text{BF}_{01} = 5.25$.

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

In Experiment 2, we attempted to replicate the finding from Experiment 1 using a different criterion test: cued recall. Taylor et al. (2020) (Experiment 2) failed to observe a

Sans Forgetica effect using highly related cue-target pairs. However, participants were told about the upcoming test. Using the highly related word pairs from Taylor et al. (2020), we set out to examine cued recall accuracy along with JOLs and RTs, with low testing expectations.

Experiment 2

Methods

Participants

One hundred and sixteen participants (N = 116) participated through Prolific (Prolific.co), and comleted the study through Gorilla (Anwyl-Irvine et al., 2020). 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 el al. (2020, Experiment 2). Twenty highly associated word pairs were used (see OSF page for stimuli characteristics).

The entire experiment can be run by following the following link: https://gorilla.sc/openmaterials/116224. Similar to Experiment 1, the experiment consisted of encoding, JOL, distractor, and test phases. At study, participants were not told about the upcoming memory test and were told to simply read the cue-target pairs. Participants were presented with a series of 20 word pairs, one at time. Typefaces were randomly intermixed. Participants were told to press the continue button after they had read each word. 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. At test, the cues from each word pair were presented individually and the participants had to type in the corresponding target (or guess if they could not remember). Responses were not time-limited. Stumuli presented at test were presented in a different typeface (Open Sans) so as not to reinstate context at test.

Scoring

Typed responses were scored with 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

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

JOLs

Figure 2b shows JOL responses. The analysis of JOls revealed that partcipants' JOLs were lower for Sans Forgetica than Arail (65.83 vs. 70.84), $M_{\text{diff}} = -5.02$, $t(108) = -3.12$, $SE = 1.61$, 95 CI% [0.030, 0.114], $p = .002$, $d_{\text{avg}} = 0.15$.

Reaction Times

Figure 2c shows log-transformed RTs.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 ~ 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, recognition memory and cued recall were enhanced when stimuli were presented in Sans Forgetica, but only when participants were not told about an upcoming memory test. Moreover, in both experiments Sans Forgetica produced lower JOLs and longer study times overall thereby suggesting that Sans Forgetica is perceptually disfluent (see Eskenazi & Nix, 2020 further evidence for this with eye tracking).

While it might be tempting to use this as evidence for the adoption of Sans Forgetica as a study tool, the current findings 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 its use in the educational domain impractical. Students always know about upcoming tests. Second, looking at the mnemonic effect sizes of the Sans Forgetica effect (Experiment 1: $d = 0.31$; Experiment 2: $d = 0.25$), the effects are quite small in nature. It is unclear if these effects would replicate in an educational setting where effect sizes are a known to be a lot smaller (Butler et al., 2014).

Conclusion

Recent reports have recommended that teachers and students use perceptual disfluency to enhance learning. Although we have shown that a simple perceptual manipulation (i.e., placing font in Sans Forgetica) can enhance learning in a very simplified context (i.e., list learning), its efficaciousness 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.

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 manuscript, collected data, and conducted all statistical analyses.

R and R package acknowledgements

This paper was written in R-Markdown. In RMarkdown, the text and the code for analysis may be included in a single document. The document for this paper, with all text and code, can be found at: . The results were created using R (Version 4.0.2; R Core Team, 2019) and the R-packages *afex* (Version 0.27.2; Singmann et al., 2019), *BayesFactor* (Version 0.9.12.4.2; Morey & Rouder, 2018b), *carData* (Version 3.0.4; Fox et al., 2019), *coda* (Version 0.19.3; Plummer et al., 2006), *cowplot* (Version 1.1.0; Wilke, 2020), *data.table* (Version 1.13.0; Dowle & Srinivasan, 2020), *dplyr* (Version 1.0.2; Wickham et al., 2019), *effects* (Version 4.2.0; Fox & Weisberg, 2018; Fox, 2003; Fox & Hong, 2009), *emmeans* (Version 1.5.0; Lenth, 2020), *forcats* (Version 0.5.0; Wickham, 2019a), *ggplot2* (Version 3.3.2; Wickham, 2016), *ggpol* (Version 0.0.6; Tiedemann, 2019), *ggrepel* (Version 0.8.2; Slowikowski, 2020), *here* (Version 0.1; Müller, 2017), *janitor* (Version 2.0.1; Firke, 2020), *knitr* (Version 1.29; Xie, 2015), *lattice* (Version 0.20.41; Sarkar, 2008), *lme4* (Version 1.1.23; Bates et al., 2015), *lubridate* (Version 1.7.9; Grolemund & Wickham, 2011), *Matrix* (Version 1.2.18; Bates & Maechler, 2019), *modelbased* (Version 0.1.2; Makowski et al., 2020), *MOTE* (Version 1.0.2; Buchanan et al., 2019), *papaja* (Version 0.1.0.9997; Aust &

Barth, 2020), *patchwork* (Version 1.0.1; 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 et al., 2018), *Rmisc* (Version 1.5; Hope, 2013), *see* (Version 0.5.2; Lüdecke et al., 2020), *stringr* (Version 1.4.0; Wickham, 2019b), *tibble* (Version 3.0.3; Müller & Wickham, 2019), *tidyr* (Version 1.1.2; Wickham & Henry, 2019), *tidyverse* (Version 1.3.0; Wickham, 2017), and *WRS2* (Version 1.1.0; Mair & Wilcox, 2020).

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*. https://github.com/crsh/papaja

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

Bates, D., & Maechler, M. (2019). *Matrix: Sparse and dense matrix classes and methods*. 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

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.

Buchanan, E. M., Gillenwaters, A., Scofield, J. E., & Valentine, K. D. (2019). *MOTE: Measure of the Effect: Package to assist in effect size calculations and their* *confidence intervals*. http://github.com/doomlab/MOTE

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

Champely, S. (2020). *Pwr: Basic functions for power analysis*. https://CRAN.R-project.org/package=pwr

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. (2020). *Data.table: Extension of 'data.frame'*. https://CRAN.R-project.org/package=data.table

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

Eitel, A., & Kühl, 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

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

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software*, *8*(15), 1–27. 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. 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*. https://CRAN.R-project.org/package=carData

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

Ginn, J., & Silge, J. (2020). *QualtRics: Download 'qualtrics' survey data*. 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. http://www.jstatsoft.org/v40/i03/

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

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

Lenth, R. (2020). *Emmeans: Estimated marginal means, aka least-squares means*. 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'.* 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). Lawrence Erlbaum Associates Publishers.

Mair, P., & Wilcox, R. (2020). Robust Statistical Methods in R Using the WRS2 Package. *Behavior Research Methods*, *52*, 464–488.

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

Morey, R. D., & Rouder, J. N. (2018a). *BayesFactor: Computation of bayes factors for common designs*. https://CRAN.R-project.org/package=BayesFactor

Morey, R. D., & Rouder, J. N. (2018b). *BayesFactor: Computation of bayes factors for common designs*. https://CRAN.R-project.org/package=BayesFactor

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

Müller, K., & Wickham, H. (2019). *Tibble: Simple data frames.* 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

Nicholas P. Maxwell, E. M. B., Mark J. Huff. (2020). *Lrd: A package for processing lexical response data.*

Olejnik, S., & Algina, J. (2003). *Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs* (Nos. 4; Vol. 8, pp. 434–447).

https://doi.org/10.1037/1082-989X.8.4.434

Pedersen, T. L. (2019). *Patchwork: The composer of plots*. 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. https://journal.r-project.org/archive/

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

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

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

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

Sagan, C. (1980). *Broca's brain: Reflections on the romance of science*. https://books.google.com/books?hl=en%7B/&%7Dlr=%7B/&%7Did=GlXPqexwO28C%7B/&%7Doi=fnd%7B/&%7Dpg=PR4%7B/&%7Dots=65nePfKWk5%7B/&%7Dsig=CTTgqKJLaozsFvFqBYjBd%7B/_%7DEOkxE

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

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2019). *Afex: Analysis of factorial experiments*. https://CRAN.R-project.org/package=afex

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2020). *Afex: Analysis of factorial experiments*. https://CRAN.R-project.org/package=afex

Slowikowski, K. (2020). *Ggrepel: Automatically position non-overlapping text labels with 'ggplot2'*. https://CRAN.R-project.org/package=ggrepel

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

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2007). Expectation of a final cumulative test enhances long-term retention. *Memory and Cognition*, *35*(5), 1007–1013. https://doi.org/10.3758/BF03193473

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'*. https://CRAN.R-project.org/package=ggpol

Weinstein, Y., Gilmore, A. W., Szpunar, K. K., & McDermott, K. B. (2014). The role of test expectancy in the build-up of proactive interference in long-term memory. *Journal of Experimental Psychology: Learning Memory and Cognition*, *40*(4), 1039–1048. https://doi.org/10.1037/a0036164

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

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

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

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

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

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

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

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

Wilke, C. O. (2020). *Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'.* https://CRAN.R-project.org/package=cowplot

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.). Chapman; Hall/CRC. https://yihui.name/knitr/

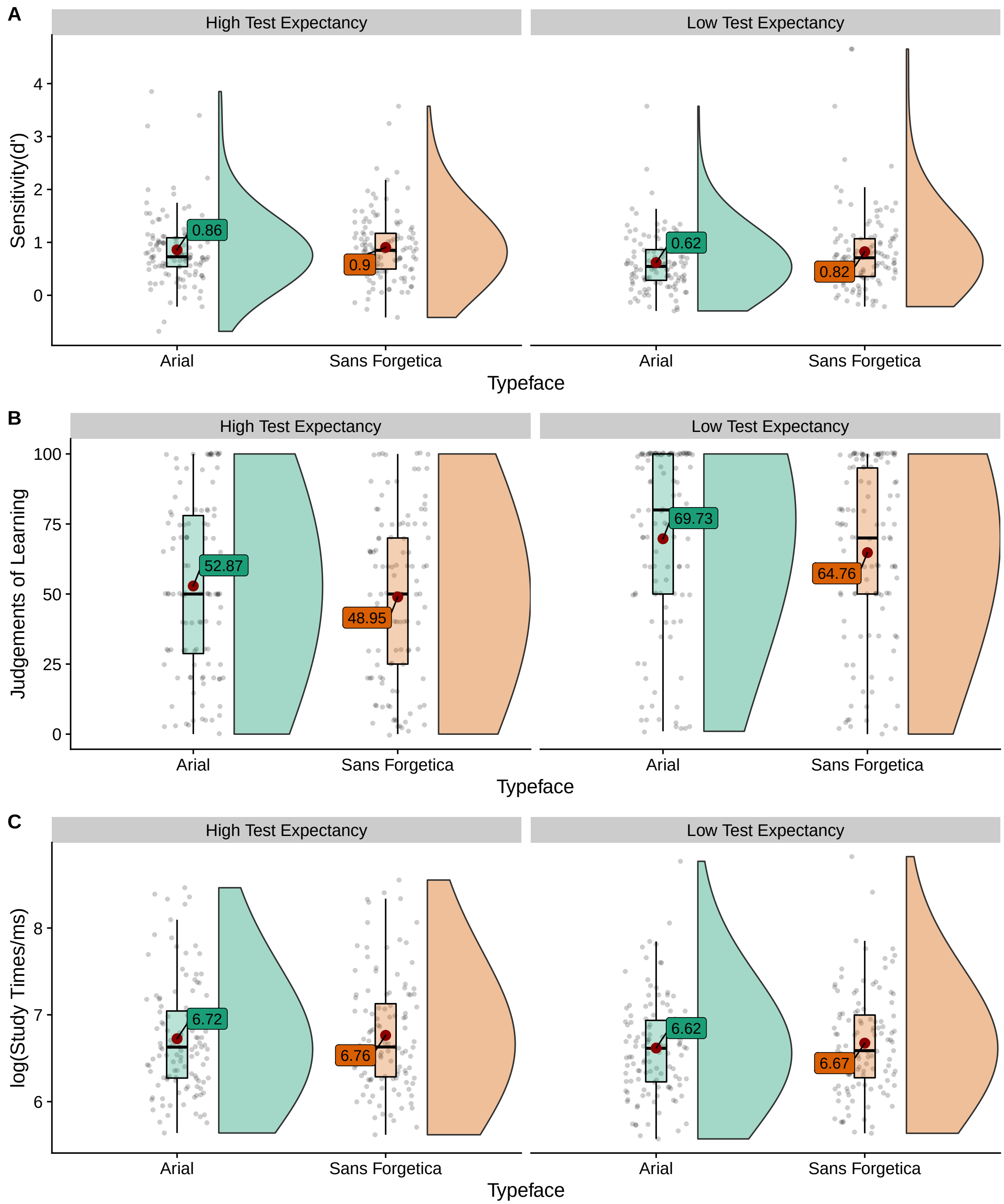

Figure 1

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

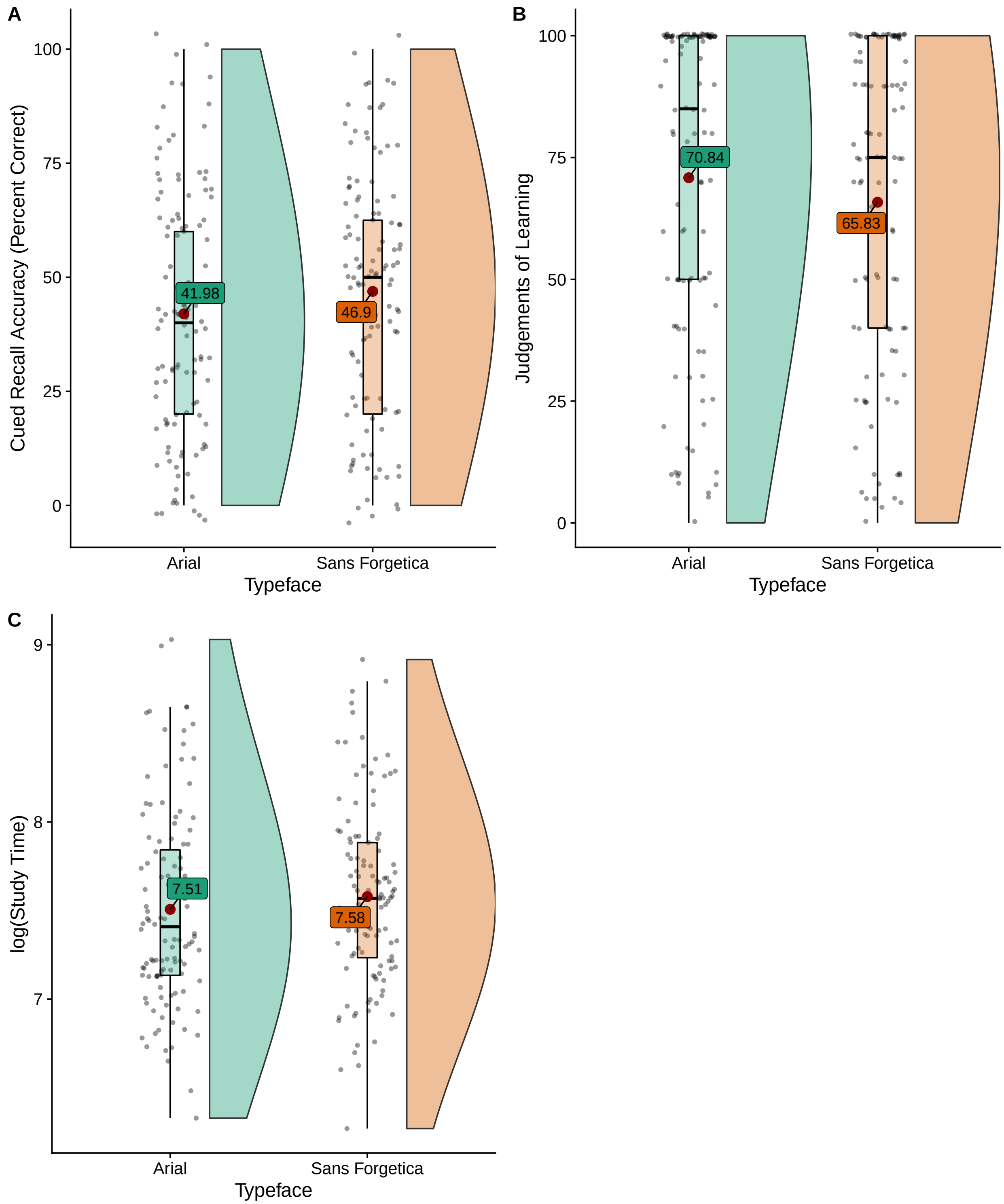

Figure 2

Raincloud plots (Allen et al., 2019) depicting raw data (dots), box plots (with mean (red dot)), and half violin kernel desntiy plots.A. Recall performance. B. Judgements of Learning. C. Study times (log transformed)