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

Author Note

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

Abstract

Recent work examaining the mnemonic effects of Sans Forgetica has been mixed. A possible explanation for this is whether participants were told about an upcoming test or no: testing expectancy. Here we report two experimets investigating the role of testing expectancy using a yes/no recognition memory test (Experiment 1, $N = 231$) and a cued recall test (Experiment 2, $N = 116$). In Experiment 1, Sans Forgetica overall eliciated lower judgements of learning and longer study times, but Sans Forgetica only improved improved memory when there was low test expectancy (compared to high test expectancy). In Experiment 2, using only a low test expectancy design, we found a similar pattern of results to Experiment 1. That is, Sans Forgetia elicted lower JOLs and longer study times, and produced better cued recall. Herein we have shown that Sans Forgetica can produce mnenemonic benefits, but only when testing expectancy is low. Caution should be taken in intreprting these results, however. Not only was the effect size small, but low testing expetcnay is not educationally realistic. Echocing previous failures to replicating the Sans Forgetica effect, students wanting to remember more and forget less should stick to other desirable difficultues shown to enhance memory.

Keywords: Disfluency

Word count: 3500

Surprise! Low Testing Expectancy Moderates the Sans Forgetica Effect

The influential desirable difficulty principle suggests that making learning harder not easier, such as having students take a test over information previously studied, can have noticeable and lasting impacts on student achievement (Bjork & Bjork, 2011; see Sotola & Crede, 2020 for a recent meta-analysis). Recently, the concept of desirable difficulties has been extended to include subtle perceptual manipulations that are difficult to encode (e.g., atypical fonts, blurring, handwritten cursive; **???**; **???**; Geller et al., 2018). One such manipulation garnering increased attention is the Sans Forgetica typeface. 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 disfluent perceptual characteristics of the typeface are purported to stave off forgetting and enhance learning. However, as the famous astronomer Carl Sagan once said, "Extraordinary claims require extraordinary evidence (Sagan, 1980).

There is a growing evidence that perceptual disfluency manipulations are simply not desirable for learning (see for meta-anlyssi). Does the same hold true for Sans Forgetica? In two independent studies, Taylor, Sanson, Burnell, Wade, and Garry (2020) and Geller, Davis, and Peterson (2020) set out to examine whether Sans Forgetica is *really* a desirable difficulty. In the first conceptual replications 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 (**???**), that 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 a Sans Forgetica memory effect.

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. Using eye-tracking, Eskenazi and Nix (2020) had participants learn the spelling and meaning for 15 low-frequency words each presented in the context of two sentences. Both 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 suggest that the Sans Forgetica may be fickle, with positive effects potentially bounded by specific conditions. Probing into Eskenazi and Nix (2020), a critical difference between their study and (**???**) and Geller et al. (2020), is testing expectancy.In Eskenazi and Nix (2020), they did did not tell their participants about the upcoming tests. Thus, one common design feature that may moderate whether we see a Sans Forgetica effect is high testing expectancy. Eitel and Kühl (2016) posited that testing expectancy may be an important moderator of the perceptual disfluency effect. They 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 would countervail the effects of perceptual disfluency by eliciting additional 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) did not find evidence for this,Geller and Still (2018), using a masking disfluency manipulation, demonstrated in a yes/no recognition memory test that indeed only under low testing expectancy does a disfluency effect occur. Given this, it is possible, then, that a Sans Forgetica effect might arise when participants have low test expectancy.

Experiment 1

Experiment 1 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 aggregate judgments of learning (i.e., a subjective memory prediction about future memory performance taken after all items are studied) and study times. We preregistered that if participants were not told about a memory test we would see a memory boot 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

Sample size, experimental design, hypotheses, outcome measures, and analysis plan for Experiment 1 were can be found on the Open Science Framework (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), we presented all participants with 188 words, 94 at study (47 in each typeface condition) and 188 at test (94 old and 94 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. This lead to the creation of 4 counterbalanced lists. 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: a study phase,JOL phase, distractor phase, and test phase. 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. 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. 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.

Analytic Strategy. For both experiments, an alpha level of .05 is maintained. A variation of Cohen's 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; 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, Bolker, Westfall, Aust, and Ben-Shachar (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).

Results and Discussion

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. Hits and false alarms along with sensitivity (d’) can be seen in Figure 1. Participants that were told about a memory test had better discrimination 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, 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, $F(1, 229) = 14.297$, $p < .001$, $d =$ 0.31. In the high test expectancy group, there were no d‘ differences between the two typefaces, $F(1, 229) = 0.716$, $p = .398$, $\text{BF}_{\text{O1}} = 5.83$.

#High Testing Data Load

#Combine

```
## # A tibble: 462 x 11
##    participant_pri~ condition1 testexpect    cr    fa   hit  miss    hr     zhr
##               <int> <chr>      <chr>      <int> <dbl> <int> <int> <dbl>   <dbl>
##  1          1531474 Arial      low           37 0.213    21    26 0.447 -0.134
##  2          1531474 Sans Forg~ low           36 0.234    20    27 0.426 -0.188
##  3          1531487 Arial      low           25 0.468    20    27 0.426 -0.188
##  4          1531487 Sans Forg~ low           26 0.447    23    24 0.489 -0.0267
##  5          1531488 Arial      low           40 0.149    20    27 0.426 -0.188
##  6          1531488 Sans Forg~ low           34 0.277    32    15 0.681  0.470
##  7          1531494 Arial      low           47 0.01     42     5 0.894  1.25
```

```
## 8          1531494 Sans Forg~ low          47 0.01     42     5 0.894  1.25
## 9          1531503 Arial      low          30 0.362    18    29 0.383 -0.298
## 10         1531503 Sans Forg~ low          12 0.745    32    15 0.681  0.470
## # ... with 452 more rows, and 2 more variables: zfa <dbl>, dprime <dbl>
```

```
##
## Univariate Type III Repeated-Measures ANOVA Assuming Sphericity
##
##                          Sum Sq num Df Error SS den Df  F value    Pr(>F)
## (Intercept)             296.652      1  166.184    229 408.7834 < 2.2e-16 ***
## testexpect                2.980      1  166.184    229   4.1058  0.043896 *
## condition1                1.818      1   38.786    229  10.7344  0.001215 **
## testexpect:condition1     0.735      1   38.786    229   4.3369  0.038405 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Anova Table (Type 3 tests)
##
## Response: dprime
##                  Effect     df  MSE        F  ges p.value
## 1            testexpect 1, 229 0.73   4.11 * .014    .044
## 2            condition1 1, 229 0.17 10.73 ** .009    .001
## 3 testexpect:condition1 1, 229 0.17   4.34 * .004    .038
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1
```

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 gorup provided 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, $\mathrm{BF}_{01} = 7.28$.

Study Times. Although not pre-registered, 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. Given reactions times are notoriously positively skewed, we also log transformed the data (see Fig.1C for reaction time data) to better approximate a normal distribution. Evidence for Testing Expectancy influencing study times was inconclusive, $F(1{,}229) = 1.97$, $\eta_g^2 = .008$, $p = .162$, BF = 1.822. Typeface did influence reading times. Log-transformed study times were higher 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, $\mathrm{BF}_{01} = 5.25$.

Dicussion

The results from Experiment1 are clear-cut. 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 the Taylor et al. (2020) and Geller et al. (2020) failure to replicate was high test expectancy. Telling participants about a test lead to deeper processing for both Sans Forgetica and Arial typefaces, reducing any benefit from the Sans Forgetica typeface. This replicates what Geller and Still (2018) found with a masking manipulation. We also found that participants gave lower JOLs to Sans Forgetica and had longer study times compared to Arial. 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). In the current experiment, a within-subject manipulation of typeface was used whereas in Geller et al. (2020) (Experiment 2) and Taylor et al. (2020) used a between-subjects typeface manipulation.The finding of lower JOls to disfluent stimuli compared to more fluent stimuli is inline with other studies that used a within-participant manipulations (). In relation to study times, Geller et al. (2020) did not study time differences between typefaces. To examine this further, in Experiment 2 we examine memory for Sans Forgetica in a cued recall task and collect JOLs and study times ti see if we can replicate the basic finding herein.

```
##
## Univariate Type III Repeated-Measures ANOVA Assuming Sphericity
##
##                      Sum Sq num Df Error SS den Df    F value    Pr(>F)
## (Intercept)         20706.2      1  168.431    229 28152.2648 < 2.2e-16 ***
## testexpt                1.1      1  168.431    229     1.5354    0.2166
## condition               0.3      1    1.797    229    33.0251 2.884e-08 ***
## testexpt:condition      0.0      1    1.797    229     1.1292    0.2891
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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) with a paired *t*-test.

Materials and Procedure. The materials were adopted from Taylor el al. (2020, Experkment 2). Twenty highly associated word paris, were used (taken from the University of Florida norms).

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 phase 1, participants were presented with a series of 20 word pairs, presented one at time. 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 Phase 2, participants were presented with the same distractor task as Experiment 1. Finally, in the third 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 stud 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. Looking at particpants JOLs to each Typeface, 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], $p = .002$, $d_{\text{avg}} = 0.15$. 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 ~ 3% of the data. We also log transformed the data (see Fig.1C for reaction time data). A paired t-test on mean log RTs showed that reading times were larger 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], $d_{\text{avg}} = 0.13$.

General Discussion

Herein we have shown a boundary condition for the Sans Forgetica effect: testing expectancy. To summarize our findings, In Experiment 1 using a a recognition memory Sans Forgetica exerted a positive effect on memory when p were not told about upcoming memory test. In experiment 21 Similar to other perceptual disfluency manipulations (masking, handwritten cursive) sans forgetica seemed to be o jefgive

Contrary to Experiments 1-3, when testing expectancy was low, we observed better memory for materials in Sans Forgetica. This provides a potential boundary condition for the Sans Forgetica effect. That is, when testing expectancy is high (e.g., Experiments 1-3) we do not see a Sans Forgetica effect. However, we do when testing expectancy is low. This might offer a potential explanation for why there is mixed evidence on the effectiveness of Sans Forgetica to enhance memory (See Eskenazi & Nix, 2020). The results herein might explain why they did find a positive effect for Sans Forgetica in a subset of their participants. Despite this, given the small effect size and the fact that studying is almost always done intentionally, their is really no evidence that it should be used as a study tool.

RTs (one possible is optimal study hypothesis switching from harder stimuli to stumuli they know). JOLs would contradict this.

References

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

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.

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

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

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). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of bayes factors for common designs*. Retrieved from 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*.

Sagan, C. (1980). *Broca's brain: Reflections on the romance of science*. Retrieved from 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

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

Sotola, L. K., & Crede, M. (2020). Regarding Class Quizzes: a Meta-analytic Synthesis of Studies on the Relationship Between Frequent Low-Stakes Testing and Class Performance. *Educational Psychology Review*, 1–20. https://doi.org/10.1007/s10648-020-09563-9

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

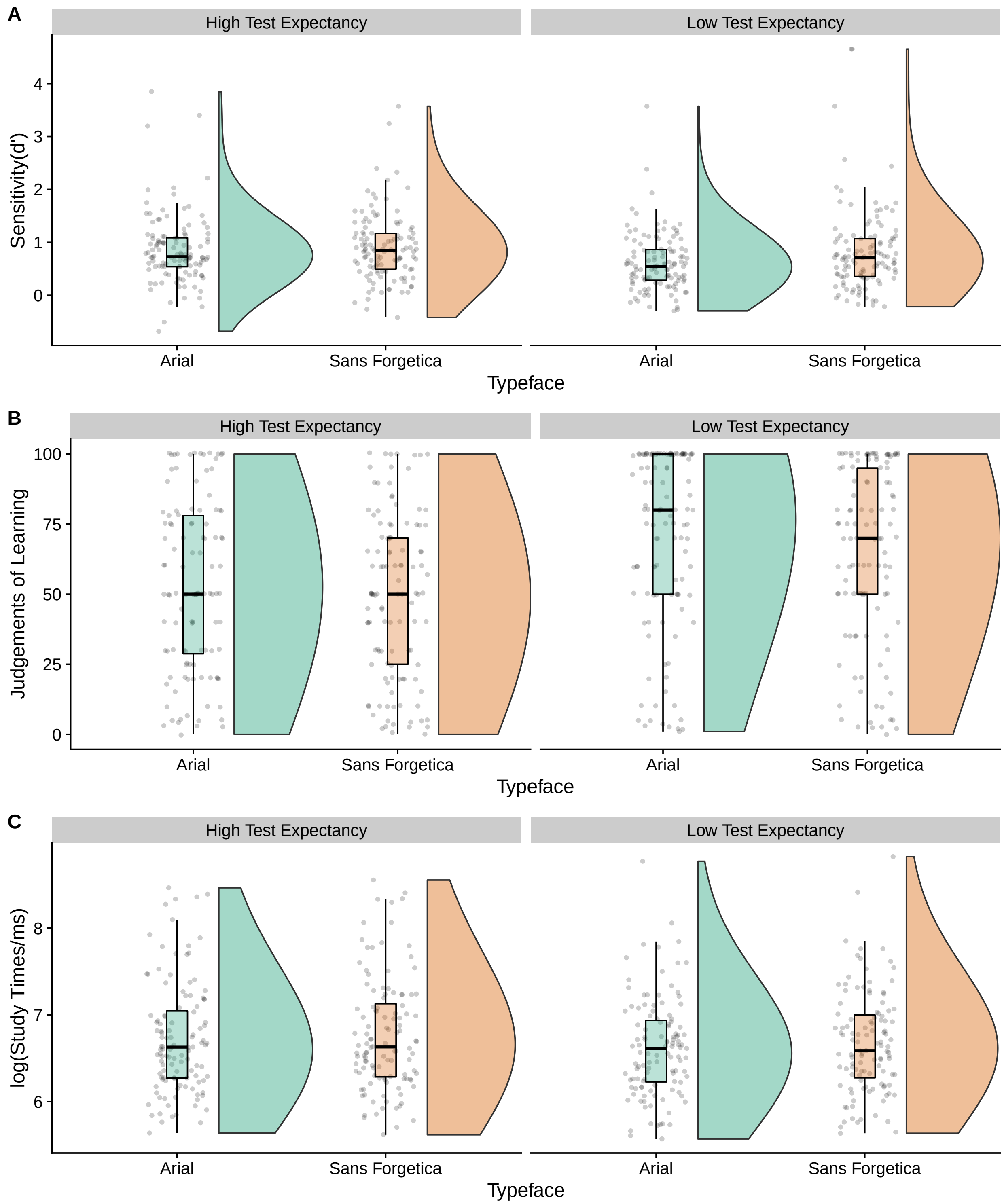

Figure 1. Raincloud plots (Allen et al., 2019) depicting raw data (dots), box plots, and half violin kernel desntiy plots.A.Memory sensitivity (d') as a function of Typeface and Testing Expectancy. B. Judgements of Learning as a function of Typeface and Test Expectany. C. Study times (log transformed) 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

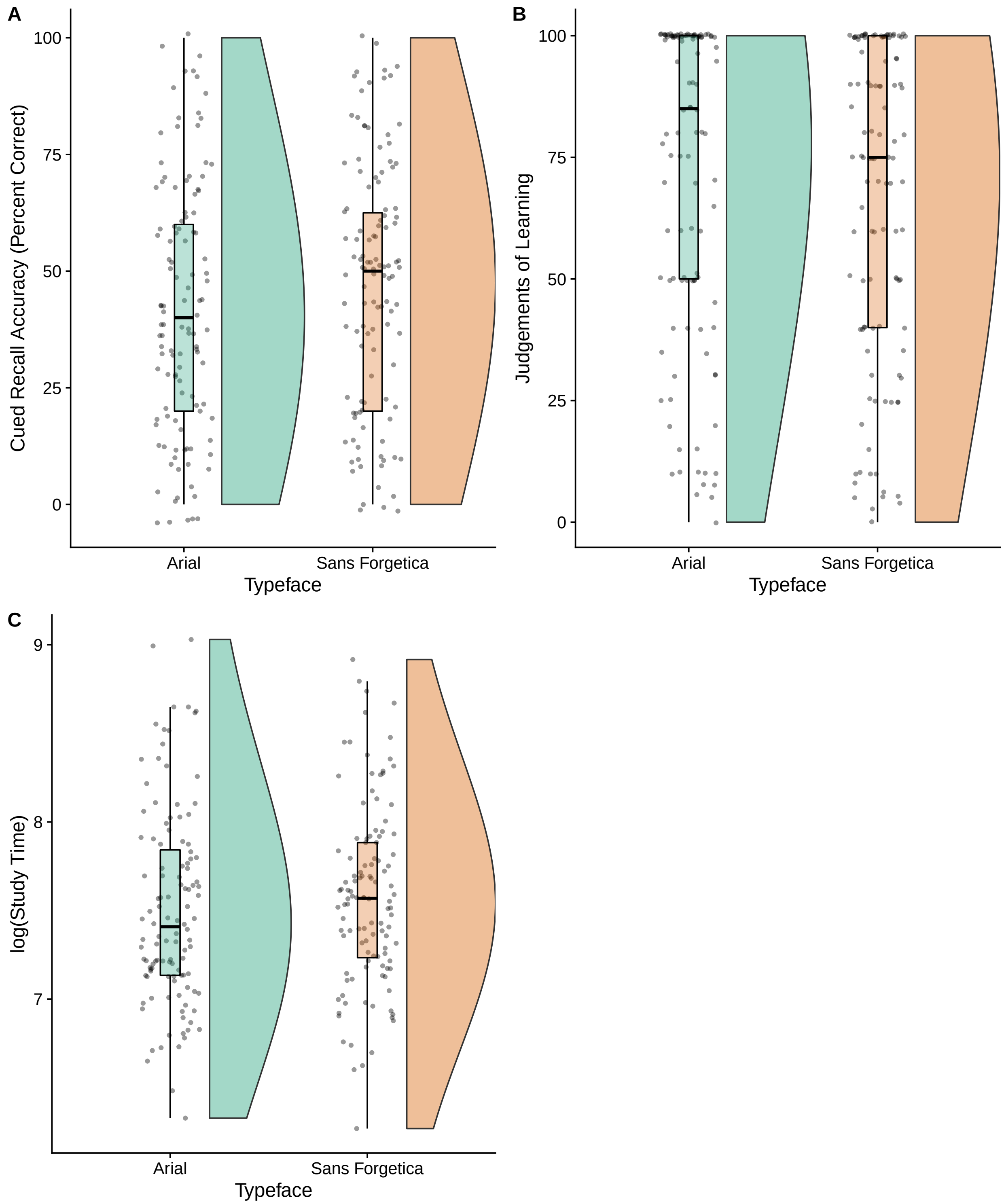

Figure 2. Raincloud plots (Allen et al., 2019) depicting raw data (dots), box plots, and half violin kernel desntiy plots.A.Memory sensitivity (d') as a function of Typeface and Testing Expectancy. B. Judgements of Learning as a function of Typeface and Test Expectany. C. Study times (log transformed) 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