Is This Going to Be on The Test? Test Expectancy Moderates the Disfluency Effect with Sans Forgetica

Jason Geller[1,2] & Daniel Peterson[3]

[1] University of Iowa
[2] Rutgers University Center for Cognitive Science
[3] Skidmore College

Presenting information in a perceptually disfluent format sometimes enhances memory. Recent work examining one type of perceptual disfluency manipulation, Sans Forgetica typeface, has yielded discrepant findings; some studies find support for the idea that the novel, disfluent typeface improves memory while others do not. To explore this discrepancy, the current study examined a boundary condition that determines when disfluency is and is not beneficial to learning. Specifically, we investigated whether knowledge about an upcoming test (high test expectancy) versus not (low test expectancy) helps clarify when mnemonic benefits arise for perceptually disfluent stimuli. In Experiment 1 (preregistered, N = 231), we found that Sans Forgetica is a memory-improving desirable difficulty, but only when there was no expectation of a final test. In Experiment 2 (preregistered N = 232), we conceptually replicated these results using a cued recall test. In Experiment 3 (preregistered, N = 232), we ruled out a time-on-task explanation for these outcomes while replicating the results of Experiment 2. Though these data provide some evidence of San Forgetica's mnemonic benefits, caution should be taken in interpreting these results. Not only were the effect sizes moderate in size, but low testing expectancy may not be realistically achievable in actual educational contexts. Though more research is warranted, we echo our prior arguments that students wanting to remember more and forget less should stick to other, more empirically supported desirable difficulties shown to enhance memory.

Keywords: Disfluency, Desirable Difficuties, Recognition, Recall
Word count: 9875

Imagine if you could remember more and forget less just by making the perceptual features of to-be-learned material harder. While this runs counter to the widely held belief that learning should be fluent (easy) and errorless (Pan et al., 2020), the concept of *desirable difficulties* (E. L. Bjork & Bjork, 2011) indicates that making encoding more disfluent (hard) and error-prone can sometimes help learners process the information more deeply and make it more likely they will retrieve the information at a later time. This general finding has been shown across a wide variety of encoding contexts (e.g., spacing and interleaving, Shana K. Carpenter, 2014). One provocative line of research that has piqued the interest of researchers and the media is the influence of extraneous factors, such as the perceptual format of to-be-learned material (e.g., size, font/typeface, or clarity), on memory. In some cases, making to-be-learned material perceptually disfluent (hard to read) has been shown to be desirable for memory—a phenomenon dubbed the perceptual interference effect (Nairne, 1988), or as it will be called henceforth, the perceptual disfluency effect (Geller et al., 2018). While perceptual disfluency has the potential to be valuable (and easy to implement), a recent meta-analysis has called into question whether perceptual disfluency is really desirable for learning (H. Xie et al., 2018, c.f., Weissgerber et al., in press). The current research aims to investigate under what conditions disfluency is and is not beneficial for learning using Sans Forgetica as a proxy for perceptual disfluency.

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

Sans Forgetica

A typeface known as Sans Forgetica has garnered increased attention in the media as a way to stave off forgetting and enhance memory. A typeface developed by a team of psychologists, graphic designers, and marketers, Sans Forgetica consists of intermittent gaps and back-slanted letters (see Figure 1 for an example; Earp, 2018). The disfluent perceptual characteristics are thought to provide the optimal level of disfluency to produce a desirable effect on memory. This has led to extensive press coverage from major news outlets (e.g., NpR, Washington post), and to the development of browser extensions and OS applications that allow users to place content in the novel typeface. The question, of course,

is whether Sans Forgetica merits such attention. As Carl Sagan famously said, "Extraordinary claims require extraordinary evidence" (Sagan, 1980).

Two recent studies provide some initial evidence against the aforementioned claim. Taylor et al. (2020) and Geller et al. (2020) set out to examine whether Sans Forgetica is really desirable for learning. In one of the first studies to look at the mnemonic benefits of Sans Forgetica (N = 882 across 4 experiments), Taylor et al. (2020) found that while Sans Forgetica was perceived as more disfluent by participants (Experiment 1) there was no evidence that it yielded a mnemonic boost in cued recall with strongly related cue-target pairs (Experiment 2) compared to a fluent (Arial) typeface, or when learning simple prose passages (Experiments 3-4). Shortly after the publication of this paper, Geller et al. (2020) contributed to the debate with three preregistered experiments (N = 820) finding, similar to Taylor[2020], Sans Forgetica did not enhance memory for weakly related cue-target pairs (Experiment 1), a complex prose passage (Experiment 2), or a yes/no recognition memory test (Experiment 3). Taken together, two independent laboratories conducting seven experiments with well over 1500 participants make for a compelling argument that there is little, if any, evidence that Sans Forgetica qualifies as a desirable difficulty.

Effects of Perceptual Disfluency on Learning

While there is evidence that Sans Forgetica does not enhance memory, there is a growing literature suggesting that other types of perceptual disfluency can improve learning. In a seminal study, Diemand-Yauman et al. (2011) used difficult-to-read fonts (i.e., Comic Sans, Bodoni MT, Haettenschweiler, Monotype Corsiva) and found those fonts enhanced learning and retention in both the laboratory (Experiment 1) when learning about space aliens, and in the classroom (Experiment 2) where students studied powerpoints in difficult fonts across several different content areas (i.e., Ap English, Honors English, Honors physics, Regular physics, Honors US History, and Honors Chemistry). Since then, there have been a number of follow-up studies showing a positive effect of disfluency with a wide array of perceptual manipulations such as high-level blurring Sungkhasettee et al. (2011), handwritten cursive (Geller et al., 2018), and other unusual or difficult-to-read fonts (Weissgerber & Reinhard, 2017; Weltman & Eakin, 2014).

However, there is not uniform support for this idea. For instance, Rhodes and Castel (2008) showed that words in a smaller-sized font (18 point) were judged as being more disfluent compared to words printed in a larger-sized font (48 point), but the smaller font did not lead to better memory — recall differences between the smaller and larger size fonts were negligible (see Hunter Ball et al., 2014; Kornell et al., 2011; Mueller et al., 2014; Susser et al., 2013, for similar failures to replicate the font size effect; but see Halamish, 2018 for moderating conditions of the font size effect). In another study, Yue et al. (2013) examined the perceptual disfluency effect using a low-level (minimal) blur manipulation. They examined the effect of blurring across several factors: type of task (recall vs. recognition), study duration (500 ms vs. 2 s), and design (within- vs. between-item lists). None of their experiments revealed a memory benefit for low-level blurring (but see Rosner et al., 2015, for evidence with a high-level blurring manipulation). Failures to replicate the disfluency effect also extend to many other types of perceptual manipulations (e.g., hard-to-read fonts, Magreehan et al., 2016; hard-to-hear auditory information, Rhodes & Castel, 2009) and more complex learning situations (e.g., in the classroom, Shana K. Carpenter et al. (2013); longer learning materials; Rummer et al. (2016); Strukelj et al. (2016)].

Complicating matters even further, in some instances, perceptual disfluency can harm learning. Yue et al.[2103, Experient 1a and 1b], found that a low-level blurring manipulation hurt recall compared to a clear, normal, font. Similarly, in the aforementioned Taylor et al. (2020) exploration of Sans Forgetica, outcomes from Experiment 2 suggested not only was the novel typeface not beneficial for learning, it actually impaired memory for briefly presented (500 ms) cue-target pairs.

Because of these mixed findings, a number of studies have begun to more specifically investigate those conditions under which perceptual disfluency does and does not enhance learning. Lehmann et al. (2016), for example, observed perpetually disfluent fonts only improved learning for individuals with high working memory capacity. Further, Geller et al. (2018) demonstrated that the level of perceptual disfluency matters. Using handwritten cursive, they varied the disfluency level of cursive (i.e., easy-to-read and hard-to-read). They found that cursive stimuli (overall) produced better memory (this memory benefit occurred in blocked and mixed designs and over a 24-hour retention interval). However, in a small-scale meta-analysis they observed an inverted U-shaped pattern wherein easy-to-read cursive produced better memory than type-print and hard-to-read cursive, despite the hard-to-read cursive being more disfluent. This suggests that not all disfluency manipulations are created equal; there is an optimal level of disfluency (also see Seufert et al., 2017). Finally, Weissgerber and Reinhard (2017) found that time of test influences whether disfluency enhances memory. They used hard-to-read fonts and tested participants at two time points spaced two weeks apart. On the immediate test, hard-to-read font did not produce better memory compared to transposed-letter (e.g., jugde for judge) and normal font conditions. At the second time point, however, material in a hard-to-read font produced less forgetting than the other two conditions suggesting that there might be a disfluency sleeper

effect of sorts, where the benefits of perceptual disfluency are seen only after a longer retention interval (Oppenheimer & Alter, 2013).

Theoretical Accounts of Perceptual Disfluency

Despite these null (and sometimes negative) effects, the positive findings reported suggest that under some conditions perpetual disfluency can be desirable for learning. What is the proposed mechanism underlying such an effect? The perceptual disfluency effect can be explained against the backdrop of traditional dual process (e.g., System 1 and System 2; Evans, 2016), depth of processing (Craik & Lockhart, 1972), and metacognition models. The most popular account is the metacognitive account of perpetual disfluency [Alter (2013); Alter et al. (2007); Diemand-Yauman et al. (2011)). This account refers to the idea that the difficulty encountered during encoding, as a result of perceptual disfluency, forces more System 2 processing, which is slow, effortful, and deep. What is critical here is not the objective disfluency of the material, but the subjective disfluency—that is, the experience of disfluency. It is the experience of disfluency that is hypothesized to stimulate metacognitive processes (monitoring and control) which serves to strengthen memory. It is also important to note that this account does not differentiate between disfluency manipulations (see Weissgerber et al., 2017). That is, anything that is perceived as disfluent should engender better memory.

An alternative account is the compensatory processing account (Hirshman et al., 1994; Mulligan, 1996). The compensatory processing account is heavily influenced by a classic model of word recognition—the interactive activation model (McClelland & Rumelhart, 1981). Within the compensatory processing account, the disfluency effect is tied to processes occurring during the word identification process. Specifically, difficulty in identifying a stimulus increases the amount of top-down feedback from a higher-level (i.e., lexical/semantic) to a lower-level (i.e., features and orthography). Strong evidence for this account comes from studies using masking to impede word recognition. Masking involves presenting a word very quickly (100 ms) and masking it with either forward or backward hashmarks (Nairne, 1988). The rapid presentation of the word along with the presentation of the mask renders visual information insufficient to recognize the word correctly, leading to greater higher-level processing. It is this feedback that results in better memory for stimuli. While more research is needed on the mechanism(s) of perceptual disfluency, it is clear that both the metacognitive account and compensatory processing account emphasize the importance of higher-level semantic or metacognitive processes in producing the positive effects of perceptual disfluency on memory.

Disfluency and Sans Forgetica: A potential Moderator

The literature reviewed above provides ample evidence that presenting materials in perceptually degraded formats can enhance memory and learning outcomes and act as a desirable difficulty, but also that the effect may be fickle. This has led to the exploration of different moderating or boundary conditions of the perceptual disfluency effect.

Related to the current research, a recent publication demonstrated that Sans Forgetica may indeed be optimal for learning, but only when spelling ability is taken into account. Eskenazi and Nix (2020) had participants learn the spelling and meaning for low-frequency words presented in sentences while their eye movements were being recorded. For half the participants, the to-be-learned material was presented in Sans Forgetica while for the other half, it was presented in a more fluent (Courier) typeface. 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. Critically, the authors reported that Sans Forgetica was indeed perceptual disfluent (i.e., the gaze duration was longer in the Sans Forgetica condition) and that it had a positive effect on memory for words and their meanings. However, spelling ability moderated this effect: only good spellers benefited from Sans Forgetica.

While spelling ability could moderate the mnemonic benefit of Sans Forgetica, there is another possibility. Probing into the design features of Eskenazi and Nix (2020), one critical difference between their design and a recent failure to replicate (Geller et al., 2020) was testing expectancy. Eskenazi and Nix (2020) surprised participants with the orthographic and semantic tests whereas participants in Geller et al. (2020) were explicitly told their memory was going to be assessed. In fact, one common feature of studies showing a desirable effect of perceptual disfluency on memory is low testing expectancy [e.g., Geller et al. (2018); Hirshman and Mulligan (1991); Mulligan (1996); Hirshman et al. (1994), Westerman and Greene (1997); but see Rosner et al. (2015), Experiment 3A; Sungkhasettee et al. (2011)). Accordingly, it is important to examine the role of testing expectancy in relation to the perceptual disfluency effect and Sans Forgetica.

Testing expectancy is known to exert a positive influence on 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 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 were unable to find an overall disfluency effect, nor did they find evidence that test expectancy moderated the disfluency effect. Following up on this, Geller and Still (2018), with a stronger perceptually disfluent manipulation (i.e., masking), demonstrated that testing expectancy can moderate the disfluency effect. Looking at the impact of item-by-item judgments of learning (JOLs) and list-wide JOLs, which are normally confounded with test expectancy, they found that under conditions where there was low testing expectancy and list-wide JOLs were used, a disfluency effect appeared. Given this, it is possible, then, the failure to find some disfluency effects (such as with Sans Forgetica) might only arise under low test expectancy. The proposed experiments more directly test this hypothesis.

The Current Experiments

The empirical work reported here was designed to investigate the effect of Sans Forgetica on memory for words and whether observation of a perceptual disfluency effect depends on testing expectancy. To this end, the present article focused on the procedures used by Geller et al. (2020) and Eskenazi and Nix (2020) with an eye towards those features on which the two studies methodologically differed. Namely, the present studies attempted to examine if perceptual disfluency is really a desirable difficulty, but is countervailed by other memory influences, such as testing expectancy, which might negate the effect. If testing expectancy is found to moderate the disfluency effect, it would have important theoretical implications as it would provide an important moderating factor for researchers doing work in this domain. Further, it would support accounts suggesting that encoding difficulty brought forth by perceptual disfluency arises from an attentional mechanism that leads to deeper, more effortful, processing. Conversely, if we do not find a disfluency effect with Sans Forgetica that would also be useful from a theoretical perspective. The failure to find a disfluency effect would further drive the nail into the coffin of perceptual disfluency as a desirable difficulty. To this end, the current research aims to examine testing expectancy as a potential boundary condition of the disfluency effect in recognition memory and cued recall using Sans Foregtica.

Experiment 1

In Experiment 1 we examined whether the impact of Sans Forgetica on memory is moderated by test expectancy. Using an old/new recognition test we manipulated testing expectancy by alerting only half the participants that their memory was to be assessed. In addition, we collected list-wide JOLs (a subjective general prediction for each typeface that assesses future memory performance) and study times as a manipulation check to ensure Sans Forgetica is perceptually disfluent. The choice to use list-wide JOLs was largely influenced by recent findings suggesting a reactive effect of JOLs on memory (Janes et al., 2018; Myers et al., 2020; Soderstrom et al., 2015). The very act of making a JOL for each world mitigates the beneficial effects of perceptual disfluency on memory (Besken & Mulligan, 2013).

In our preregistration, we predicted an interaction between Typeface (Arial vs. Sans Forgetica) and Test Expectancy. Specifically, we anticipated seeing a memory boost for Sans Forgetica, but only under low test expectancy (vs. high test expectancy). This was based on previous studies demonstrating perceptual disfluency effects under low test expectancy (e.g., Geller et al., 2018; Hirshman & Mulligan, 1991; Mulligan, 1996), but not under high test expectancy (e.g., Geller et al., 2020). Finding a null effect of perceptual disfluency in the high test expectancy group would replicate the findings from Geller et al. (2020; Experiment 3). Further, we predicted that we would not see JOL differences as a function of Typeface or Testing Expectancy. Finally, with respect to study times, we predicted we would see longer study times for Sans Forgetica, but only in the low test expectancy group.

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 for Experiments 1 can be found at https://osf.io/cqp6s/.

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 et al., 2020). The targeted sample size was based off a previous experiment (Geller et al., 2020), Experiment 1), wherein we calculated power to detect a medium sized interaction effect (d = 0.35) using a similar design to the current study. Data collection resulted in the collection of 231 participants. participants were compensated for their time. We used prolific's costume prescreening measures and included participants that were native English speakers, from the United States, had an approval rating between 80% and 100%, and did not participate in any prior studies conducted by the researchers. ### Materials

Stimuli included 188 single-word nouns taken from Geller et al. (2018). All words were from the English Lexicon project database (Balota et al., 2007). We controlled for both word frequency (all words were high frequency; mean log HAL

frequency = 9.2) and length (all words were four letters). 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) × 2 (Testing Expectancy: High vs. Low) mixed analysis of variance (ANOVA).

Procedure

Similar to Geller et al. (2020; Experiment 3), a total of 188 words were divided across four lists (94 words each; 47 in each typeface condition). This was done so each word appeared in each 2 (old/new) x 2 (Arial/Sans Forgetica) condition. This ensured that each word served equally often as a target and a foil in both typefaces across participants. In the first two lists, 94 words were chosen to be "old" (47 in Arial and 47 in Sans Forgetica) and 94 words were chosen to be "new" (47 presented in Arial and 47 presented in Sans Forgetica) and were only presented during the test phase. In the other two lists, items presented as "new" were presented as "old" and vice versa. 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.

Participants were randomly assigned to one of two groups: the high test expectancy group, or the low test expectancy group. 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). Specifically, those in the high test expectancy group received the following study description: "In this study your memory will be tested for words in different typefaces. In the first part, you will study words. In the second part, your memory will be tested for the words you studied." They were also explicitly told before the experiment that their memory for the words were going to be assessed. In the low test expectancy group, participants received a different study description: "In this study you will be reading words in different typefaces." Further, the experiment instructions before the experiment made no mention of any memory test.

The experiment consisted of four phases: encoding phase, JOL phase, distractor phase, and test phase. During the encoding 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. Though the presentation of the words was a single, heterogeneous mix of Arial and Sans Forgetica words, the JOL phase required them to provide two list-wide JOLs wherein they retrospectively judged on a scale from 0 (not at all likely)-100 (most likely) how successful they would be recalling, as a whole, words presented in Arial and Sans Forgetica. Then, during a three-minute distractor, participants wrote down as many US state capitals as possible. Finally, participants were given 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 of the new words was presented in either Arial typeface or Sans Forgetica typeface. All words were individually randomized for each participant during both the study and test phases and progress was self-paced. After the experiment, participants were debriefed. The entire experiment lasted approximately 15 minutes.

Analysis Plan

For all experiments reported in this paper, we employed a 2 × 2 mixed analysis of variance (ANOVA). We report a variation of Cohen's d (d_{avg}; Buchanan, De Deyne, et al., 2019) and generalized eta-squared (η_g^2; Olejnik & Algina, 2003) as measures of effect size. Alongside traditional analyses that utilize null hypothesis significance testing (NHST), we also report the Bayes Factor (BF) for reported null effects. As a rule of thumb, BFs greater than or equal to 3 provide substantial evidence, while BFs greater than or equal to 10 provide strong evidence for one model over another model (Jarosz & Wiley, 2014). 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 & Rouder, 2018a). All figures were generated using ggplot2 (vers. 3.3.0; Wickham, 2016a). See the appendix for a list of all R packages used.

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 the ceiling or floor were changed to .99 or .01. Figure 1a presents d' values along with difference scores (Figure 1b). The analysis revealed that when told about a memory test, participants had better discriminatory ability than those not told about a memory test, $M_{\text{diff}} = 0.16$, $F(1, 229) = 4.11$, $p = .04$, $\eta_g^2 = .014$. Individuals were better at discriminating target words presented in Sans Forgetica than Arial, $M_{\text{diff}} = 0.12$, $F(1, 229) = 10.73$, $p = .001$, $\eta_g^2 = .010$. This was qualified by an interaction between Test Expectancy and Typeface, $F(1, 229)$

= 4.34, , p = .038, η_g^2 = .004. planned comparisons showed that individuals in the low test expectancy group had better recognition memory for words presented in Sans Forgetica compared to Arial, $F(1, 229)$ = 14.297, p < .001, d_{avg} = 0.31. In the high test expectancy group, there was substantial evidence for no difference between typefaces, $F(1, 229)$ = 0.716, p = .398, d_{avg} = 0.07, BF_{01} = 5.83.

JOLs

JOL responses are presented in Figure 1c along with difference scores (Figure 1d). We excluded seven participants for not providing JOLs to each typeface.Using the same model as above, participants in the high testing expectancy group gave higher JOLs than the low testing group, M_{diff} = 16.2, $F(1,221)$ = 16.01, p < .001, η_g^2 = .065. Arial elicited higher JOLs than Sans Forgetica, M_{diff} = 4.0, $F(1,221)$ = 27.05, p < .001, η_g^2 = .004. There was no interaction between Testing Expectancy and Typeface, $F(1,221)$ = 0.13, p = .715, η_g^2 < .001. Compared to a main effects-only model, there was substantial evidence for no interaction (BF = 7.28).

Study Times

Although not preregistered, study times less than 150 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.[1] Given the heavy positive skew of the data, we log-transformed study times to better approximate a normal distribution (see Fig.1e). Evidence for testing expectancy effects on log-transformed study times were inconclusive, $F(1,229)$ = 1.97, p = .162, η_g^2 = .008, BF = 1.822. Typeface did influence study times: study times were slower for Sans Forgetica than Arial, $F(1,229)$ = 30.91, p < .001, η_g^2 = .001. There was no interaction between Testing Expectancy and Typeface, $F(1,229)$ = 1.10, p = .296, η_g^2 < .001. Compared to a main effects-only model, there was substantial evidence that there was no interaction between Testing Expectancy and Typeface (BF = 5.25).

As predicted, memory sensitivity for Sans Forgetica was higher when testing expectancy was low, but not when testing expectancy was high. High test expectancy could explain why Geller et al. (2020; Experiment 1) failed to find a disfluency effect with Sans Forgetica. We also found subjective and objective evidence that Sans Forgetica is in fact perceptually disfluent. Participants gave lower JOLs to stimuli studied in the Sans Forgetica typeface, regardless of test expectancy. That is, not only did the novel typeface improve recognition memory, but participants also subjectively rated it as an inferior context for word learning. These findings are inconsistent with the predictions preregistered 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 is in line 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)). In relation to study times, 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 of Sans Forgetica on study times. It is quite possible that not correcting for the skew of raw data or omitting outliers lead to the null effect of study time observed in Geller et al. (2020, Experiment 3). Indeed, reanalyzing the study time data from Geller et al., (2020, Experiment 3) with a similar procedure outlined above showed larger study times for Sans Forgetica (p = .049, one-tailed).

The finding that test expectancy moderates the disfluency effect in recognition contradicts a finding from Rosner et al. (2015) (Experiment 3a). In that particular experiment, they used a high-level blurring manipulation and manipulated test expectancy, but did not find the critical interaction. Given the novelty of the current findings, in Experiment 2, we attempted to replicate this pattern of results using a different criterion test: cued recall

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

In Experiment 2, we used weakly related cue-target pairs from Geller et al. (2020; Experiment 1). In that experiment, participants were told about the upcoming memory test, and there was strong evidence against there being a Sans Forgetica effect (BF > 100). In the present experiment, we set out to examine whether this null effect persists regardless of test expectancy. That is, when test expectancy is low, will we again observe a Sans Forgetica effect with cued recall?

[1] The decision to omit these observations did not meaningfully impact any of the conclusions reported here.

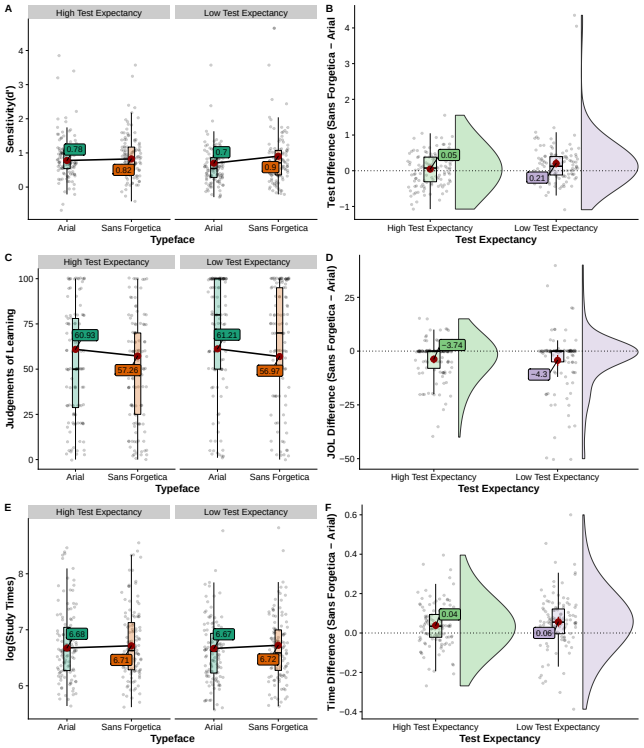

Figure 1

A. Participant accuracy (dots), box plots (medians and interquartile ranges), and label means for memory sensitivity (d') as a function of Typeface and Testing Expectancy in Experiment 1. B. Raincloud plots (Allen et al., 2019) for difference scores, with labeled means and bootstrapped 95% CIs, as a function of Test Expectancy in Experiment 1. C. Participant accuracy (dots), box plots (medians and interquartile ranges), and labeled means for JOLs as a function of Typeface and Testing Expectancy in Experimenr 1. D. Raincloud plots (Allen et al., 2019) for JOL difference scores, with labeled means and bootstrapped 95%CIs, as a function of Test Expectancy in Experiment 1. E. Participant accuracy (dots), box plots (medians and interquartile ranges), and labeled means for study times (log-transformed) as a function of Typeface and Testing Expectancy in Experiment 1. F. Raincloud plots (Allen et al., 2019) for study time difference scores, with labeled means and bootstrapped 95% CIs, as a function of Test Expectancy in Experiment 1

Methods

The preregistered analysis plan for Experiment 2 can be found here: https://osf.io/3xak9. All raw and summary data, materials, and R scripts for pre-processing, analysis, and plotting for Experiment 2 can be found at https://osf.io/cqp6s/.

Participants

We preregistered and collected a sample size of 232 participants. Participants were recruited on Amazon's Mechanical Turk (MTurk) platform, all of whom completed the experiment through Pavlovia (Pavolvia.org). In order to participate in the study, participants had to be native-English speakers, live in the United States, and no record of participating in previous studied offered by the researcher.

Design

Per our pre-registration, accuracy, JOLs, and study times were analyzed with a mixed factorial design with typeface (Arial vs. Sans Forgetica) manipulated within-participants and test expectancy (High vs. Low) manipulated between participants.

Materials and procedure

Materials and Procedure Experiment 2 was programmed in PsychoPy (Peirce et al., 2019) and hosted on Pavoliva (Pavolvia.org). The materials were adopted from Geller et al. [2020, Experiment 1; also see Shana K. Carpenter et al. (2006)). Participants were presented with 24 weakly related cue-target pairs. The pairs were all nouns, 5–7 letters and 1–3 syllables in length, high in concreteness (400–700), high in frequency (at least 30 per million), and had similar forward ($M = 0.031$) and backward ($M = 0.033$) association strengths. Two counterbalanced lists were created for each testing condition (high and low-test expectancies) so that each target could be presented in each typeface condition (Arial vs. Sans Forgetica) without repeating any items for an individual participant.

A version of the experiment can be run by following the following link: https://run.pavlovia.org/Jgeller112/sf_low_cb1. The experiment consisted of four phases: encoding phase, JOL phase, distractor phase, and test phase. Similar to Experiment 1, some participants were told about an upcoming memory test while others were not. During the encoding phase, each participant was presented with a series of word pairs randomly, one at time with the cue always presented in Arial on the left hand side and the target word presented in either a disfluent typeface (Sans Forgetica) or a fluent typeface (Arial), on the right hand side. Typefaces of the target words were randomly intermixed. The encoding phase was self-paced: Participants were instructed to press a button of the screen after reading each word. Like Experiment 1, participants then made two list-wide JOLs. Following a short distractor task (3 min), participants were given a cued recall test which began with instructions for the test. Each trial started with the presentation of a cue from the encoding phase, in lowercase letters, to participants one at a time. Participants were instructed to type in the corresponding target (or guess if they could not remember). The test phase was self-paced.

All cues were presented in Arial font. The entire experiment lasted approximately 10 minutes.

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 threshold of 80% was used to determine whether a typed response was correct or not.

Results and Discussion

Cued Recall

Figures 3a shows performance in the cued-recall test (Figure 3a) along with difference scores (Figure 3b). Participants in the high test expectancy group performed better than participants in the low test expectancy group, M_{diff} = 20%, F(1, 230) = 38.26, p < .001,η_g^2 = .126. Participants recalled more target words in Sans Forgetica than Arial, M_{diff} = 5%, F(1, 230) = 13.57, p < .001, η_g^2 =.008. This was qualified by an interaction between Test Expectancy and Typeface, F(1, 230) = 10.74, p = .001, η_g^2 = .006. A Bayesian analysis revealed that the interaction model was strongly preferred to the full model (BF = 21.77). Planned comparisons showed that individuals in the low test expectancy group recalled more words presented in Sans Forgetica than Arial, t =4.92, p < .001, d_{avg}= 0.38; In the high test expectancy group, there was substantial evidence that there was no difference between Sans Forgetica and Arial, t = 0.287, p = .778, d_{avg} = 0.02, BF_{01} = 9.31.

JOLs

Figures 3c and 3d show JOLs (Figure 3b) as well as difference scores (Figure 4c). Using the same model as above, participants in the high test expectancy group gave higher JOLs than the low test expectancy group, M_{diff} = 5.91, F(1,229) = 13.57, p < .001, η_g^2 = .028. Arial elicited higher JOLs than Sans Forgetica, M_{diff} = 15.15, F(1,229) = 87.05, p < .001, η_g^2 = .161. There was an interaction between Testing Expectancy and Typeface, F(1,229) = 13.65, p < .001, η_g^2 < .029. A Bayesian analysis revealed that the interaction model was strongly preferred to the main effects-only model (BF > 100). Planned comparisons revealed that the JOL effect was larger in the low test expectancy group (d_{avg} = 1.21) than in the high test expectancy group (d_{avg} = 0.72).

Study Times

Figures 3e and 3f show log-transformed RTs (Figure 3e) and difference scores (Figure 4f). Like Experiment 1, we excluded study times less than 150 ms and study times greater than 2.5 SD above the mean per condition for each participant. The outlier procedure removed ~ 2% of the data. Study times were overall larger for the high test expectancy group compared to the low test expectancy group, M_{diff} = 0.34, F(1,230) = 17.02, p < .001, η_g^2 = .068. Cue-target pairs yielded larger study times for Sans Forgetica compared to Arial, M_{diff} = 0.06, F(1,230) = 27.74, p < .001, η_g^2 = .002. There was no interaction between Testing Expectancy and Typeface, F(1,230) = 0.39, p = .533, η_g^2 < .001. A main effects-only model was strongly preferred over the interaction model (BF = 6.03).

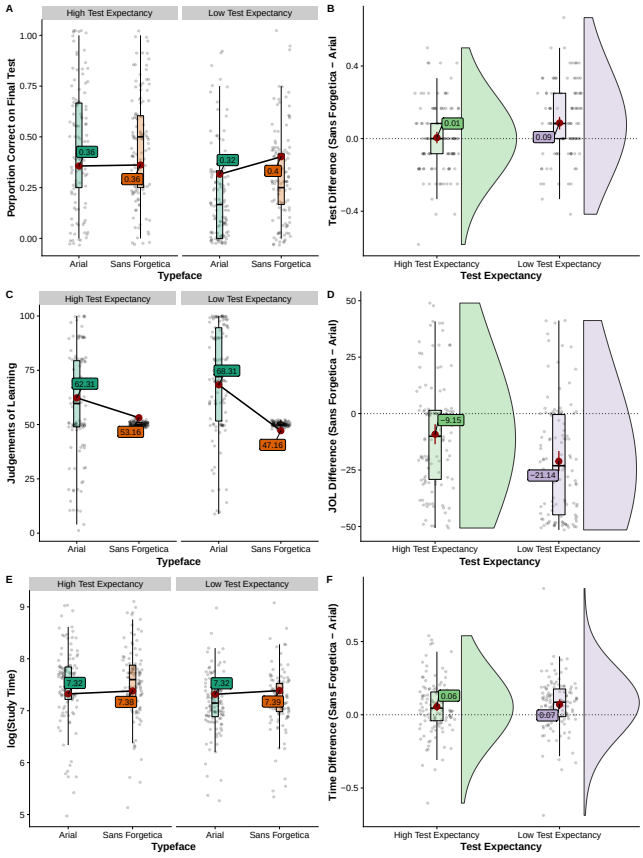

Figure 2

The results complement those from Experiment 1 well and suggest the disfluency effect with Sans Forgetica is not unique to a particular criterion test. Using cued recall, we once again demonstrated that Sans Forgetica can constitute a desirable difficulty, but only when test expectancy is low. Importantly, the effect we observed was rather modest; Sans Forgetica conferred roughly a 5% increase in cued recall performance above and beyond Arial, a more fluent typeface. When looking at the low test expectancy group alone we observed a 9% increase. Furthermore, we once again showed longer study times and lower JOLs for words studied in Sans Forgetica. There are, however, a couple points of divergence that merit mention. Compared to Experiment 1, JOLs for the Sans Forgetica condition were tightly bunched around the middle of the response scale.[2] This could reflect uncertainty

[2]Experiments 2 and 3 used a slider scale that ranged from 0-100

around how well participants remember Sans Forgetica target words. Additionally, study times for Sans Forgetica were longer for the high test expectancy group compared to the low test expectancy group. This most likely reflects participants studying word pairs longer in preparation for an upcoming test.

Experiment 3

In Experiments 1 and 2, we observed a benefit for Sans Forgetica under low test expectancy. Although this result constitutes an example of a desirable difficulty effect as a result of perceptual disfluency, the mechanisms underlying such effects remain an open issue. Our preferred interpretation is that encoding difficulty from the typeface is an attentional response eliciting deeper processing that in turn leads to better remembering. However, another possible explanation is that Sans Forgetica is remembered better simply because participants spend more time processing them, as indexed by slower study times during encoding in both Experiments 1 and 2.[3].

To examine if time-on-task can account for the desirable effect of Sans Forgetica on recall, we manipulated time spent encoding by having participants either encode stimuli at their own pace (self-paced), or by removing control over how long the stimuli were studied for. If time-on-task moderates the Sans Forgetica effect, we expect there to be no effect on memory when time is constrained to be equal between Arial and Sans Forgetica. However, when encoding is self-paced, we would expect there to be better memory for Sans Forgetica compared to Arial. Corroborating this, Kühl et al. (2014) showed that self-paced study produced better learning outcomes compared to constrained study time. Because of this, we hypothesized that we would observe a disfluency effect for Sans Forgetica only when study time was self-paced.

In Experiment 3 we choose to keep testing expectancy low and manipulate time-on-task (self-paced vs. 3 s)[4]. This design also served to replicate the novel findings from Experiment 2 showing that low test expectancy is essential for the Sans Forgetica memory effect. In addition, we examine list-wide JOLs. Due to the experiment design, we could not analyze study times as they could only be collected in the self-paced group.

Methods

The preregistration for this experiment can be found here: https://osf.io/hjnk5. All raw and summary data, materials, and R scripts for pre-processing, analysis, and plotting for Experiment 3 can be found at https://osf.io/cqp6s/.

Participants

We preregistered and collected a sample size of 232 participants. Participants were recruited on Prolific , all of whom completed the experiment through Pavlovia (Pavolvia.org). Using prescreening questionnaires on Prolific, we limited our sample to participants to those residing in the USA, native English speakers, and had no record of participating in previous studies by the first author.

Design, Materials, and Procedure

The design, materials, and procedure are identical to Experiment 2, with one exception—instead of manipulating test expectancy (no participants were informed of the impending memory test in Experiment 3), study time was manipulated (Self-paced vs. 3 s) between participants. In the self-paced group (like in Experiment 2), participants were given as long as they wanted to process the cue-target pairs. In the 3 s group, cue-target pairs were presented for 3 seconds.

Results and Discussion

Cued Recall

Figures 4a shows performance on the cued recall test (Fig. 4a) along with difference scores (Figure 4b). The analysis revealed that there was no reliable difference between the Self-paced and Timed groups on cued recall, M_{diff} = 2%, $F(1, 230) = 0.369$, $p < .544$, $\eta_g^2 = .055$. Individuals were better at recalling target words presented in Sans Forgetica than Arial, $M_{\text{diff}} = 5\%$, $F(1, 230) = 15.03$, $p < .001$, $\eta_g^2 = .013$. This was no interaction between Time on Task and Typeface, $F(1, 230) = 1.13$,$p = .289$, $\eta_g^2 < .001$. A Bayesian analysis revealed that a main effects-only model was preferred to the interaction (BF = 5.50).

JOLs

Figures 4c and 4d show participant-level JOLs (Figure 4c) and difference scores (Figure 4d). Using the same model as above, participants in the Timed group gave higher JOLs than in the Self-paced group, $M_{\text{diff}} = 5.91$, $F(1,230) = 17.43$, $p < .001$, $\eta_g^2 = .055$. Arial typeface elicited higher JOLs than Sans Forgetica typeface, $M_{\text{diff}} = 9.7$, $F(1,230) = 48.81$, $p < .001$, $\eta_g^2 = .048$. There was an interaction between Time on Task and Typeface, $F(1,230) = 27.17$, $p < .001$, $\eta_g^2 < .027$. A Bayesian analysis revealed that the interaction model was strongly preferred to the main effects-only model (BF = 57.24). Simple effects revealed that the JOL effect (Arial < Sans Forgetica) was larger in the self-paced group ($d_{\text{avg}} = 1.22$) than in the timed group ($d_{\text{avg}} = 0.10$; $\text{BF}_{01} = 1.466$).

in increments of 10 while Experiment 1 had participants type in a number between 0-100.

[3] A simple time-on-task account does a poor job of explaining the lack of a Sans Forgetica effect we observed in Experiments 1 and 2 when participants were told about a memory test.

[4] Three seconds was chosen by looking at overall study times for Experiment 2 (M = 2,192 ms). Given this, we thought 3 s would be more than sufficient to allow identification of the cue-target pairs

Taken together, the results from Experiment 3 are clear. Cued recall performance was better overall for Sans Forgetica—it did not matter if encoding was self-paced or timed. This contradicts a study by Kühl et al. (2014) showing that self-paced study produces better learning outcomes compared to constrained study time. It is important to note that our study used simple learning materials whereas Kühl et al., used more complex materials (i.e., multimedia slides about lightening construction). With more complex materials, a time limit might hurt rather than help recall. Despite this, the findings from Experiment 3 nicely replicated the findings from Experiment 2 under low test expectancy. From this, it is clear that a simple time-on-task account cannot explain these findings. A time on task account would predict better memory in the self-paced group because they can spend longer encoding each pair. While the interaction was not significant, looking at the effect sizes between groups, the disfluency effect was larger in the timed group (d_{avg} = 0.32) than the self-paced group (d_{avg} = 0.15). We return to this issue in the general discussion.

Turning to JOLs, we replicated the outcomes from Experiments 1 and 2 showing that participants judged Sans Forgetica as less memorable (lower JOLs). This difference was larger in the self-paced group than in the timed group. While the reason for this is not clear, one possible explanation could be that during self-paced encoding, individuals are more uncertain about whether they will remember disfluent targets because they were not restricted by a time limit and could advance at their own pace. This fact is highlighted by JOLs in that condition clustering around the middle point of the scale.

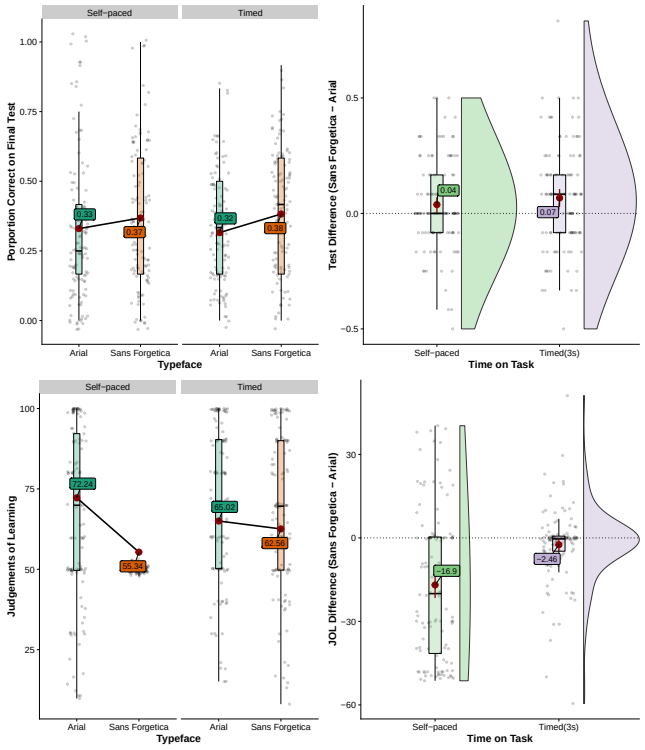

Figure 3

A. Participant accuracy (dots), box plots (with medians and interquartile ranges), and labeled means for cued recall as a function of Typeface and Time-on-Task in Experiment 3. B. Raincloud plots (Allen et al., 2019) for cued recall difference scores, with labeled means and bootstrapped 95% CIs, as a function of Time-on-Task in Experiment 3. C. Participant accuracy (dots), box plots (with medians and interquartile ranges), and labeled means for JOLs as a function of Typeface and Time-on-Task in Experiment 3. D. Raincloud plots (Allen et al., 2019) for JOL difference scores, with labeled means and bootstrapped 95% CIs as a function of Time-on-Task in Experiment 3

General Discussion

Sans Forgetica has garnered substantial attention from both the media and the scientific community as of late. The present experiments attempted to reconcile the mixed findings in the literature as it relates to Sans Forgetica and more broadly, perceptual disfluency. Following up on recent calls to examine boundary conditions of the perceptual disfluency effect (R. A. Bjork & Yue, 2016; Dunlosky & Mueller, 2016), we focused on one boundary condition: testing expectancy. To summarize, we found evidence that testing expectancy moderates the perceptual disfluency effect. Sans Forgetica produced lower JOLs and longer study times across (Experiments 1 and 2) and enhanced memory in recognition (Experiment 1) and cued recall (Experiment 2) only when participants were not told an upcoming memory test. Experiment 3 revealed this effect does not seem to be a solely mediated by time-on-task.

These outcomes conflict with some recent findings. First, Rosner et al. (2015, experiment, 3A) did not find a moderating role for test expectancy in recognition memory using a high-level blurring manipulation—low and high test expectancy elicited a similar benefit. Despite this, those findings have not been replicated, closely or conceptually. In the current set of experiments, we demonstrated a robust effect of test expectancy across different test formats (Experiments 1 and 2), and replicated the basic disfluency effect with low test expectancy (Experiment 3). One interesting possibility is that disfluency manipulations can have differential effects on memory. For instance, in Rosner et al., (2015; Experiment 4a), they were only able to show a desirable effect of blurring using a high-level blurring manipulation—they could not find a recognition benefit using a low-level blurring manipulation. Similarly, Geller et al. (2018) showed that easy-to-read cursive words produced stronger memory effects than

hard-to-read cursive words. Thus, the blurring manipulation used by Rosner et al. might have been a stronger cue than Sans Forgetica. An important avenue for future research would be examine different levels and types of perceptual disfluency and their role on memory.

Additionally, while we found a general benefit of Sans Forgetica under low test expectancy, Eskenazi and Nix (2020) only found a memory benefit for Sans Forgetica among those participants who were stronger spellers. Better spellers are thought to have a more precise mental lexicon which allows for more efficient processing at multiple levels of representation (i.e., orthographic, phonological, and semantic; Perfetti, 2007). When confronted with perceptual degradation, better spellers would be able to process a stimulus at a deeper level, which could give rise to better memory. The disparate findings can be reconciled by the fact that we used high frequency words in all three experiments. Presumably, these words were well known to the participant therefore allowing perceptual disfluency to be desirable for learning.

Perceptual Desirable Difficulty: A Time on Task Effect?

The result of primary interest here is that Sans Forgetica, a perceptually disfluent typeface, was associated with better recognition and recall, but only when test expectancy was low. It has been proposed that perceptual disfluency enhances memory as a result of deeper, more effortful, processing. A rather uninteresting alternative explanation is that an extended period of time dedicated to encoding is sufficient to enhance memory encoding. While a time on task account can explain the Sans Forgetica effect under low test expectancy, it is not adequate to explain some of the other findings. In both Experiments 1 and 2, Sans Forgetica produced longer study times, yet there was strong evidence that there was no perceptual disfluency effect in the high test expectancy group. Furthermore, in Experiment 3, where we directly tested time-on-task by having participants either encode cue-target pairs with a time-limit, or have encoding be self-paced, we found robust effects of perceptual disfluency on cued recall regardless of pacing. In addition, we found that the perceptual disfluency effect were larger under a time constraint than it was when encoding was self-paced. It is not clear how a time-on-task account would explain these findings.

In addition, a simple time-on-task account has been refuted in other studies. In Geller et al. (2018), for example, the authors showed that while hard-to-read cursive words engendered longer naming latencies, they did not enhance memory at test compared to an easy-to-read cursive manipulation. Similarly, Rosner et al. (2015), showed that while a low-level blurring condition produced longer naming latencies, it did not enhance memory at test . In contrast, a higher level of perceptual blur slowed naming latencies and enhanced recognition memory at test. These results suggest that perceptual degradation affects naming times in the study phase in a continuous manner, but that perceptual degradation at study must surpass some threshold to induce processing that enhances memory encoding.

Theoretical Mechanisms of the Perceptual Disfluency Effect

If perceptual disfluency is not driven by time-on-task, what then? The current findings add to our understanding of the mechanism(s) underlying the desirable effects of perceptual disfluency on memory. Eitel and Kühl (2016) postulated that if Sans Forgetica is a desirable difficulty, it fosters learning by increasing mental effort and by stimulating deeper processing. When preparing for an upcoming test (high testing expectancy), there is a high investment of effort allocated to the material, regardless of whether the to-be-learned information is fluent or disfluent—which would attenuate the effects of disfluency. Looking at both testing expectancy groups (see Figures 2b and 3b), there is some evidence for this. In both groups, recognition memory and cued recall was generally higher for Sans Forgetica, suggesting those stimuli received deeper processing, while the processing of Arial words appear to be shallower in the low test expectancy group. This suggests that with high test expectancy all words get deeper processing resulting in a smaller difference in the high test expectancy group.

Given that high testing expectancy eradicated the mnemonic benefit of Sans Forgetica, this points to a similar mechanism of action—that is, deeper, more effortful, processing at encoding. Just how this processing is carried out is still subject to debate. Geller et al. (2018) recently provided a potential answer to this question. They presented participants with varying levels of handwritten cursive stimuli (easy-to-read and hard-to-read) in order to adjudicate between current accounts of perpetual disfluency (i.e., metacognitive and compensatory processing accounts). From a metacognitive perspective, the memory benefit should be equal for easy-to-read and hard-to-read cursive words—within that account, all disfluency types are created equal (Weissgerber et al., 2017). However, the compensatory processing account suggests that the memory benefit should be greater for hard-to-read cursive stimuli. This is because during word identification hard-to-read cursive are harder and therefore should elicit more lexical/semantic processing (Perea et al., 2016). In contrast to both accounts, Geller et al. found that easy-to-read cursive were better remembered than hard-to-read cursive words, despite not being as hard. This pattern is hard for extant accounts to explain. This prompted Geller et al. to propose an alternative explanation for disfluency effects. Within their account, perceptual disfluency effects arise due to (1) increased processing difficulty during recognition (i.e., dif-

ficulty mapping letters to words) and (2) deeper processing that occurs after recognition, presumably as the result of some combination of semantic processing and metacognitive control and regulatory components. This account can explain the lack of disfluency effect in the high test expectancy group as a result of increased metacognitive monitoring and control processes eliciting attention to both types of stimuli.

A more general framework that invokes cognitive monitoring and control, such as the conflict monitoring framework (Botvinick et al., 2001), might also explain the present findings (see Geller et al., 2018; Rosner et al., 2015). Within this framework, the up- and down-regulation of monitoring and control are mediated by response ambiguity or conflict (in the current case, difficulty identifying the word). Under low test expectancy, Sans Forgetica would trigger greater control due to the difficulty associated with recognizing the stimulus—this serves to enhance memory. However, under high testing expectancy, the goal is switched to remember words for an upcoming memory test, and while Sans Forgetica is still harder, monitoring and control processes are directed to both types of stimuli, dampening/weakening the disfluency effect. The exact mechanisms underlying perceptual disfluency remain an open issue and more research is needed to better understand how perceptual disfluency enacts a desirable effect on memory.

Practical Implications

The current findings have some educational significance. While it might be tempting to conclude from these findings that Sans Forgetica should be used as study tool, the current results need to be interpreted with caution. First, and most importantly, the conclusion that Sans Forgetica is only beneficial to memory under low test expectancy makes its use in the educational domain impractical. In the classroom, students rarely encode information incidentally; learning is always purposeful and goal directed. Second, the experimental paradigms used involved simple list learning. It is not clear if Sans Forgetica would benefit learning under low test expectancy with more complex materials. Some evidence from Taylor et al. (2020, Experiments 3 and 4) suggests it might not. In those experiments, memory for factual and conceptual information was tested using more educationally realistic materials (prose passages) and displayed no mnemonic advantage for Sans Forgetica. Thus, even with low testing expectancy, Sans Forgetica did not enhance memory when the to-be-learned material was more educationally realistic. Second, the effect sizes from all three experiments were rather modest by conventional standards Funder & Ozer (2019) (Experiment 1 - d_{avg} = 0.31; Experiment 2 - d_{avg} = 0.38; Experiment 3 - Timed: d_{avg} = 0.32; Self-paced: d_{avg} = 0.15). It is unclear if the Sans Forgetica effect would replicate in educational settings where effect sizes are a known to be a smaller and more variable (Butler et al., 2014).

Finally, there is a fair amount of variability in the number of participants that benefited from perceptually disfluency. If you look at the difference scores presented in Figures 2, 3, and 4, positive effects are not seen consistently. In fact, some students are hurt by the presentation of Sans Forgetica. Before we start recommending perceptual disfluency as a potential study tool, it is critical we better understand the nature of these individual differences (i.e., why perceptual disfluency hurts some students while benefiting others).

We do acknowledge, however, that Sans Forgetica might have some practical implications. Outside the classroom, information is largely acquired incidentally, without the goal of memorization (Castel et al., 2015). If this is the case, information presented in Sans Forgetica might serve to indirectly enhance memory. For instance, one area where perceptual disfluency might be desirable is in advertising. We acquire visual information, incidentally, via billboards, online advertisements, magazines, etc. Placing this type of information in a perceptually disfluent typeface like Sans Forgetica might be helpful. However, before any definitive claims are made we need to have a better understanding of the conditions under perceptual disfluency is and is not desirable for learning.

Conclusions

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 material in Sans Forgetica) can enhance learning in a very simplified context (i.e., list learning), its efficacy as a potential learning technique is tempered by the finding that testing expectancy can nullify the effect. In educational settings, learning is explicitly goal-directed, and students accordingly encode information intentionally. Thus, Sans Forgetica (and perceptual disfluency manipulations in general) may not effectively enhance memory in ecologically valid settings. While a recent meta-analysis (H. Xie et al., 2018) claimed the disfluency effect was null and void, what is clear from the current findings is that the impact of perceptual disfluency manipulations such as Sans Forgetica, is not straightforward. Researchers should heed the call to further examine the conditions under which perceptual disfluency is and not desirable for learning.

Disclosures

Acknowledgements

This research was supported by grant number 220020429 from the James S. McDonnell Foundation awarded to the second author. We would like to Gene Brewer and two anonymous reviewers for their helpful comments on an earlier draft of the paper.

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.DJP edited the manuscript and provided feedback.

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 1.0.3; Wickham et al., 2019), *effects* (Fox, 2003; Fox & Hong, 2009; Version 4.2.0; Fox & Weisberg, 2018), *emmeans* (Version 1.5.0; Lenth, 2020), *forcats* (Version 0.5.0; Wickham, 2019a), *ggplot2* (Version 3.3.2; Wickham, 2016b), *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 1.31; Y. Xie, 2015), *lattice* (Version 0.20.41; Sarkar, 2008), *lme4* (Version 1.1.25; Bates et al., 2015), *lubridate* (Version 1.7.9; Grolemund & Wickham, 2011), *Matrix* (Version 1.2.18; Bates & Maechler, 2019), *modelbased* (Version 0.4.0; Makowski et al., 2020), *MOTE* (Version 1.0.2; Buchanan, Gillenwaters, et al., 2019), *papaja* (Version 0.1.0.9997; Aust & Barth, 2020), *patchwork* (Version 1.1.0; Pedersen, 2019), *plyr* (Wickham, 2011; Version 1.8.6; Wickham et al., 2019), *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), *report* (Version 0.2.0; Makowski et al., 2020), *Rmisc* (Version 1.5; Hope, 2013), *see* (Version 0.6.1.1; Lüdecke et al., 2020), *stringr* (Version 1.4.0; Wickham, 2019b), *tibble* (Version 3.0.6; Müller & Wickham, 2019), *tidyr* (Version 1.1.2; Wickham & Henry, 2019), *tidyverse* (Version 1.3.0; Wickham, 2017), *tinylabels* (Version 0.1.0; Barth, 2020), and *WRS2* (Version 1.1.0; Mair & Wilcox, 2020).

References

Alter, A. L. (2013). The Benefits of Cognitive Disfluency. *Current Directions in Psychological Science*, *22*(6), 437–442. https://doi.org/10.1177/0963721413498894

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming Intuition: Metacognitive Difficulty Activates Analytic Reasoning. *Journal of Experimental Psychology: General*, *136*(4), 569–576. https://doi.org/10.1037/0096-3445.136.4.569

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* (No. 3; Vol. 39, pp. 445–459). Springer New York LLC. https://doi.org/10.3758/BF03193014

Barth, M. (2020). *Tinylabels: Lightweight variable labels*. https://CRAN.R-project.org/package=tinylabels

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.

Bjork, R. A., & Yue, C. L. (2016). *Commentary: Is disfluency desirable?* (No. 1; Vol. 11, pp. 133–137). Springer New York LLC. https://doi.org/10.1007/s11409-016-9156-8

Botvinick, M. M., Carter, C. S., Braver, T. S., Barch, D. M., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. https://doi.org/10.1037/0033-295X.108.3.624

Buchanan, E. M., De Deyne, S., & Montefinese, M. (2019). A practical primer on processing semantic property norm data. *Cognitive Processing*. https://doi.org/10.1007/s10339-019-00939-6

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

Carpenter, Shana K. (2014). Spacing and interleaving of study and practice. In *Applying science of learning in education: Infusing psychological science into the curriculum.* (pp. 131–141). Society for the Teaching of Psychology.

Carpenter, Shana K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin and Review*, *13*(5), 826–830. https://doi.org/10.3758/BF03194004

Carpenter, Shana K., Wilford, M. M., Kornell, N., & Mullaney, K. M. (2013). Appearances can be deceiving: Instructor fluency increases perceptions of learning without increasing actual learning. *Psychonomic Bulletin and Review*, *20*(6), 1350–1356. https://doi.org/10.3758/s13423-013-0442-z

Castel, A. D., Nazarian, M., & Blake, A. B. (2015). Attention and incidental memory in everyday settings. In *The handbook of attention.* (pp. 463–483). Boston Review.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences, Rev. ed.* (pp. xv, 474–xv, 474). Lawrence Erlbaum Associates, Inc.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684. https://doi.org/https://doi.org/10.1016/S0022-5371(72)80001-X

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

Dunlosky, J., & Mueller, M. L. (2016). Recommendations for exploring the disfluency hypothesis for establishing whether perceptually degrading materials impacts performance. *Metacognition and Learning*, *11*(1), 123–131. https://doi.org/10.1007/s11409-016-9155-9

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

Evans, J. S. B. T. (2016). Reasoning, Biases and Dual Processes: The Lasting Impact of Wason (1960). *Quarterly Journal of Experimental Psychology*, *69*(10), 2076–2092. https://doi.org/10.1080/17470218.2014.914547

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

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

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/

Halamish, V. (2018). Can very small font size enhance memory? *Memory & Cognition*, *46*(6), 979–993. https://doi.org/10.3758/s13421-018-0816-6

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

Hirshman, E., & Mulligan, N. (1991). Perceptual interference improves explicit memory but does not enhance data-driven processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(3), 507–513. https://doi.org/10.1037//0278-7393.17.3.507

Hirshman, E., Trembath, D., & Mulligan, N. (1994). Theoretical implications of the

mnemonic benefits of perceptual interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 608–620. https://doi.org/10.1037/0278-7393.20.3.608

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

Hunter Ball, B., Klein, K. N., & Brewer, G. A. (2014). Processing fluency mediates the influence of perceptual information on monitoring learning of educationally relevant materials. *Journal of Experimental Psychology: Applied*, *20*(4), 336–348. https://doi.org/10.1037/xap0000023

Janes, J. L., Rivers, M. L., & Dunlosky, J. (2018). The influence of making judgments of learning on memory performance: Positive, negative, or both? *Psychonomic Bulletin and Review*, *25*(6), 2356–2364. https://doi.org/10.3758/s13423-018-1463-4

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting bayes factors. *Journal of Problem Solving*, *7*(1), 2–9. https://doi.org/10.7771/1932-6246.1167

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The Ease-of-Processing Heuristic and the Stability Bias: Dissociating Memory, Memory Beliefs, and Memory Judgments. *Https://Doi.org/10.1177/0956797611407929*. https://doi.org/10.1177/0956797611407929

Kühl, T., Eitel, A., Damnik, G., & Körndle, H. (2014). The impact of disfluency, pacing, and students' need for cognition on learning with multimedia. *Computers in Human Behavior*, *35*, 189–198. https://doi.org/10.1016/j.chb.2014.03.004

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

Lenth, R. (2020). *Emmeans: Estimated marginal means, aka least-squares means*. 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.

Magreehan, D. A., Serra, M. J., Schwartz, N. H., & Narciss, S. (2016). Further boundary conditions for the effects of perceptual disfluency on judgments of learning. *Metacognition and Learning*, *11*(1), 35–56. https://doi.org/10.1007/s11409-015-9147-1

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

Makowski, Dominique, Lüdecke, Daniel, Ben-Shachar, & S., M. (2020). Automated reporting as a practical tool to improve reproducibility and methodological best practices adoption. *CRAN*. https://github.com/easystats/report

McClelland, J. L., & Rumelhart, D. E. (1981). *An interactive activation model of context effects in letter perception: I. An account of basic findings.* (No. 5; Vol. 88, pp. 375–407). American Psychological Association. https://doi.org/10.1037/0033-295X.88.5.375

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

Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language*, *70*, 1–12. https://doi.org/10.1016/j.jml.2013.09.007

Mulligan, N. W. (1996). The effects of perceptual interference at encoding on implicit memory, explicit memory, and memory for source.

Journal of Experimental Psychology: Learning Memory and Cognition, *22*(5), 1067–1087. https://doi.org/10.1037/0278-7393.22.5.1067

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

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

Myers, S. J., Rhodes, M. G., & Hausman, H. E. (2020). Judgments of learning (JOLs) selectively improve memory depending on the type of test. *Memory and Cognition*, *48*(5), 745–758. https://doi.org/10.3758/s13421-020-01025-5

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* (No. 4; Vol. 8, pp. 434–447). Psychol Methods. https://doi.org/10.1037/1082-989X.8.4.434

Oppenheimer, D. M., & Alter, A. L. (2013). Disfluency sleeper effect: Disfluency today promotes fluency tomorrow. In *The experience of thinking: How the fluency of mental processes influences cognition and behavior* (pp. 85–97). https://doi.org/10.4324/9780203078938

Pan, S. C., Sana, F., Samani, J., Cooke, J., & Kim, J. A. (2020). Learning from errors: students' and instructors' practices, attitudes, and beliefs. *Memory*, *28*(9), 1105–1122. https://doi.org/10.1080/09658211.2020.1815790

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

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Perea, M., Gil-López, C., Beléndez, V., & Carreiras, M. (2016). Do handwritten words magnify lexical effects in visual word recognition? *Quarterly Journal of Experimental Psychology*, *69*(8), 1631–1647. https://doi.org/10.1080/17470218.2015.1091016

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357–383. https://doi.org/10.1080/10888430701530730

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

Rummer, R., Schweppe, J., & Schwede, A. (2016). Fortune is fickle: null-effects of disfluency on learning outcomes. *Metacognition and Learning*, *11*(1), 57–70. https://doi.org/10.1007/s11409-015-9151-5

Sagan, C. (1980). *Broca's brain: Reflections on the romance of science*. https://books.google.com/books?hl=en&lr=&id=GlXPqexwO28C&oi=fnd&pg=PR4&ots=65nePfKWk5&sig=CTTgqKJLaozsFvFqBYjBd_EOkxE

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

Seufert, T., Wagner, F., & Westphal, J. (2017). The effects of different levels of disfluency on learning outcomes and cognitive load. *Instructional Science*, *45*(2), 221–238. https://doi.org/10.1007/s11251-016-9387-8

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

Soderstrom, N. C., Clark, C. T., Halamish, V., & Bjork, E. L. (2015). Judgments of learning as memory modifiers. *Journal of Experimental Psychology: Learning Memory and Cognition*, *41*(2), 553–558. https://doi.org/10.1037/a0038388

Strukelj, A., Scheiter, K., Nyström, M., & Holmqvist, K. (2016). Exploring the lack of a disfluency effect: evidence from eye movements. *Metacognition and Learning*, *11*(1), 71–88. https://doi.org/10.1007/s11409-015-9146-2

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

Susser, J. A., Mulligan, N. W., & Besken, M. (2013). The effects of list composition and perceptual fluency on judgments of learning (JOLs). *Memory & Cognition*, *41*(7), 1000–1011. https://doi.org/10.3758/s13421-013-0323-8

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

Weissgerber, S. C., & Reinhard, M. A. (2017). Is disfluency desirable for learning? *Learning and Instruction*, *49*, 199–217. https://doi.org/10.1016/j.learninstruc.2017.02.004

Weltman, D., & Eakin, M. (2014). Incorporating Unusual Fonts and Planned Mistakes in Study Materials to Increase Business Student Focus and Retention. *INFORMS Transactions on Education*, *15*(1), 156–165. https://doi.org/10.1287/ited.2014.0130

Westerman, D. L., & Greene, R. L. (1997). The effects of visual masking on recognition: Similarities to the generation effect. *Journal of Memory and Language*, *37*(4), 584–596. https://doi.org/10.1006/jmla.1997.2531

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

Wickham, H. (2016a). *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/

Yue, C. L., Castel, A. D., & Bjork, R. A. (2013). When disfluency is-and is not-a desirable difficulty: The influence of typeface clarity on metacognitive judgments and memory. *Memory and Cognition*, *41*(2), 229–241. https://doi.org/10.3758/s13421-012-0255-8