Sans Forgetica is Not the “Font” of Knowledge: Disfluent Fonts are Not Always Desirable Difficulties

Elizabeth L. Wetzler1, Aryn A. Pyke1, and Adam Werner1

Abstract
Subsequent recall is improved if students try to recall target material during study (self-testing) versus simply re-reading it. This effect is consistent with the notion of “desirable difficulties.” If the learning experience involves difficulties that induce extra effort, then retention may be improved. Not all difficulties are desirable, however. Difficult-to-read (disfluent) typefaces yield inconsistent results. A new disfluent font, Sans Forgetica, was developed and alleged to promote deeper processing and improve learning. Although it would be invaluable if changing the font could enhance learning, the few studies on Sans Forgetica have been inconsistent, and focused on short retention intervals (0–5 minutes). We investigated a 1-week interval to increase practical relevance and because some benefits only manifest after a delay. A testing-effect manipulation was also included. Students \( (N = 120) \) learned two passages via different methods (study then re-study vs. study then self-test). Half the students saw the passages in Times New Roman and half in Sans Forgetica. Recall test scores were higher for passages learned via self-testing than restudying, but the effect of font and the interaction were nonsignificant. We suggest that disfluency increases the local (orthographic) processing effort on each word but slowed reading might impair relational processing across words. In contrast, testing and generation effect manipulations often engage relational processing (question: answer; cue: target)—yielding subsequent benefits on cued-recall tests. We elaborate this suggestion to reconcile conflicting results across studies.

Keywords
disfluency, testing effect, desirable difficulties, levels of processing, generation effect

Educational Impact and Implications Statement
This study assessed whether having students read content passages in a new difficult-to-read font, Sans Forgetica, would lead to better recall on a test 1 week later than if students read the same content in a more conventional typeface (Times New Roman). Despite the font designers’ claims that Sans Forgetica is a desirable difficulty that enhances memory retention, the new font did not improve recall test performance. Thus, students and educators should not assume Sans Forgetica will improve the learning of facts and relations.

Introduction
One objective of education is to help learners internalize and retain important factual information. Such information can be integrated meaningfully into their knowledge bases and subsequently accessed and applied to interpret new information critically and to inform decision making and problem solving. As such, pedagogical practices and factors that facilitate learning, retention, and recall are of key interest to educators and educational psychologists. It has long been known that learners’ success at retrieving content from memory tends to increase with the number of exposures to the information (e.g., Logan & Klapp, 1991). However, beyond the number of exposures, the nature of these exposures and of the learning task can also influence success at subsequent recall. There is a broad notion that “desirable difficulties” can facilitate learning (Bjork & Bjork, 2011). In brief, sometimes when the learning experience is characterized by difficulties that induce extra effort, later recall may be improved.

1United States Military Academy, West Point, NY, USA

Corresponding Author:
Elizabeth L. Wetzler, Department of Behavioral Sciences and Leadership, Official Mail & Distribution Center, United States Military Academy, 646 Swift Road, West Point, NY 10996, USA.
Email: elizabeth.wetzler@westpoint.edu
To further investigate which difficulties are desirable, and why, we simultaneously explored two difficulties: the well-established testing-effect (Roediger & Karpicke, 2006) and a relatively new difficult-to-read ("disfluent") font, Sans Forgetica, that was allegedly designed to evoke effortful reading and promote learning (Francis, 2018). Desirable difficulties involving typeface manipulations are sometimes referred to as “disfluency effects” (Kühl & Eitel, 2016). Prior attempts to obtain disfluency effects have been at best inconsistent (see reviews by Kühl & Eitel, 2016 and Weissgerber & Reinhard, 2017), and, at the time we collected our data, there were no published studies assessing the mnemonic benefits of Sans Forgetica specifically. A few studies have since been published which investigate this font, however, there are inconsistencies in those results, with some studies yielding null effects of Sans Forgetica (Geller et al., 2020; Taylor et al. 2020) and another reporting benefits (Eskenazi & Nix, 2021). Our contribution is distinctive in several respects. To endeavor to reconcile prior inconsistent findings, the current study serves to test possible hypotheses (outlined below) as to why some studies on disfluency—and Sans Forgetica in particular—failed to find a benefit, despite the possibility that a benefit may exist. Additionally, we investigated whether a disfluent font difficulty would interact with a testing effect difficulty. Thus, this study contributes to the growing body of evidence that boundary conditions may constrain the benefits of disfluent fonts.

One possible skeptical interpretation of prior null effects is that something about the materials or procedure somehow precluded the detection of desirable difficulty effects in general. To address this issue, the current study included a testing effect manipulation (study then re-study vs. study then self-test) as well as manipulating font (Sans Forgetica vs. Times New Roman), to ensure that the materials and procedure sufficed to produce a known desirable difficulty effect (i.e., the testing effect). The inclusion of a testing-effect manipulation was further motivated by a desire to enable us to compare the effect sizes of any learning benefits for each factor, and to explore the possibility of an interaction, which we further motivate in the Current Research section.

A second possible reason for prior null effects is that they may arise from the use of an insufficient retention interval since memory benefits are sometimes only detectable (or are stronger) after a delay (Toppino & Cohen, 2009; Weissgerber & Reinhard, 2017). Indeed, two very recent Sans Forgetica studies, which reported null effects, suggested that using a longer retention interval should be a key aspect for future research (Geller et al., 2020; Taylor et al., 2020). These studies, and prior studies involving font/disfluency manipulations in general, tend to use retention intervals of under 10 minutes (see meta-analysis by Xie et al., 2018). We used a 1-week retention interval to: (i) enable the detection of possible “long-term” benefits; and (ii) to increase the pedagogical relevance of the design. Longer-term benefits might arise from impacts of a desirable difficulty on consolidation and forgetting processes (vs. impacts localized to the encoding stage). Third, another possible concern about why some prior studies may have failed to detect an effect is that the effect might have been “drowned out” by another factor. Desirable difficulties work by inducing extra effort on the part of learners, but if effort might already be elevated—for example, because subjects expect to be tested on the content—it has been hypothesized that this might over-shadow the effects of the increased effort that might have been elicited by a disfluent font (Eitel & Kühl, 2016). To control for this possibility, in the current study, the delayed test on the content was unexpected. This design choice corresponds to another suggestion for future research in a recently published study (Geller et al., 2020).

Finally, the broader theoretical contribution of this paper is a proposed account in the Discussion to situate our findings in the context of desirable difficulty effects and provide a theoretical framework to reconcile our results (no memory benefits of disfluency, see also Geller et al., 2020; Taylor et al., 2020) with those of other studies which reported disfluency benefits (e.g., Eskenazi & Nix, 2021).

To provide some background, we briefly review some prior findings of studies investigating potential desirable (and undesirable) difficulties, with a focus on generation and testing effects and typeface/font manipulations.

**Established Desirable Difficulties: Generation and Testing Effects**

A long-established desirable difficulty effect is the generation effect, wherein subsequent recall is improved for material that was actively generated by the subject versus passively read (Slamecka & Graf, 1978; for a meta-analytic review, see Bertsch et al., 2007). For example, in the case of antonym word pairs, in the initial exposure phase, one can manipulate whether subjects are required to generate (some of) the content themselves (e.g., HOT: C___) versus passively reading the content (HOT: COLD). In subsequent cued and free recall tests, performance is superior for content that had been self-generated versus passively read (Bertsch et al., 2007; Slamecka & Graf, 1978). Another operationalization of the generation effect is to present words with missing letters (e.g., GUY: G_RL) versus complete words (e.g., GUY: GIRL; Bertsch et al., 2007; Geller et al., 2020).

Another established desirable difficulty is the testing effect, wherein subsequent recall is improved if students deliberately try to recall information—such as facts and relations—during their study period (i.e., self-testing) as opposed to simply re-reading the key materials (Karpicke & Roediger, 2008). The testing effect has been established for a variety of target content and types of tasks including learning: (i) the foreign counterparts of English words (Carrier & Pashler, 1992; Gaspelin et al., 2013; Karpicke & Roediger, 2008); (ii) answers to arithmetic problems (Pyke et al., 2019); and (iii)
semantic facts from prose passages (Einstein et al., 2012; Roediger & Karpicke 2006).

We suggest that for both the generation and testing effects, the desirable difficulty seemingly lies in requiring the subject to retrieve content from memory (vs. read/observe it). Note that in most generation effect studies, the task is essentially one of cued recall. For example, subjects may be provided with a cue (e.g., HOT: C_____ and also often a generation rule (e.g., that the words are antonyms). Thus, in the “learning” phase, the subject is required to recall a word in memory cued by the constraints that it is the [opposite of hot] and [starts with C]. Testing effect studies differ in that the subject’s initial exposure to the material-to-be-remembered (e.g., word pairs) must be via passive observation—that is, they have to see it at least once to be able to then practice recalling it (self-testing). This allows the content to be more arbitrary (e.g., SHOE: MELON), since it need not be generated from a pre-existing relation in memory (like HOT: COLD). However, as in the generation effect, in the self-testing part of the learning phase, subjects also typically engage in cued recall. For example, students might practice recall by testing themselves using flash cards with a cue on one side (e.g., SHOE) and the target on the other (MELON), or, for content in the form of text passages, they might try to answer questions about the content where the words in the question serve as retrieval cues.

Thus, in the above view, engaging in a process of retrieval is instrumental to these effects. As such, it may not be extra effort/difficulty per se that yields memory benefits, but effort invested specifically in retrieval of the target content. This mechanistic account is compatible with evidence that not all difficulties are desirable.

**Not All Difficulties are Desirable**

As an example of an undesirable difficulty, memory benefits tend to be elusive in cases where the added difficulty involves dividing a learner’s attention during initial exposure to the material (Fernandes & Moscovitch, 2000), or during recall practice (Gaspelin et al., 2013; see also Craik et al., 1996). Dividing the learner’s attention during recall practice would obviously limit the time/effort that could be invested specifically in recalling the target information, which seems self-defeating if the locus of the memory benefit lies in the effort invested in recalling the target content. As another example, math students can gain exposures to answers to arithmetic facts (e.g., $3 \times 4 = 4 + 4 + 4$) by computing the answer themselves or by using a calculator. Although self-computation is a more difficult and effortful way to practice, the act of self-computation itself does not provide a benefit for committing the problem-answer association to memory (Pyke & LeFevre, 2011). That said, to avoid arduous computation, students may be motivated to first attempt to recall the answer before trying to compute it; this recall attempt, rather than effort spent on computation itself, can facilitate fact learning (i.e., a testing effect, Pyke et al., 2019). The above examples of difficulties (mental computation and divided attention) may have been undesirable because they did not optimally direct effort to be invested in using or strengthening associations between elements of the target content (cue: target; question: answer).

**Font Manipulations: Desirable or Undesirable Difficulties?**

A purpose of the current study was to investigate a new potential desirable difficulty in the form of a font explicitly designed to be disfluent, Sans Forgetica (Francis, 2018). As shown in Figure 1, Sans Forgetica includes a backwards slant, as well as gaps and other irregularities in the letters. The designers, who are cognitive psychologists, claim—in absence of publishing any peer-reviewed results—that because of these properties, this font will improve retention (Sansforgetica.rmit). That said, as suggested by Geller et al. (2020), at first blush it seems possible that Sans Forgetica, which presents letters with missing parts (i.e., gaps), might induce similar generation effect benefits as presenting words with missing letters (HOT: C______). We will revisit this analogy in the Discussion to reconsider whether this apparent similarity it is likely to invoke common cognitive mechanisms.

It would obviously be of great practical value if learning could be enhanced simply by changing materials into this font. However, prior results on font manipulations have been at best inconsistent, with some studies reporting a beneficial learning effect (e.g., Diemand-Yauman et al., 2011; Halamish, 2018), but with many, arguably the “silent majority,” failing to yield one (see reviews by Kühl & Eitel, 2016 and Weissgerber & Reinhard, 2017; and a meta-analysis by Xie et al., 2018; Yue et al., 2013). For example, visually degrading (i.e., blurring) a text to induce disfluency can sometimes enhance memory (Rosner et al., 2015), but does not always show expected benefits (Yue et al., 2013). In terms of comparisons across font types, Diemand-Yauman et al. (Study 1, 2011) reported that when tested after a 15-minute distractor task, learners were better able to recall facts from passages that had been presented in “disfluent” fonts, specifically Comic Sans and Bodoni, versus in Arial. In that study, the disfluent fonts also differed in size (12 point) and ink saturation (75% greyscale) in comparison to the Arial control (16 point, black).
In all, disfluency effects are not always readily obtained nor replicated (Weissergerber & Reinhard, 2017; Xie et al., 2018). For example, Kühl and Eitel (2016) summarized the results from a special issue dedicated to research on disfluency outcomes, and in all 13 studies disfluency did not yield an overall benefit to performance. In that issue disfluency was operationalized via one or more of the following manipulations: making the text smaller, grey (vs. black), blurred, italicized, and/or in a different font than the Arial control (e.g., Times New Roman, Comic Sans, Brush Script, or Haettenschweiler).

In terms of Sans Forgetica specifically, after the data for the current study were collected, a study by Eskenazi and Nix (2021) was published suggesting that Sans Forgetica might induce desirable difficulties, relative to Courier, in a lexical acquisition task. Subjects had to learn the spelling and infer the meaning of 15 low frequency words, each presented in the context of two sentences. The efficacy of orthographic learning (spelling) was then assessed by having learners choose the correct spelling of each word from among four options in a multiple-choice recognition test. Learning of semantic meanings was assessed by presenting the word as a cue and having subjects recall the definition that they had inferred from the context sentences. These researchers reported a benefit for both orthographic and semantic learning when the original sentences were presented in Sans Forgetica (vs. Courier), but only among subjects that were high- (but not low-) skilled at spelling. Although it was not entirely clear whether these researchers checked if the skill levels among “high-skill” subjects were matched across the two font groups, individual difference may moderate disfluency effects. Nonetheless, the non-null benefits reported by Eskenazi and Nix seem to support the optimism for Sans Forgetica alluded to by the generation effect analogy (Sans Forgetica: letters missing parts; Generation Effect: words missing letters).

More recently, however, Taylor et al. (2020) found that although participants reported experiencing Sans Forgetica as disfluent (Expt. 1), there was no evidence that Sans Forgetica yielded a boost in recall relative to Arial (Expt. 2–4). In fact, they found in the second experiment that memory for word pairs was impaired when the pairs had been presented in Sans Forgetica as opposed to Arial and tested 10, 20, or 30 seconds later. The third and fourth experiments used prose passages, and they found no effect of font type on memory for either factual or conceptual information on recall tests that occurred after a 5-minute delay.

Similarly, after a 2- or 3-minute retention interval, Geller et al. (2020) also reported null effects from using Sans Forgetica (vs. Arial), despite obtaining memory benefits due to other manipulations like the generation effect (cue: target pairs where learners had to mentally fill in the vowels in the target word; Expt. 1), and pre-highlighting parts of passages (Expt. 2).

Taken together, the positive effect detected in the lexical acquisition study (Eskenazi & Nix, 2021) and the negative or null effects in other Sans Forgetica studies (Geller et al., 2020; Taylor et al., 2020) suggest that Sans Forgetica (like many other disfluencies) may be a fickle difficulty, with positive effects potentially bounded by very specific conditions.

The Current Research

Our primary objectives were to assess the learning benefits of Sans Forgetica for enhancing memory in an educational setting and to explore possible interactions with the testing effect. Using an established paradigm and materials known to elicit the testing effect (Einstein et al., 2012; Roediger & Karpicke, 2006), we added a font manipulation that allowed us to present the to-be-learned passages in their entirety in either Sans Forgetica or a more conventional font, Times New Roman. In contrast to many prior disfluency studies, we also ensured we used a delay interval (1 week) long enough that should allow desirable difficulties to emerge. Besides being delayed, the recall test was also unexpected to avoid the possibility that test expectancy might inflate effort during initial exposure to an extent that might drown out effects of interest.

Including a testing effect manipulation was partly as a sanity check to ensure our procedure and materials could produce an expected effect. However, we also wanted to determine whether there would be a potential interaction between learning method (study then re-study vs. study then self-test) and font type (Sans Forgetica vs. Times New Roman).

We expected to obtain a main effect of testing and, if the claims of the Sans Forgetica developers were justified, a main effect of font—that is, better recall for content presented in Sans Forgetica. Assuming that both factors (self-testing and disfluency) would contribute benefits, participants who used the study then self-test method combined with materials presented in Sans Forgetica were expected to score higher than all other groups on the unannounced recall test 1 week later. There was also, however, reason to believe that type of font (Sans Forgetica vs. Times New Roman) and method of study (study then re-study vs. study then self-test) might interact. Specifically, if studying by (re)reading is more difficult in the case of Sans Forgetica, and thus more effective, then (re)reading content in Sans Forgetica might be almost as difficult/effective as trying to recall the content during study (i.e., self-testing). If so, testing-effect benefits might be reduced or eliminated for material presented in the Sans Forgetica font. Consequently, the greatest gains for Sans Forgetica were expected in the study-then-(re)study condition, for two reasons. First, in the study-then-(re)study condition, learners interact for twice as long with the font as they do in the study-then-self-test condition. In the latter case, they do not see the
passage (font) during the second half of the learning interval while they are self-testing. Second, given the known benefits of testing, the study-then-self-test condition should already yield good performance, leaving less room for an additional contribution of disfluency.

Materials and Methods

Participants

Participants were eight class sections of freshman students (N=120) taking a general psychology course during the fall semester in 2019. Half the class sections (four of eight) saw the stimulus passages in Times New Roman font (N=62) and half the class sections saw the passages in Sans Forgetica (N=58). Consistent with the composition of the student body at the institution, the sample included 92 men and 28 women. Assignment of class sections to font type was random. The Institutional Review Board approved all procedures.

An a priori power analysis was conducted using G*Power 3.1.9.7 (Faul et al., 2007) to calculate the sample size needed to obtain a medium-sized effect for the between subjects font factor (Sans Forgetica vs. Times New Roman) in an ANOVA that also included a within subjects factor (method of study: study-(re)study vs. study-test). With alpha set at .05 and using a medium effect size, results indicated that a sample of 108 participants would be sufficient for achieving power of .90. To allow for possible attrition and to keep the sample size relatively balanced among groups, we oversampled and included 120 participants. A medium effect size was chosen because we were interested in testing for an effect sufficient to have pedagogical relevance.

Materials

The materials to be learned were two short passages originally designed to assess comprehension in English-as-a-Second-Language students (Rogers, 2001), but that have also since been used and validated in experimental studies on the testing effect (Einstein et al., 2012; Roediger & Karpicke 2006). One passage was about sea otters (275 words) and one was about the sun (256 words). Each passage was associated with a quiz consisting of 12 short-answer questions. For example, the sun passage contains the information that: “The sun today is a yellow dwarf star”; and its quiz contains the question: “What type of star is the sun today?”. The sea otter passage contains the information: “Sea otters dwell in the North Pacific”; and its quiz contains the question: “Where do sea otters dwell?”

Procedure

The current procedure largely replicated that in Einstein et al. (2012) and occurred in two phases: Learning (Session 1) and Recall Quiz (Session 2), which were conducted 1 week apart. These activities then served as a basis for discussion in a subsequent lesson on study habits and research methods (debriefing lesson).

Learning session. During a psychology class on learning, each student was given both passages (sun and sea otter), each printed on its own page of paper, to learn sequentially. A key difference from Einstein et al. (2012) was that half our participating class sections (4) were given these passages in Sans Forgetica font and the other sections received them in Times New Roman font. Learning method, however, was a within-subjects variable—each student learned one passage via a study then (re)study method and one via study then self-test. Each student was allocated 8 minutes of total learning time per passage. For both learning conditions, the first 4 minutes were spent studying a passage (i.e., reading it and taking notes in the margin if desired). In the study then (re) study condition, the next 4 minutes were spent doing more of the same, but in the study then self-test condition, they flipped the passage over and out of view and spent 4 minutes recalling and writing down information they could remember on a blank page. Both the association of passages with learning methods and the order that the learning methods were executed were counterbalanced across class sections. After completing the learning sessions, students indicated on a scale of 1 to 5 the extent to which they agreed that the font the passages were written in was easy to read, with 1 reflecting that they strongly disagreed and 5 indicating that they strongly agreed. Students were not told there would be a subsequent recall quiz on these materials.

Recall Test Session

One week after the learning session, in a class on memory, two unexpected recall tests were administered—each with 12 short-answer questions per passage. The students took these tests in the same order in which they had initially read the passages and were given 8 minutes to complete each test. The short-answer tests were identical to those used by Einstein et al. (2012). Our method involved an unexpected (vs. expected) test because it has been hypothesized that students may consciously exert extra effort learning in anticipation of a test, and this may overshadow the effects of extra effort induced by disfluency (Eitel & Kühl, 2016).

Debriefing Lesson

Pedagogically, the above activities served as the basis for a discussion in a subsequent lesson about: (i) which study habits students used and which they found most effective; (ii) the testing effect; (iii) research methods and the importance of replication. Learning, memory, and research methods are key topics in psychology and this hands-on demonstration served as an excellent springboard for discussion.
Table 1. Results of t-tests Comparing the Two Groups (Sans Forgetica and Times New Roman) for Equivalency of Final Course Grade and Effectiveness of the Disfluency Manipulation.

| Font type (group) | Sans Forgetica | Times New Roman | Statistics |
|-------------------|----------------|-----------------|------------|
|                   | M    | SD   | M    | SD   | t(118) | p    | Cohen's d |
| Comparing overall course grade for two groups (%) | 82.93 | 7.39 | 84.10 | 6.30 | 0.93* | .355 (ns) | 0.2* |
| Comparing ratings for font readability | 3.05 | 1.33 | 4.63 | 0.77 | 7.87 | <.001 | 1.5 |

*df=117 due to one cell missing data.

**This check to ensure groups were effectively equivalent was ns, but we included Cohen’s d for completeness (no effect yields a near—zero “effect size”).

Results

We first used final course grades in Freshman Psychology to confirm that random assignment did not yield font groups that differed significantly in academic ability. As summarized in Table 1, results from an independent groups t-test indicated that students in the Sans Forgetica (M=83%) and Times New Roman (M=84%) font groups performed similarly in the course overall. In terms of a disfluency manipulation check, Table 1 also summarizes results from an independent groups t-test, which indicated that students who read the passages in Times New Roman reported the font to be significantly easier to read than did participants who read the passages in Sans Forgetica. These results suggest that students who read the passages in Sans Forgetica did experience the font as disfluent.

A mixed model analysis of variance (ANOVA) was then used to assess the main effects of type of font (Sans Forgetica vs. Times New Roman) and learning method (study then re-study vs. study then self-test) and the interaction between them on recall test performance after a 1-week delay. In this analysis, type of font was a between groups factor and learning method was a repeated measure. The pattern of results is illustrated in Figure 2 and the statistics are summarized in Table 2. The analysis revealed a significant main effect for learning method on recall scores. Consistent with prior research, recall was significantly better after study then self-test learning than after study then (re)study learning. However, there was neither a significant effect of font type, nor a significant interaction.

Discussion

The results from this study replicated the expected testing effect (Einstein et al., 2012; Roediger & Karpicke, 2006); recall was better after study then self-test learning than study then (re)study learning. This positive finding confirmed that the procedure and materials were sufficient to allow the detection of a known desirable difficulty. The results, however, did not provide support for the hypothesis that Sans Forgetica would induce a desirable difficulty to improve the cued recall of semantic facts relative to the more conventional font, Times New Roman. Our null results for Sans Forgetica are consistent with findings from Taylor et al. (2020) and Geller et al., (2020). However, our research also extended the scope of these prior studies by providing evidence that null effects persist despite: (i) using a longer retention interval (1-week vs. a few minutes) to increase pedagogical relevance, and to allow for the fact that some desirable difficulty effects emerge or become stronger after delays (Toppino & Cohen, 2009; Weissgerber & Reinhard, 2017); and (ii) using an unexpected/unannounced test to alleviate concerns that test expectancy could generally elevate students’ effort to an extent that would drown out the effects of extra effort induced by the disfluent font (Eitel & Kühl, 2016).

A potential limitation of our study is that the sample was comprised of students who were of typical college age and who were mostly male (about 75%). That said, the institution draws students from all 50 U.S. states, is racially and ethnically diverse, and is representative of different social class backgrounds, so in these important respects, the sample is by no means homogeneous. Students in our sample can also be in any academic major at the institution. Students in our sample exhibited the testing effect, which is characteristic in studies on learning with other samples. We did not find evidence of a ceiling effect for memory scores, which also would suggest these students do not disproportionately represent high performers. Although we believe these results would generalize to college students, results from the current study may not generalize to other groups, such as students of different ages, novice readers, or students learning to read prose in a second language.

These accumulating null effects may seem surprising from the perspective of two frameworks. The first is a proposed metacognitive explanation for disfluency effects, which suggests that a learner’s awareness of increased difficulty should promote additional processing effort, and thus improved recall (Alter et al., 2007). Although our participants reported finding Sans Forgetica more difficult to read than Times New Roman, they did not exhibit any recall benefits. This result suggests that awareness of perceptual difficulty may not
always have a clear causal link to recall benefits (see also Weissgerber & Reinhard, 2017).

A second “framework” that seems challenged by the current data is the apparent analogy between Sans Forgetica and the generation effect. The former involves missing parts of letters (i.e., gaps) that the reader must mentally fill in, and the latter often involves missing letters in words that the reader must mentally fill in. Why then, in a similar context, might the generation manipulation serve as a successful desirable difficulty while Sans Forgetica did not (Geller et al., 2020)? The remainder of the discussion will be devoted to reconciling the apparent inconsistency raised by the generation effect analogy, and also to reconciling these accumulating null findings for Sans Forgetica (current study; Geller et al., 2020; Taylor et al., 2020) with the benefits reported by Eskenazi and Nix (2021) for Sans Forgetica, and with benefits sometimes reported for other disfluency manipulations (e.g., Diemand-Yauman et al., 2011).

**Semantic-Relational Retrieval Might Drive Desirable Difficulty Effects**

A unifying mechanistic account would be valuable to predict which difficulties might be desirable and which might not. A
consideration of processing at a more mechanistic level is necessary to form a clearer comparison between the processing of Sans Forgetica and classical desirable difficulties like the testing and generation effects. These classical desirable difficulties have in common that they induce retrieval processes during the “learning” of the target content. For the generation effect, one might retrieve an antonym for a cue (HOT: C____) or retrieve a concept to complete a sentence context “The water coming out of the tap was c____.” For the testing effect, the learner might practice retrieving a learned associate (MELOM) for a cue (SHOE) or might retrieve answers to questions about the content in a text passage (e.g., Who turned on the tap?). In our context of interest, this recovery effort in the learning phase leverages associative relations (e.g., words co-occurring as a pair) or semantic relations (which may take the form of declarative facts). Thus, if the goal is for a student to learn such relational information, our relational-retrieval hypothesis is that some desirable difficulties arise when, during learning, students expend effort engaging in retrieval involving associative or semantic relations. This view is compatible with claims that a “deeper” (semantic) level of processing is associated with better subsequent memory (Craik & Tulving, 1975). Here, the specific type of processing we are implicating is retrieval of relational information. Our hypothesis is also compatible with the theory of transfer appropriate processing (Morris et al., 1977), which suggests that the ability to apply or recall knowledge is facilitated if the learning context and context of application are similar (i.e., require/involve similar processes, here: retrieval involving relations between concepts).

Revisiting a Generation-Effect Analogy for Sans Forgetica: Semantic Versus Perceptual Processing

If inducing semantic/relational retrieval is necessary or at least sufficient to be a desirable difficulty, is there reason to suspect that Sans Forgetica might not induce relational retrieval, but a generation manipulation would? In contrast to our generation effect examples, which involved the use of semantic relations to retrieve a word (HOT:C____; “The water coming out of the tap was c____”), we suggest that the type of information necessary to mentally fill in the gaps in Sans Forgetica letters could be very local in scope—that is, the rest of the letter in question is likely sufficient to recognize the letter as a whole (see Figure 1). We acknowledge that some readers might engage in holistic processing to recognize each word as a whole (Allen et al., 1995)—but even so, the information required is local to the word. Readers can readily interpret single words in isolation written in Sans Forgetica and are unlikely to require information from other concepts in the sentence to enable the recognition of a current word. Thus, in keeping with being local in scope, we suggest that the necessary processing to interpret a word in a disfluent font like Sans Forgetica is orthographic (vs. semantic) in nature. In contrast, in many generation effect examples, information from other words in the context is often important to guide and constrain the completion of the word in question (e.g., “The water from the tap was c____” vs. just “c____”). Thus, we argue that this kind of generation task typically evokes the retrieval of semantic relations, whereas a disfluent font need not.

We do acknowledge that a memory process is typically involved to read a single word. Assuming that the word is known to the reader, the orthographic cue supports the retrieval—or more likely recognition—of a lexical representation in memory. The association being leveraged or strengthened by this process would be between the visible (perceptual) representation of the word and its mental orthographic representation. However, the task demands or “difficulties” of interpreting known words written in this disfluent font would likely neither specifically require nor motivate the use of semantic-relational memory.

Not only do we suggest that the semantic relations within a sentence are likely unnecessary to the task demands of reading words in Sans Forgetica, we further suggest that such semantic relations might be less well apprehended and less accessible to bring to bear when read in a disfluent versus a fluent font. Specifically, disfluent fonts, including Sans Forgetica, tend to slow the pace of reading (Eskanazi & Nix, 2021; Xie et al., 2018), and slower reading can interfere with comprehension, likely because it is harder to maintain and integrate information over longer durations in working memory (Chang, 2010). As such, in contrast to the expectations of the Sans Forgetica designers, slower processing by no means guarantees “deeper” (i.e., semantic) processing.

Thus, despite apparent similarities to a generation manipulation involving missing letters, we suggest that the extra processing time and effort induced by the Sans Forgetica is not invested in relational retrieval (in contrast to testing and generation effects), which may explain why it does not yield benefits on a subsequent cued recall. We further suggest that this account might more broadly explain why disfluency effects in general are often very hard to obtain (Kühl & Eitel, 2016; Xie et al., 2018). Disfluent fonts might make heavy demands on local (e.g., word-level) orthographic processing, but may not automatically induce—and may even impair—bringing semantic-relational memory to bear to read the words.

Boundary Conditions: When Might Disfluent Fonts Yield Benefits?

Having suggested that disfluent fonts may fail to yield benefits because they might not necessarily induce extra semantic-relational processing, how then can we account for findings that Sans Forgetica and other disfluent fonts occasionally yield memory benefits? We suggest three main possibilities. The first and most mundane possibility is that some such findings may reflect Type I errors.
Are some disfluency effects due to distinctiveness versus difficulty? A second possibility is that, rather than functioning as a desirable difficulty per se, atypical fonts and sizes may sometimes signal novelty—making material in those typefaces more perceptually distinctive, salient, and attention-grabbing. On this view, the atypical font would function in a similar fashion as emphasizing phrases using bold, italics, and/or pre-highlighting (Diemand-Yauman et al., 2011). Note that the locus of this possible distinctiveness effect would be based on perceptual versus semantic mechanisms.

If disfluencies sometimes achieve effects via perceptual distinctiveness, a possible issue arises in contexts where all of the stimulus material is presented in the distinctive/disfluent font as opposed to presenting only key facts and phrases in the distinctive font. Distinctiveness is presumably useful when it serves to direct attention to key elements of content—likely at a cost of allocating relatively less attention to other content elements. It is unclear how useful such “distinctiveness” would be if applied to every element of content. Imagine if every sentence in a textbook were pre-highlighted. Thus, it may not be surprising that null effects are often found in studies, like the current one, where learners in the disfluent group saw the whole passage in the disfluent font. As such, the font would neither serve to make key elements distinctive (distinctiveness effect) nor, as we suggest above, serve as a desirable difficulty that promotes semantic-relational processing.

That said, there is evidence to suggest that distinctiveness benefits might occur in contexts where only some of the content appears in a disfluent font. For example, Diemand-Yauman et al. (2011, Expt. 2) reported a disfluency benefit when distinctive fonts were used in supplementary course materials like worksheets and slides. In another between-groups study, Geller et al. (2020, Expt. 2) had subjects read a passage that was mostly in Arial but contained 11 critical phrases that were either in Arial (control group), in Sans Forgetica (disfluent group) or pre-highlighted (highlighting group). After a 3-minute distractor task, subjects had to recall target key words from these critical phrases on a fill-in-the-blank recall test. Numerically the highlight group outperformed the disfluent group which outperformed control (118) = −.128, p = .898. Thus, in that study, however, target key words may also have been bolded in all three conditions which may have somewhat diminished the distinctiveness effects of the font/highlighting manipulation.

Do novel fonts facilitate learning novel words? Another context in which disfluent fonts sometimes yield a benefit is when the task involves learning unfamiliar words (Diemand-Yauman et al., 2011; Eskenazi & Nix, 2021). For example, Eskenazi and Nix reported that high-skilled spellers were better able to learn the spellings and meanings of new-to-them words (e.g., otiose) based on their use in two sentences if the sentences were presented in the disfluent Sans Forgetica font versus in Courier. Another study by Diemand-Yauman and colleagues reported that subjects were more adept at learning the properties of hypothetical types of aliens with novel names (e.g., The pangerish has blue eyes) when the information was presented in a disfluent font (Comic Sans or Bodoni) as opposed to in Arial. The tasks in these studies seem to involve two main demands: (i) perceptually processing the novel word and creating a new mental orthographic representation for it (and perhaps also phonological); and (ii) processing the semantic information in the rest of the sentence and mentally associating it with the newly created mental representation of the word.

We suggest that disfluency could facilitate the first demand—the perceptual processing and mental orthographic representation of a new word. When in a disfluent font, the novel word is “doubly distinctive”: Not only is the specific sequence of letters unfamiliar to the reader (precluding automatic processing), but it is also in a distinctive font. This distinctiveness and the slowed perceptual processing pace afforded by a disfluent font could lead to the allocation of extra attention and time to encode the orthographic properties of the new word. A disfluent font could thus potentially promote a robust orthographic mental representation, which could facilitate a learner’s ability to later spell the word, which could explain the benefits for learning the spelling of novel words reported by Eskenazi and Nix. However, note that their spelling benefits had noteworthy boundary conditions—that is, they occurred on a recognition (four-alternative multiple choice) rather than a recall test, and only occurred for the subset of subjects who were independently assessed as high-skilled at spelling. Diemand-Yauman et al. did not assess subjects’ ability to spell their novel words (alien species) since those words served as part of the cue in their cued-recall task (e.g., “what color eyes does the pangerish have?”). In terms of the present study, our materials conveyed new facts but possibly included only one novel word (mustelids, which refers a family of carnivorous mammals including sea otters). Cued-recall of this concept—which included reproducing its spelling—was very poor (6% correct) and was not better in our Sans Forgetica versus Times New Roman condition, t(118) = −.477, p = .638. Even when we compared the percent of subjects who seemed at least on the right spelling track (e.g., knew it started with “m”), performance did not differ across font (Times New Roman: 42%; Sans Forgetica: 43%), t(118) = −.128, p = .898. Thus, in all, it is advisable to exercise caution before assuming that a disfluent font will facilitate learning the spelling of a novel word for students in general—not just high-skilled spellers—and in cases where the spelling must be recalled versus just recognized.

In terms of learning semantic meaning versus spelling, how might we explain why the recall of the meaning (or semantic properties) of the novel words may have been improved for content presented in a disfluent font (Diemand-Yauman et al.,
in the file drawer (Kühl & Eitel, 2016). In all, it may be unpublished studies that obtained null-effects and ended up significant disfluency effects may belie a silent majority of not reliably afford the same benefits as difficulties like self-automatically induce semantic-relational processing, they do even interfere with) “deeper” semantic and relational processing of sentences relative to fluent fonts. In the context of Eskenazi and Nix (2021) and Diemand-Yauman et al. (2011, Expt. 1), the task itself—learning a new word and its meaning—is inherently relational. So, we suggest that the inherent task demands motivated relational processing, regardless of the fact that parsing a disfluent font itself might not automatically evoke deeper semantic-relational processing. That said, as we have suggested previously, the slowed reading pace for disfluent fonts may not be ideal for comprehending and integrating content across the sentence in working memory (Chang, 2010). This may explain why, for Eskenazi and Nix, the recall benefit for new word meanings was restricted to high-skilled subjects. Further research is necessary to determine whether disfluency benefits are reliable for learning new words.

Conclusions

The results of the current study, which support and extend prior research (Geller et al., 2020; Taylor et al., 2020), suggest caution in assuming that Sans Forgetica will serve as a desirable difficulty in an educational setting. Even though we avoided test expectancy and used a 1-week delay interval that should allow desirable difficulties to emerge, Sans Forgetica did not promote the recall of semantic knowledge above and beyond that afforded by the Times New Roman font. We did obtain the expected testing effect, which we attribute to the fact that the “difficulty” in the self-testing phase requires the application of retrieval processes operating on semantic/relational knowledge (our relational-retrieval hypothesis). In contrast, we suggest that disfluencies tend to evoke extra perceptual-orthographic (vs. semantic-relational) processing. Note that words in a disfluent font can typically be read in insolation, so disfluency need neither necessitate nor motivate readers to bring semantic information from the prior sentence context to bear to inform the recognition of the current word. The slower processing induced by typical disfluencies need not guarantee (and may even interfere with) “deeper” processing, especially processing that integrates content across multiple words or sentences. Thus, we suggest that since disfluencies need not automatically induce semantic-relational processing, they do not reliably afford the same benefits as difficulties like self-testing. As others have argued, published studies reporting significant disfluency effects may belie a silent majority of unpublished studies that obtained null-effects and ended up in the file drawer (Kühl & Eitel, 2016). In all, it may be prudent to exercise caution before rushing to put any disfluency manipulations into pedagogical practice.

Authors’ Note

Results from this study were presented at the 2020 Annual Conference of the Eastern Psychological Association. The Institutional Review Board approved all procedures: project control number BSL-20-003. Wetzler, Werner, Pyke, Feltner, Thomson. The views expressed herein are those of the authors and do not reflect the position of the United States Military Academy, the Department of the Army, or the Department of Defense.

Acknowledgments

The authors are grateful for the assistance of MAJ David Feltner in data collection.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Elizabeth L. Wetzler https://orcid.org/0000-0001-5266-4747

References

Allen, P. A., Wallace, B., & Weber, T. A. (1995). Influence of case type, word frequency, and exposure duration on visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 21*(4), 914–934.

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

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

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*(6), 633–642. https://doi.org/10.3758/BF03202713

Chang, A. (2010). The effect of a timed reading activity on EFL learners: Speed, comprehension, and perceptions. *Reading in a Foreign Language, 22*(2), 284–303.

Craik, F. I. M.,Govoni, R., Naveh-Benjamin, M., & Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General, 125*(2), 159–180. https://doi.org/10.1037/0096-3445.125.2.159

Elizabeth L. Wetzler https://orcid.org/0000-0001-5266-4747
