The Effects of Font Disfluency on Reading Retention

Marissa Bailey, Lauren Duty, Austin Greenway

Abstract

The font disfluency effect is the theory that low legibility fonts with higher reading difficulty promote higher cognitive engagement and therefore increase content retention (Bjork, 1994). Some researchers believe that this theory can be applied when choosing fonts to increase a reader's ability to retain information. If this is the case, the application of this theory could produce a widespread impact on several fields, including education, marketing, and design. Several studies have shown font disfluency to be effective (Bjork, 1994; Oppenheimer et al., 2010; Sungkhasettee et al., 2011), but several studies have also shown it to be ineffective (Eitel & Kühl, 2016; Rummer et al., 2016; Taylor et al., 2020). In an attempt to learn more about the effects of font disfluency on reading retention, we conducted a study involving 64 participants in which we administered a timed reading test in four different fonts styles to evaluate font disfluency and rank reading difficulty, differing versions of a multiple-choice reading retention test to compare participant scores to font styles and difficulty rankings, and a post-test interview to assess participant perceptions of font and performance. Our results may indicate that there is a correlation between the legibility of a font style and how much content the reader retains.

Introduction

In today's society, textual communication is everywhere, and it is typically designed for a distinct purpose: reaching, persuading, and convincing a target audience. In addition to these goals, text is often meant to be remembered. For a body of text to fulfill this purpose, selecting the correct typeface is vital. A study entitled, "The Taste of Typeface" has explored the ways in which people associate tastes with different shapes and fonts (Velasco et al., 2015). When choosing font styles, one must consider how particular fonts may be associated with other ideas, emotions, and experiences. Aside from how meaning is derived from fonts, audience engagement can be largely impacted by the choice of typeface. One study testing the difference between handwritten text and typed text concluded the following: fonts that mimic handwriting elicit the action of approach and therefore haptic engagement (Izadi & Patrick, 2020). There are numerous examples of why font choice is important, but in this study, we are focusing on how fonts affect content retention. Specifically, we are seeking to understand whether font dysfluency has any effect on how much content readers can remember.

What is font disfluency, and how might it affect the amount of information people retain? The font disfluency effect is the theory that low legibility fonts with higher reading difficulty promote higher cognitive engagement and therefore increase content retention (Bjork, 1994). Some researchers believe that this theory can be applied when choosing fonts to increase a reader's ability to retain information. If this is the case, the application of this theory could produce a widespread impact on several fields, including education, marketing, and design. In two related studies on disfluency, researchers found that harder-to-read fonts increased retention rates and that perceptual disfluency can successfully function in education as a desired difficulty (Oppenheimer et al., 2010). Further research on the concept of desirable difficulty has shown the potential benefits of applying font disfluency. A study in 2011 tested twenty undergraduate students from the university of California (Sungkhasettee et al., 2011). The methodology required each participant to study lists of words. These lists were presented in two different formats: upright and inverted. The results of the study found that recall performance was better for inverted words across all lists (Sungkhasettee et al., 2011).

Although recent studies have provided promising results on the use of font disfluency, there is still doubt surrounding the validity of this theory. Some researchers argue that there is a difference between disfluent difficulty and desirable difficulty. A study on fonts and memory notes that, “Of course, not all difficulties are desirable, and desirable difficulties are notoriously fickle” (Taylor et al., 2020). Upon further investigation, it seems as though several researchers would agree in stating that applying desirable difficulties is not generally effective. One study hypothesized that disfluent text paired with high test expectancy would elicit more mental effort, increased retention, and better test scores (Eitel & Kühl, 2016). However, the researchers found that disfluency was not effective and could even be a drawback under those experimental conditions (Eitel & Kühl, 2016).

In fact, many researchers have found flaws in studies that support disfluency. Several disfluency-supporting studies have tested their participants using word lists rather than paragraphs, which does not mimic real-world contexts. Additionally, it has been noted that the test content in certain studies was not only disfluent but also unusual. In 2016, a study was conducted in response to this flaw and the difference in methods, produced countering results (Rummer et al., 2016). In contrast to studies mentioned earlier, this study on disfluency and learning outcomes found that the use of disfluent text in educational settings does not produce learning advantages (Rummer et al., 2016).

The conflicting results of many of the previously mentioned studies makes the effectiveness of font disfluency unclear. While some researchers advocate for the use of harder-to-read materials, others discredit the idea based on their own results. In the following study, we conducted tests to further explore the theory of font disfluency. Our research aims to determine if there is a relationship between font legibility and how much information readers remember. To produce reliable results, we have drawn methods and best practices from past studies to design valid experiments.

Most of the fonts we used were chosen from a population of fonts on Google Fonts based on possessing the highest frequency of the characteristics of their font style. Old Standard TT had the highest number of serifs per letter, Zen Maru Gothic had the fewest number of letter extensions, Cherish had the highest frequency and longest length of flourishes per letter. Sans Foregtica was chosen as a display font because it was used in a previous unpublished study by researchers at the Royal Melbourne Institute of Technology (RMIT) for its properties that supposedly create desirable reading difficulty (RMIT, 2018). We decided to include Sans Forgetica in our study to test the validity of RMIT’s claim.

In previous research studies, researchers used word pairs or highly unusual words when designing their reading retention tests. Conversely, we chose to use paragraphs to mimic the style of reading that participants would normally engage in. We created a paragraph geared towards a low reading level (it rates between third and fifth grade depending on the readability scale being used) to control for the difficulty of the content. We chose to base our reading comprehension questions on the adjectives and adverbs in the sentences to test for content retention as opposed to using nouns and verbs that may have only been effective for testing concept retention.

Methods

To answer our research question most effectively, our team employed quantitative experimental research with test participants and then conducted post-test interviews. Our research team utilized convenience sampling to recruit testing subjects for our study. Due to the time and locational constraints of our study, this type of sampling was used to recruit as many participants as possible and increase the reliability of the trends found from our data points. In the following paragraphs, we will disclose the nature of our quantitative experimental research method:

We recruited participants in the Atrium building on campus. We tested during four different sessions which took place across varying times but were mostly conducted during midday. Once recruited, our participants were seated in a controlled environment (a quiet, well-lit room) and given a consent cover letter that discussed the following sections:

- Title of Research Study
- Researcher’s Contact Information
- Description of Project
- Explanation of Procedures
- Risks or Discomforts
- Benefits
- Compensation
- Confidentiality.

Upon agreeing to the terms of the cover letter, participants began the timed reading section of our test. Participants were informed that their reading would be timed and then were sequentially given four different printed paragraphs, each containing sixty words, and our researchers recorded their reading speeds. Each paragraph was printed in a different style from one of the following four font styles: Old Standard TT (serif font), Zen Maru Gothic (sans serif font), Sans Forgetica (display font), and Cherish (script font). The order of the font styles received by the participants was varied in aggregates of eight to ensure that the specified order of the fonts read did not affect the participants’ performance, thus bolstering the internal validity of our research. Upon completion of the first reading, the participants continued the process until each of the four paragraphs in the varying font styles were read and their times were documented.

After the completion of the timed reading section, participants immediately began the reading retention section of the test. The participants were informed that they would receive a sheet of paper containing a paragraph printed in one of the four font styles that they had seen in the previous section. The order of the font styles received by the participants in the second section was operated by a schedule to ensure that our team gathered equivalent data points on each of the four font styles. This section, as it was explained, would require the participant to read the paragraph at their own untimed pace. After completing the reading, the paragraph was collected, and participants were given a multiple-choice questionnaire containing five questions that tested their ability to recall certain adjectives and adverbs from the paragraph that they had read. At the end of the multiple-choice section, the participants were given a demographic questionnaire so that the research team might recognize existing patterns in the data found based on their personal information (age, self-identification, ethnicity, education level).

Upon completing the reading test, we conducted our second research method: a post-test interview. Our researchers followed an interview schedule containing five questions regarding the test that the participants had just completed and their perceptions of the fonts they had read. The interview questions were asked in a funnel sequence, beginning with broad questions first, followed by specific, closed-ended questions regarding their experience during the first research method employed in our study. Our researchers also noted any additional comments participants made about their feelings towards specific fonts. The nature of the interview schedule allowed our researchers to ask follow-up questions, thus gaining greater insights into our findings.

After completing our quantitative experimental research and post-test interviews for our sixty-four participants, our team gathered and organized the data. Our research team was then able to recognize trends in the data, allowing us to answer our research question most effectively.

Results and Discussion

The results of our findings do seem to indicate that there may be a link between the reading difficulty of a font style and the reader's retention of content, as shown below in *Table 1*.

Table 1: Participant Score Distribution by Test Version

Test# SANS	Score	Test # SERF	Score	Test # SCRP	Score	Test # FORG	Score
SANS-03	1	SERF-05	0	SCRP-04	1	FORG-02	2
SANS-05	1	SERF-11	0	SCRP-06	1	FORG-03	2
SANS-16	1	SERF-07	1	SCRP-05	2	FORG-10	2
SANS-07	2	SERF-04	2	SCRP-07	2	FORG-12	2
SANS-10	2	SERF-06	2	SCRP-11	2	FORG-15	2
SANS-14	2	SERF-08	2	SCRP-14	2	FORG-01	3
SANS-15	2	SERF-14	2	SCRP-16	2	FORG-06	3
SANS-01	3	SERF-15	2	SCRP-01	3	FORG-09	3
SANS-04	3	SERF-01	3	SCRP-02	3	FORG-11	3
SANS-06	3	SERF-02	3	SCRP-10	3	FORG-13	3
SANS-08	3	SERF-09	3	SCRP-12	3	FORG-04	4
SANS-09	3	SERF-10	3	SCRP-03	4	FORG-05	4
SANS-11	3	SERF-03	4	SCRP-08	4	FORG-08	4
SANS-12	3	SERF-12	4	SCRP-09	4	FORG-14	4
SANS-02	4	SERF-16	4	SCRP-13	4	FORG-16	4
SANS-13	4	SERF-13	5	SCRP-15	5	FORG-07	5
Average Score	2.50	Average Score	2.50	Average Score	2.81	Average Score	3.13

Table 1: SANS tests were administered in the sans serif font (Zen Maru Gothic), SERF tests were administered in the serif font (Old Standard TT), SCRP tests were administered in the script font (Cherish), and FORG tests were administered in the display font (Sans Forgetica). Columns are ordered by average score from lowest to highest.

The results in the table above also seem to indicate that Sans Forgetica did have the highest retention rate among the chosen fonts, which is consistent with the study conducted at the Royal Melbourne Institute of Technology that led to the creation of Sans Forgetica. These findings are in opposition with other studies conducted that specifically involved Sans Forgetica and font disfluency (Geller & Peterson, 2021; Taylor et al., 2020), but this may be due to differences between those studies and our study. Those studies use of word pairs (Taylor et al., 2020) and word lists (Geller & Peterson, 2021) while our study used full paragraphs. Both other studies were conducted online and only compared Sans Forgetica with one other font style while our study was conducted in a physical testing space where participant environment was controlled, and we compared four font styles of varying disfluency levels.

During our timed reading test, the reading order was varied and seemed to have no impact on participant reading times. Our findings do indicate that font difficulty levels varied based on the participant, therefore the fonts were shown to be more or less difficult for different individual participants, as shown below in *Table 2*.

Table 2: Font Difficulty Rank by Individual Participant Reading Times

	Zen Maru Gothic (Sans Serif)	**Old Standard TT (Serif)**	**Cherish (Script)**	**Sans Forgetica (Display)**
Ranked Highest Difficulty	0	1	55	8
Ranked Lowest Difficulty	25	37	0	2

Table 2: Numbers in each section represent the number of participants whose scored reflected the difficulty rating in column one for the font in row one

For most participants, the highest difficulty font was Cherish (script) and the lowest difficulty font was Old Standard TT (serif). Sans Forgetica (display) was the second highest difficulty font and Zen Maru Gothic (sans serif) was the second-lowest difficulty font. These ratings were not consistent for all participants, and due to these inconsistencies, the font difficulty ratings varied for each participant. Previous studies conducted on font disfluency have not accounted for individual differences in font difficulty rankings, which may possibly have confounded results.

To account for varying difficulty rankings among participants, we also analyzed our results by correlating the individual participant's font difficulty ranking with the version of the test they received so that we could review the scores based on how the test version matched up with each participant's personal difficulty ratings. The results are shown below in *Table 3*. When disregarding the font style itself and considering font difficulty among individual participants, the results show that the users typically scored higher when given the version of the test in the font that was rated as the second most difficult for them. Likewise, participants typically scored lower when given the version of the test in the font that was ranked lowest difficulty for them.

Table 3: Participant Score Distribution by Ranked Difficulty

Difficulty: Highest	Score	Difficulty: High	Score	Difficulty: Easy	Score	Difficulty: Easiest	Score
FORG-07	5	FORG-04	4	SANS-02	4	SERF-13	5
SCRP-15	5	FORG-14	4	SANS-13	4	SERF-12	4
FORG-05	4	FORG-16	4	SERF-03	4	SERF-16	4
FORG-08	4	SCRP-03	4	SANS-01	3	SANS-04	3
SCRP-09	4	SCRP-08	4	SERF-02	3	SANS-06	3
SCRP-13	4	FORG-01	3	SERF-10	3	SANS-08	3
SCRP-01	3	FORG-06	3	SANS-07	2	SANS-09	3
SCRP-02	3	FORG-09	3	SANS-14	2	SANS-11	3
SCRP-10	3	FORG-11	3	SERF-04	2	SANS-12	3
SCRP-12	3	FORG-13	3	SERF-14	2	SERF-01	3
FORG-15	2	FORG-02	2	SERF-15	2	SERF-09	3
SCRP-05	2	FORG-03	2	SANS-16	1	FORG-12	2
SCRP-07	2	FORG-10	2			SANS-10	2
SCRP-11	2	SANS-03	1			SANS-15	2
SCRP-14	2					SERF-06	2
SCRP-16	2					SERF-08	2
SCRP-04	1					SANS-05	1
SCRP-06	1					SERF-07	1
						SERF-05	0
						SERF-11	0
Average Score	2.89	Average Score	3.00	Average Score	2.67	Average Score	2.45

Table 3: There are uneven numbers of participants in each category because tests were distributed based on font, not difficulty.

Several other data points were also collected to either validate or invalidate the findings of our timed reading and reading comprehension tests. During the interview portion of our test, we measured the participant's perception of which font was the most difficult, the participant's perception of which font was most enjoyable to read, and the participant's perception of how much information they retained based on the font style. These factors correlated in any way to our other results, which may imply that the participant's perception of the font did not impact their retention. Many participants who perceived Cherish as having the highest difficulty rating had higher reading times on Sans Forgetica and vice versa. Participants who received the version of the test in the font they perceived as most enjoyable to read did not do any better on average than other participants who received the same test version.

We also collected data on age, gender, and education. Education and age both showed a correlation with test scores, which indicates that older participants and those with higher education levels tended to get higher test scores; however, these results should be interpreted with caution. One reason to interpret our findings with caution is that higher education is inherently linked to higher age due to the time it takes participants to obtain higher education. On average, participants who had the highest levels of education fell into higher age ranges. Furthermore, many participants self-identified that their highest level of education obtained was "high school graduate" despite the fact that they were college students and "some college" was the most accurate response. Some participants who were graduating seniors struggled to choose between "some college" and "4-year degree" because, although they had intellectually obtained a "4-year degree" education level, they had not received their diploma. Therefore, it could be that students who had obtained "some college" self-identified at a lower education level because they were not academically confident. Our findings for education and age should also be considered in the context of our study's skewed sampling bias, which is discussed in further detail later in this section.

While our results would seem to indicate a correlation between the reading difficulty of a font style and the reader's retention of content, there were several issues with our study. The largest potential issue with our study is that participants knew what they were being tested for. The information given about the study during the consent process may have impacted their expectations for the subsequent reading tests and therefore primed them to respond to the tests in certain ways. Many of the participants made comments during their interviews to explain why they thought they did or did not remember more, and some of those comments were oddly similar to the phrasing in the consent letter. This could be coincidence, but the possibility exists that this information may have influenced how participants responded to the test. Due to the limitations of our research context, we were not able to use deception while conducting our research, but future researchers may consider using some deception during the testing process to avoid priming participants.

Another potential issue with our study is that majority of our participants were white male college students, and because most of our participants were chosen from a similar geographic location, many of them shared the same areas of study: computer science or engineering. Both of these sample characteristics may have skewed our results. For example, we noticed that three themes continued to appear during our interviews, especially among the majority demographic of our sample:

- Participants self-identified as poor readers or did not do much reading outside of school-related settings.
- Participants thought Sans Forgetica was "cool" or "interesting" because it reminded them of certain font styles from video games.
- Participants did not like the script font Cherish because they did not know how to read cursive.

These interview responses indicate that our study may have achieved different results if our sample group had contained equal representations for age, gender, and education as opposed to our heavily skewed white male STEM-major college student sample. Sample groups who identify as avid readers, spend more time reading, don't play video games, or know how to read cursive may have responded to the font styles differently. While some of our interview data seems to indicate that the participant's perception of the font did not influence their performance on the test, further research with other sample groups is recommended to confirm if our results are valid.

Regardless of whether our results are shown to have external validation across other populations, or at least external validation for other populations of college students, it is still questionable as to whether or not it would benefit readers to engage with texts that intentionally promote font disfluency. During our interviews, many participants discussed the two highest difficulty fonts, Cherish and Sans Forgetica, with mostly very negative descriptions.

Examples of these descriptions include:

- “Felt like every hole in a letter was a hole in my brain” (Sans Forgetica)
- “Looked like a bunch of shapes” (Sans Forgetica)
- “Weird. Made my brain want to fill in the letter so I got stuck on every sentence.” (Sans Forgetica)
- “Too fancy. My brain kind of filled in as I was reading but it was still too difficult” (Sans Forgetica)
- “Almost illegible” (Cherish)
- “Miserable” (Cherish)
- “Didn’t like it” (Sans Forgetica)
- “I don’t really do cursive” (Cherish)
- “[Looked like] fake cursive” (Cherish)
- “No word space” (Cherish)
- “My brain rebelled against me” (Sans Forgetica & Cherish)

Our participants were tasked with reading very short samples of these fonts, just one 60-word paragraph in each style and an additional 72-word paragraph in one of the font styles. The paragraphs were very short and had an average reading level between third and fifth grade, so participants did not read for long and did not struggle to read or understand the material itself, only the font style. In a real-world setting, such as a textbook in which readers may read for hours and may struggle to memorize new vocabulary terms or comprehend new concepts, would it be beneficial to their learning to increase the difficulty of the text? Would increasing the difficulty level make the material so difficult for readers to understand that their “brain would rebel” as one participant put it? Further research with longer and more difficult material may give better insight into the potential applications of our findings and confirm whether readers would actually benefit from reading materials in disfluent fonts or if comprehension would decline due to frustration.

References

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Eitel, A., & Kühl, T. (2016). Effects of disfluency and test expectancy on learning with text. *Metacognition & Learning*, *11*(1), 107–121. https://doi.org/10.1007/s11409-015-9145-3

Geller, J., & Peterson, D. (2021). Is this going to be on the test? Test expectancy moderates the disfluency effect with sans forgetica. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001042.supp (Supplemental)

Izadi, A., & Patrick, V. M. (2020). The power of the pen: Handwritten fonts promote haptic engagement. *Psychology & Marketing*, *37*(8), 1082–1100. https://doi.org/10.1002/mar.21318

Oppenheimer, D., Diemand-Yauman, C., & Vaughan, E. (2010). Fortune Favors the Bold (and the Italicized): Effects of Disfluency on Educational Outcomes. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *32*. https://escholarship.org/uc/item/4wd1s7hj

Royal Melbourne Institute of Technology. (2018). *Sans Forgetica*. Sans Forgetica-RMIT. https://sansforgetica.rmit.edu.au/

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

Sungkhasettee, V. W., Friedman, M. C., & Castel, A. D. (2011). Memory and metamemory for inverted words: illusions of competency and desirable difficulties. *Psychonomic Bulletin & Review*, *18*(5), 973–978. https://doi.org/10.3758/s13423-011-0114-9

Taylor, A., Sanson, M., Burnell, R., Wade, K. A., & Garry, M. (2020). Disfluent difficulties are not desirable difficulties: the (lack of) effect of Sans Forgetica on memory. *Memory (Hove, England)*, *28*(7), 850–857. https://doi.org/10.1080/09658211.2020.1758726

Velasco, C., Woods, A. T., Hyndman, S., & Spence, C. (2015). The Taste of Typeface. *I-Perception*, *6*(4), 1–10. https://doi.org/10.1177/2041669515593040