The Journal of Experimental Education

Why dysfluent Font does not aid Second Language Learning

Submission ID	244575092
Article Type	Learning, Instruction, and Cognition
Keywords	Desirable difficulty, Dysfluency effect, Fonts, Se cond language learning, Judgments of Learning
Authors	Elina Tsigeman, Maksim Likhanov, Lidia Kalinni kova, Anastasiia Nikulenko, Elena Filimonova, Y ulia Kovas

Why dysfluent Font does not aid Second Language Learning

Elina Tsigeman[1], Maxim Likhanov[2], Lidia Kalinnikova[3], Anastasiia Nikulenko[3], Elena Filimonova[3], Yulia Kovas[4*]

[1] Laboratory for Social and Cognitive Informatics, HSE University, Saint Petersburg, Russia;

[2] State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University;

[3] Foreign Language Training Center, ITMO University, Saint Petersburg, Russia;

[4] HSE University, Moscow, Russia

[*]Correspondence: Yulia Kovas, e-mail: yuliakovas@icloud.com

Abstract

Multiple studies show that reading in difficult-to-read (dysfluent) fonts can enhance memory and comprehension of learned material, but it is unclear whether this effect extends to second language (L2) learning. This study investigated the effect of dysfluent fonts on L2 text memorization and comprehension, controlling for learners' individual differences (gender, L2 anxiety, L2 proficiency and L1 vocabulary size) in a sample of 480 students. We found no positive dysfluency effect on either short- and long-delayed information retention or comprehension, as well as on metacognitive judgments. Furthermore, learners' individual differences did not moderate the effect. The study provides compelling evidence that the dysfluency effect on memorisation observed in L1 does not extend to L2 learning contexts. Further research in this area is needed to assess whether there is a particular combination of individual differences, dysfluent font complexity and other characteristics that will benefit learners.

Keywords: Desirable difficulty; Dysfluency effect; Fonts; Second language learning; Judgments of Learning.

Introduction

The *dysfluency effect*, a type of desirable difficulty effect, refers to the improved learning outcomes achieved through the intentional perceptual distortion of learning materials (Weissgerber & Reinhard, 2017). This manipulation was applied to maths reasoning (Meyer et al., 2015), analytical reasoning (Alter et al., 2007; Cui et al., 2022) and most often – to information retention and retrieval (Cui et al., 2021; Cui & Liu, 2022; Geller et al., 2018; Rummer et al., 2016; Taylor et al., 2020). In these studies the dysfluency effect was elicited by altering font contrast (Diemand-Yauman et al., 2011) and size (Katzir et al., 2013), employing handwriting-like fonts (French et al., 2013; Geller et al., 2018), blurring words (Rosner et al., 2015), or using specially designed dysfluent fonts such as Sans Forgetica (Eskenazi & Nix, 2021). These studies have reported positive effects of small to medium size (*Cohen's d* = 0.3 – 0.7). However, other research employing similar methodologies has not observed this effect (Geller et al., 2020; Rummer et al., 2016; Taylor et al., 2020; Wetzler et al., 2021; Yue et al., 2013), including one meta-analysis (Xie et al., 2018, but see some discussion in Weissgerber et al., 2021).

The majority of the available studies examine the dysfluency effect within the context of reading and retention of information in the first language (L1) in adult populations (87.2%; meta-analysis by Xie et al., 2018). According to previous eye-tracking research, fluent readers (most adults) process text holistically; and ignore and skip dysfluency in text. For example, skilled readers process highly predictable and short words without fixating on them (Rayner, 1998; Rayner et al., 2011). Only a handful of studies explored dysfluency effect in younger (i.e., less experienced) samples. A positive effect of dyisfluency was found for secondary school children (5th-6th and 9th-11th graders; Beege et al., 2021; French et al., 2013; Katzir et al., 2013). This pattern of results may indicate that the positive effect of dysfluency may manifest itself in less experienced readers, such as children or adults learning a new language as L2. At the same time, reading in L2 is associated with inherent challenges. For example, reading in L2 is perceived as a more complex task than reading in L1, especially if the languages differ in scripts significantly (e.g., Cyrillic vs. Latin; Bermúdez-Margaretto et al., 2022). In addition, reading in a L2 may be associated with greater anxiety (Teimouri et al., 2019). Thus, dysfluency may create "difficulty overdose" instead of desirable difficulty, if reading proficiency is too low.

To our knowledge, the only study that tested an effect of dysfluent font on learning in L2 reported no positive effect in a small sample of university students (Berezner & Gorbunova, 2021b). The current study will investigate the effect of dysfluency in a large sample of university students, accounting for a number of moderators and learners' individual differences, which are discussed below.

Moderators of the Effect

Theoretical and empirical research on dysfluency effect in L1 context has identified a number of moderators (Kühl et al., 2014; Oppenheimer & Alter, 2014) that could also influence the effect in the L2 context. For example, according to *compensatory processing account* (Mulligan, 1996), hard-to-process materials activate additional higher-level processing resources and thus lead to enhanced elaboration and retention, suggesting better performance on recall and comprehension tasks. However, research has shown mixed findings with regards to the effects of task type: recognition, recall (exact reproduction), comprehension (understanding, including non-exact reproduction), and transfer (extrapolation of information that was not explicitly stated in learning materials); on dysfluency effect. Empirical studies that report no improvement in comprehension (that require high-order processing) of dysfluent materials challenge this account (Eitel et al., 2014; Eitel & Kühl, 2016; Strukelj et al., 2016).

Some studies showed that dysfluency effect is restricted to recall tasks (Seufert et al., 2017; Weissgerber & Reinhard, 2017). However, other studies indicated that dysfluency does not affect recall (Rummer et al., 2016), including cued recall (Magreehan et al., 2016) and free recall (Berezner & Gorbunova, 2021b; Eitel et al., 2014; Eitel & Kühl, 2016; Huff et al., 2022; Strukelj et al., 2016). Some studies even found that dysfluency interfered with semantic processing and led to decrease in recall (Taylor et al., 2020; Yue et al., 2013) and increase in false memories (Sanchez & Naylor, 2018). Research with other types of tasks is scarce and showed null or negligible positive effect for transfer tasks (the tasks with supposedly deeper level of processing than recall; Eitel et al., 2014; Eitel & Kühl, 2016); and positive effect on recognition (French et al., 2013). Some studies calculate mean performance score to assess learning outcomes, without analyzing multiple-choice, open-ended and transfer items separately (e.g. Pieger et al., 2018). Thus, there is a clear gap in the understanding of what tasks performance dy~~i~~sfluency manipulation can improve. In the current study we will investigate the moderating effect of test type on dysfluency effect in L2

Another moderator of the dysfluency effect in L1 research might be the delay in testing of learnt materials: short- vs. long-delay. To date, most studies used short time delay (< 1 hour in duration; Geller et al., 2020; Lehmann et al., 2016; Rummer et al., 2016; Seufert et al., 2017; Strukelj et al., 2016; Taylor et al., 2020; Yue et al., 2013) with controversial results. Some studies show positive short-delay dysfluency effect (Beege et al., 2021; Eskenazi & Nix, 2021; French et al., 2013; Lehmann et al., 2016; Sungkhasettee et al., 2011) while many others demonstrate null effect (Berezner & Gorbunova, 2021b, 2021a; Eitel & Kühl, 2016; Geller et al., 2020; Huff et al., 2022; Rummer et al., 2016; Taylor et al., 2020). Only a handful of studies tested delayed dysfluency effects (e.g. 2 weeks in Weissgerber & Reinhard, 2017). Three of these studies found a positive dysfluency effect (Diemand-Yauman et al., 2011; Geller et al., 2018; Weissgerber & Reinhard, 2017); while one study found null dysfluency effect on recall after 1 week delay (Wetzler et al., 2021). One study also indicated that dysfluency does not enhance memory; however, it does prevent forgetting, as evidenced by comparable recall in the immediate and delayed tests in the dysfluency condition (and decline in recall in fluent condition; Weissgerber & Reinhard, 2017). However, this effect might be linked with the effect of the first test and related additional recollection, testing or practice effects (Toppino & Cohen, 2009). The current study will assess dysfluency effect in a between-participant design to avoid distortions caused by first testing, by comparing groups who had either short- or long-delayed testing.

Individual differences of learners might also moderate the dysfluency effect in L2 learning. For example, according to *compensatory processing account* (Mulligan, 1996), only those who are able to apply additional processing resources (e.g. working memory) to the task can benefit from dysfluent font (Lehmann et al., 2016; although see Strukelj et al., 2016). In contrast, people with lower resources may not benefit from dysfluency because dysfluency may create an "overdose" of difficulty for those who already struggle with the task (e.g., learners with low L2 proficiency). For example, it was shown that individuals with higher verbal ability benefit from dysfluency more than those with lower ability (Eskenazi & Nix, 2021).

On the same lines, anxiety level might also moderate dysfluency effect, especially in L2 learning. Previous studies demonstrated that anxiety impairs working memory resources and has a detrimental impact on performance on a range of tasks (Moran, 2016). Students with higher anxiety perceive learning materials as more difficult (r = .31, von der Embse et al., 2018) and reading in L2 may add to this anxiety (Teimouri et al., 2019). Coupled with the dysfluency manipulation, which is designed to make materials to be perceived as more difficult, it may result in an excessive level of difficulty, particularly in students with high anxiety levels. However, to our knowledge, previous studies have not yet tested anxiety as a moderator of dysfluency effect on learning outcomes in L1 or L2 contexts.

Individual differences in cognitive resources and anxiety are also linked to gender. Previous studies showed that females on average demonstrate higher verbal ability (Petersen, 2018; Reilly et al., 2019; Stoet & Geary, 2013), but also report higher neuroticism (worry and anxiety; Murphy et al., 2021). In light of this differences, and potential advantages of dysfluency for learners with greater resources and lower anxiety, it could be hypothesized that the effect of the dysfluency manipulation might be different for males and females. However, previous studies have not yet tested whether there are gender differences in dysfluency effect.

Finally, the precise mechanisms through which dysfluency affects learning remain unclear. The central theory, the *metacognitive account*, posits that dysfluency functions as a meta-cognitive cue of the difficulty level of the materials in question (Alter et al., 2007). The cue impacts self-perceived estimates of memory and performance – Judgments of Learning (JOLs; Pieger et al., 2016), likely reducing them or rendering them more accurate. Overall, the research indicated that participants in dysfluent conditions report lower JOLs and show longer reading times in comparison to participants in fluent conditions (see meta-analysis by Xie et al., 2018). However, it is not clear whether the JOLs are more accurate in dysfluent conditions and whether they actually affect memory performance (Magreehan et al., 2016; Sungkhasettee et al., 2011; Yue et al., 2013). The current study will test whether this theoretical explanation is indeed responsible for the effect observed in research.

Present Study

The current study aims to investigate dysfluency effect in L2 learning. We extended previous studies by examining dysfluency effect on information retention and comprehension in L2, as well as several potential moderators of this effect, including learners' individual differences. The primary objective of the present study was to reproduce the dysfluency effect, observed in L1, in L2 learners.

Hypothesis **H1a** tests whether dysfluent font improves the overall memorization of information in L2 in comparison to control conditions (fluent fonts). Hypothesis **H1b** tests whether the impact of dysfluent font on task performance varies as a function of learners' L2 proficiency. Specifically, it is expected that participants with greater L2 proficiency will benefit more from the dysfluent condition in comparison to those with lesser proficiency.

As with studies on L1, our hypothesis **H2a** seeks to ascertain whether the impact of dysfluent font on task performance varies depending on the specific task at hand. Specifically, it is expected that dysfluency will improve recognition performance, but not recall performance in L2. In light of the findings from the L1 context, our hypothesis **H2b** tests whether the impact of dysfluent font on task performance in L2 varies depending on delay between learning and testing. Specifically, it is expected that dysfluency will improve long-delay performance, but not short-delay performance.

To replicate findings in L1, hypothesis **H3a** tests whether the impact of dysfluent font on task performance varies depending on learners' verbal ability. Specifically, it is expected higher performance in dysfluent condition in participants with higher verbal ability in comparison to participants with lower ability. Our hypothesis **H3b** is that the impact of dysfluent font on task performance varies depending on learners' anxiety level. Specifically, it is expected higher performance in dysfluent condition in participants with lower anxiety in comparison to participants with higher anxiety. Our hypothesis **H3c** tests whether the impact of dysfluent font on task performance varies depending on learners' gender. Given that previous research reported higher verbal ability and higher anxiety in females, which presumably should have opposite effects on performance, this hypothesis is not a directional one.

Finally, in accordance with the metacognitive account, we hypothesized that **(H4a)** participants in the dysfluent condition will rate the task as more difficult than those in the control groups and **(H4b)** participants in the dysfluent condition will have lower self-reported estimates of performance than those in the control groups.

In order to test the above hypotheses, a three (dysfluency) by two (test time) between-participant experiment was conducted in ecologically valid conditions. The between-participant design was chosen to eliminate potential testing effect in Short-delay vs. Long-delay test condition (Wenzel & Reinhard, 2019).

Our experimental conditions consisted of Arial font and OpenDyslexic for control conditions and Sans Forgetica font as experimental condition. Arial font is a standard easy-to-read font previously used in research (Sanchez & Naylor, 2018; Steindorf et al., 2023; Strukelj et al., 2016; Taylor et al., 2020). The OpenDyslexic font was developed with the objective of improving the legibility of written text for individuals with reading difficulties. One study demonstrated improved reading patterns (i.e. visual ease) with OpenDyslexic in individuals both with and without dyslexia (Franzen et al., 2019). Conversely, another study found no advantage of OpenDyslexic on objective (reading rate or accuracy) or subjective (preferences) levels (Wery & Diliberto, 2017). Sans Forgetica, conversely, was designed to improve memory retention by introducing perceptual dysfluency, which theoretically engages deeper cognitive processing and slows reading down (Earp, 2018). However, the majority of empirical studies have not demonstrated its positive effect on retention (Berezner & Gorbunova, 2021b; Cushing & Bodner, 2022; Eskenazi & Nix, 2021; Geller et al., 2020; Taylor et al., 2020; Wetzler et al., 2021), and some studies have indicated that its efficacy may be contingent upon specific conditions, such as the absence of testing expectations (Geller & Peterson, 2021). In this study, we employ OpenDyslexic and Sans Forgetica to compare their effects on reading and memory, motivated by their distinct design objectives and features and the mixed findings in existing literature. We further aimed to explore conditional factors and clarify the applicability of these fonts in L2 learning context.

Materials and Methods

Participants

The study involved 480 university students (176 females; 8 participants did not indicate their gender; 17-37 years, *Mean age* = 19, *SD* = 1.93). All participants were native Russian language speakers (L1), learning English as a foreign language (L2) at school (starting from 2nd grade) and at the university. L2 proficiency was assessed both via level assigned after standardized knowledge test on the university entry (analogous to TOEFL) and via self-perceived L2 level on four items asking participants to estimate how well they think they can read, write, speak, and comprehend in English (on a scale from 1 – poorly to 5 – very well). Figure 1 illustrates the distribution of participants according to their level of L2 proficiency assessed with standardized knowledge test. The sample exhibited a proficiency level between A2 (pre-intermediate) and C1 (advanced) levels of L2 proficiency according to the Common European Framework of Reference for Languages, with 95.5 per cent of the sample falling within this range. Of this group, 40.6 per cent exhibited a B2 (upper-intermediate) level of proficiency.

[Figure 1 near here]

Procedure

Participants provided their informed consent before data collection. Participants did not receive any compensation for participation. The study was approved by the Tomsk State University Interdisciplinary Research Ethics Committee (№ 12022020-30).

The data were collected via an online form in ecologically valid conditions, either as part of the students' English language classes or as a part of homework assignment. Participants were randomly assigned to either short-delay or long-delay test condition. Subsequently, participants were randomly assigned to one of the three dysfluency and testing conditions (see

Table 1 for conditions and *N*s of participants in each condition). All participants completed the recognition questions and the recall tasks (within-participant condition). The procedure took approximately 45 minutes. The experiment consisted of five stages (shown on **Figure 2**).

[Figure 2 near here]

Half of the participants performed the Stages 4 and 5 right after the filler tasks (Short-delay test group). The other half of the participants completed memory test and subjective questions two weeks after Stage 3 (Long-delay test group). To ensure participants completed Stages 4 and 5 at the same time, they received the link for these stages only after 2 weeks have passed. Participants could not access the text after they finished Stages 1-3.

[Table 1 near here]

As can been seen from Table 1, Font groups significantly differed in *N*s as can be expected by random condition allocation. Further, attrition in Long-delay test varied from 6 % to 29 % with greater attrition in Experimental group.

Materials

Learning Materials

All students read the same text (858 words) in English on ground water, which was utilized in previous studies (Geller et al., 2020; Yue et al., 2015). Reading was self-paced. In Control 1 group, the whole text was typed in Arial font. For Control 2 and Experimental condition, ten critical phrases (1.6% of the whole text volume), each containing a different keyword, were selected from the passage (e.g., the term *permeability* was the keyword in the phrase: “permeability is a measure of how well the spaces are connected”) and were presented in OpenDyslexic (Control 2) or Sans Forgetica (Experimental), while all other text in these two conditions were written in Arial. The presentation of the entire text in Sans Forgetica was deemed to be unduly challenging. To match the conditions, in Control condition 2 also only critical phrases were modified.

Memory Test - Recognition

A memory test to assess text information retention and comprehension included ten multiple-choice (recognition) questions and ten open (recall) questions (see Supplementary materials). Example item for recognition: “Water seeping down from the land surface adds to the ground water and is called ______ water”: a) recharge; b) table; c) ground; d) leaking. Each multiple-choice question had one correct response option and three distractor options. The participants were given 1 point for each correct response. The total sum of correct responses was used as a *Recognition accuracy*. Internal consistency of the recognition test was moderate (Cronbach’s α = .67; McDonald's ω = .73).

Memory Test - Recall

Each open question (see Supplementary materials) was formulated as a phrase with a missing word (e.g., “An extended period of dry weather may decrease recharge and cause the water table to _____.”). Participants were asked to type in a correct answer (a word, a collocation, or a number).

Two variables were created, using the following scoring procedures:

1) *Exact recall.* Participants were given 1 point only in case they typed the same word that appeared in the text. Internal consistency of the exact recall test was high (Cronbach’s α = .77; McDonald's ω = .84).
2) *Synonym recall.* Participants were given 1 point for each response that was either the same word as they saw in the text (e.g., “fall”) or a synonym (e.g., “drop down” or “decrease”). Internal consistency of the synonym recall test was high (Cronbach’s α = .81; McDonald's ω = .84).

Foreign Language Anxiety

A foreign language anxiety questionnaire (Yim, 2014) consisting of 14 items was used to assess level of worry during language classes and task performance in L2. Example item: "I feel anxious when the teacher asks me a question that I have not prepared for." Four items had the reversed coding. The sum of all responses varied from 14 to 70. The test was adapted to Russian using the translation/back translation procedure. Reliability of the questionnaire (*Cronbach's alpha*) in the current study was .89.

Verbal Ability

A short version of "MyVocab" test was used to assess vocabulary size in L1 as a proxy to verbal ability (Maslennikova et al., 2017). The participants were presented with 99 words in Russian, from which four were pseudowords. Participants were asked to mark all words for which they can explain at least one meaning. All the marked words except pseudowords were given 1 point. A sum of all marked items was used as a total score. Verbal ability was assessed in L1 because it was shown to moderate L2 learning (Sparks et al., 2009, 2019; Zaretsky, 2014).

Subjective Evaluation of the Text

Three questions about text difficulty and the pleasure of working with it were used to assess fluency/dysfluency of the material (see questions 1-3 in Supplementary materials).

Self-reported Accuracy

Self-reported accuracy was assessed using two items (see questions 4-5 in Supplementary materials). Item 4 assesses self-reported memory for the text overall, while item 5 assesses self-reported performance on the test questions. Retrospective subjective estimates of memory and performance were used as previous studies suggested that judgments of learning at the time of study reduce dysfluency effect on memory (Rosner et al., 2015).

Statistical Analysis

All the analyses were conducted using R (version 4.3.3; RStudio Team, 2020) and JASP (Version 0.19.0.0; JASP Team, 2024). In light of previous studies showing null effect of dy~~i~~sfluency on learning and the fact that the non-significant result in null hypothesis significance testing (NHST) cannot be interpreted as an absence of effect (Lakens, 2017), we employed Bayesian options for Repeated measures ANOVA and ANCOVA available in JASP.

To test the hypotheses **H1a, H2a, H2b** we conducted 3 (Dysfluency) by 2 (Test time) Repeated Measures Bayesian ANOVA with Question type (Recognition, Exact recall, Synonym recall) as a within-participant factor.

Further, for hypotheses **H1b, H3a, H3b and H3c** we explored interactions between Dysfluency and covariates (gender, L2 anxiety, L2 proficiency and verbal ability) in a follow-up 3-way (Condition) Bayesian ANCOVA for Short-delay group. For Long-delay group, these covariates could not be included, as the data collection was fully anonymized. This meant that for the group in the Long-delay condition their memory test performance was not linked with their data that was collected during Stage 1-3.

Finally, we conducted a series of Kruskal-Wallis χ^2 tests and correlational analysis to test hypotheses **H4a** and **H4b**.

Results

Descriptive statistics of memory performance in each condition are presented in Table 2 and raincloud plots are presented on **Figure S1** in Supplementary materials.

[Table 2 near here]

Objective Memory Performance

First, we ran a 3 (Dysfluency) by 2 (Test time: Short- vs Long-delay) Repeated Measures Bayesian ANOVA with Question type (Recognition*,* Exact recall*,* and Synonym recall) as a within-participant factor. The Bayesian analysis revealed that the null model (i.e., the model assuming no effects of Dysfluency, Test Time, or their interaction on Question Type) was more likely than any other model considered. Specifically, the Bayes factor (BF_{01}) comparing the null model to the model including the main effects of Dysfluency, Test Time, and Question Type was $BF_{01} = 1.593\times10^{-64}$, indicating that the data were more likely under the null model compared to the alternative model. Similarly, the Bayes factors (BF_{01}) for models including the interactions between Test Time and Dysfluency and Question type and Dysfluency were also smaller than for the null model (all $BF_{01} < 1$).

These results suggest that there is very strong evidence in favor of the null model, implying that neither Dysfluency, Test Time, Question Type, nor their interactions had a substantial effect on the participants' performance. The detailed results of model comparison are presented in **Table S1** (BF01) **and Table S2** (BF10) in Supplementary materials.

Regular 3 (Dysfluency) by 2 (Test time: Short- vs Long-delay) MANOVA with three dependent variables: Recognition, Exact recall, and Synonym recall yielded no interaction between Dysfluency and Test time (*Wilks'* $\lambda = 0.994$, F (6, 796) = 0.358, *ns*). The main effect of Dysfluency was also not significant (*Wilks'* $\lambda = 0.982$, F (6, 796) = 1.229, *ns*). Main effect of Test time was significant (*Wilks'* $\lambda = 0.930$, F (3, 398) = 9.970, $p < .001$), with higher scores for Short-delay test group then for Long-delay test group. Additional univariate analyses (one-way ANOVA) showed that this effect was only significant for Recognition questions (η^2_p = 0.044).

Further, we assessed correlations between three types of memory tests and possible moderators of the dysfluency effect on performance. We found strong intercorrelations between all three memory tests (.67-.96, all *p*s < .001) and all the three tests were associated with L2 proficiency (r = .30 - .36, all *p*s < .001). Performance on memory tests was not linked to participants' verbal ability or anxiety. See **Figure 3** for details.

[Figure 3 near here]

We conducted three Bayesian ANCOVAs for Recognition, Exact recall, and Synonym recall in Short-delay testing condition, including L2 proficiency, L2 anxiety, verbal ability and gender as covariates in the model. According to Bayes factor, null model was more likely than any other model that included suggested moderators in all three ANCOVAs (all $BF_{01} < 1$; see Supplementary materials, Tables S3-S5 for model comparisons).

Subjective Evaluation of the Text

We compared participants' font and text enjoyment across the three conditions using the Kruskal-Wallis χ^2 test with Dunn's post-hoc pairwise comparisons. We found significant differences in participants' ratings of font enjoyment (Kruskal-Wallis $\chi^2 = 11.673$, $p < 0.01$), with Sans Forgetica rated as less enjoyable than Arial; and OpenDyslexic as less enjoyable than Sans Forgetica (see **Figure 4**). Other comparisons (i.e., text enjoyment and text difficulty) were not significant. It should be noted that majority of people selected "Did not pay attention" option.

[Figure 4 near here]

Self-reported Estimates

We compared participants' memory estimates across the three conditions using the Kruskal-Wallis χ^2 test with Dunn's post-hoc pairwise comparisons. The results showed null effect of Dysfluency on metacognitive judgement of memory performance, namely, self-reported memory for the text and self-reported performance on the test questions.

Discussion

The present study investigated the effect of dysfluency on L2 information retention and comprehension. Using Bayesian approach, we found strong evidence in favor of a null effect of font manipulation on both short- and long-term performance of all three types of tasks. Our results are in line with previous studies that showed no positive effect of dysfluency on learning in L1 using NHST approach (see a recent meta-analysis; Xie et al., 2018). We also extended these results by showing that more fluent fonts (e.g., OpenDyslexic) also have no effect on information retention. None of the hypothesized moderators of dysfluency effect on memory performance were significant (gender, L2 anxiety, verbal ability), which is in line with aforementioned meta-analysis (Xie et al., 2018).

Null effect of dysfluency manipulation on short-term and long-term retention is also in line with recent findings (Wetzler et al., 2021). Previous findings of positive dysfluency effect on long-term retention might be explained by re-learning of the material during classes (e.g. Diemand-Yauman et al., 2011). Another explanation might be the testing effect, where one test serves as a learning event and results in better performance on the second test (Weissgerber & Reinhard, 2017). In our study we removed these confounds by using a between-participant design for Test time condition; and using a text content that is not covered by the curriculum.

Power limitations are unlikely to explain the observed results of the current experiment as previous studies showed dysfluency effect in smaller samples (Alter et al., 2007; Beege et al., 2021; Eskenazi & Nix, 2021; French et al., 2013). First explanation for such results could be that highlighting certain elements in the text with the Sans Forgetica font was not difficult enough to produce a strong dysfluency effect. In other words, the limited distortion of the learning materials with the font was unable to trigger an increase in cognitive processing or serve as a metacognitive cue to the difficulty of the materials. Although participants in the current study rated Sans Forgetica as less pleasant (and OpenDyslexic less pleasant than Sans Forgetica), participants' ratings of the difficulty and pleasantness of the text in our study did not differ. However, it may be premature to judge that the manipulation was “not difficult”. The subjective estimates of processing fluency, used in this study, might be biased (Dunlosky & Mueller, 2016). More robust measures such as gaze features (Strukelj et al., 2016) or study time (Eitel et al., 2014) might give more accurate estimates of learning materials’ difficulty. An indirect indication of the materials' dysfluency in the present study is evidenced by the attrition rate in the Sans Forgetica condition, which was higher than in the control conditions (28% vs. 12% and 16%, respectively). This could be an indirect indication of the unappealing nature of Sans Forgetica, potentially leading to participants declining to participate in an "unpleasant study."

Second explanation for the reported null effect comes from the use of a mixture of two fonts in Control group 2 and the experimental conditions. Simultaneous presentation of fluent and dysfluent font within one text might create a “distinctiveness effect” (Rummer et al., 2016; Sung et al., 2022; Wetzler et al., 2021) whereby attracting participants’ attention to the information presented in these fonts. While the effect of the font's distinctiveness may overshadow the dysfluency effect, this could not be used as an explanation of no differences between Experimental condition and Control 1 condition where the whole text was presented in one font. The highlighting role of fonts mixture should be explicitly tested in future studies, e.g., by estimating memory for highlighted and non-highlighted items presented in different fonts within one text.

Thirdly, previous studies suggest that the dysfluency effect is influenced by test expectancy. Specifically, dysfluency effect occurs only when participants are not informed about an upcoming memory test (Geller & Peterson, 2021). In other words, test expectancy serves as a metacognitive cue that leads to deeper learning (and therefore dy~~i~~sfluency does not have added benefit in this condition; Geller & Peterson, 2021). Thus, in the present study, notifying participants about the upcoming test may have diminished the effect of font on learning outcomes.

Limitations and future directions

This study addressed limitations of the previous research of dysfluent fonts by implementing between-participants design for short- and long-delay memory testing and collecting larger sample size. However, several limitations must be mentioned regarding the current study. First limitation is linked to anonymization procedure used in the current study that did not allow us to test moderators in long-delay test subgroup. Future studies might use between-participants design to test moderators' effects also on delayed performance. Secondly, dropout rate in Experimental condition deserves attention. The absence of almost 30% of participants who saw the text in dysfluent font might bias the results. Thirdly, the Recognition, Exact Recall and Synonym Recall tests had high internal consistency. Some studies have shown that higher internal consistency for a test in a particular group can lead to higher scores in that group (Ishikawa, 2023). However, our data showed that the tasks were sufficiently complicated, as indicated by absence of ceiling effect (Table 2).

Fourthly, we tested general L2 anxiety as a moderator of the dysfluency effect with null result. More specific measures, such as L2 reading anxiety scale (e.g. FLRAS; Saito et al., 1999), may demonstrate more complex relations to dysfluency in the text.

Conclusion

To sum up, our study provides compelling evidence that the dysfluency effect observed in L1 reading does not extend to L2 learning contexts. Despite some previous research suggesting that difficult-to-read fonts can enhance memory and comprehension, our findings indicate that such effects are absent when controlling for individual learner differences among L2 learners. The results of this study suggest that it is premature to implement dysfluent fonts (such as Sans Forgetica) or more fluent fonts (such as OpenDyslexic) in L2 education in order to improve learning. Further research is needed to test different dysfluent fonts, including against each other, to investigate the effect of other perceptually degraded manipulations (e.g., blurred letters) on the L2 learning.

Acknowledgements: The work of Elina Tsigeman on this paper was partially supported by the Basic Research Program at the National Research University Higher School of Economics (HSE University)

Declaration of interest statement: The authors report there are no competing interests to declare.

Data availability statement: Derived data supporting the findings of this study are available from the corresponding author Y.K. on request.

References

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

Beege, M., Krieglstein, F., Schneider, S., Nebel, S., & Rey, G. D. (2021). Is There a (Dis-)Fluency Effect in Learning With Handwritten Instructional Texts? Evidence From Three Studies. *Frontiers in Education*, *6*, 678798. https://doi.org/10.3389/feduc.2021.678798

Berezner, T., & Gorbunova, E. (2021a). The Presence of Images Does Not Moderate the Disfluency Effect. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3960675

Berezner, T., & Gorbunova, E. (2021b). *Улучшение запоминания текста посредством изменения шрифтов: Исследование Sans Forgetica*. *14*(78). https://doi.org/10.54359/ps.v14i78.155

Bermúdez-Margaretto, B., Kopytin, G., Myachykov, A., Fu, Y., Pokhoday, M., & Shtyrov, Y. (2022). Biliteracy and acquisition of novel written words: The impact of phonological conflict between L1 and L2 scripts. *Psychological Research*, *86*(3), 871–890. https://doi.org/10.1007/s00426-021-01529-y

Cui, L., Harrison, L., & Liu, J. (2022). *Sans Forgetica font does not improve Cognitive Reflection Test performance despite differences in legibility*. PsyArXiv. https://doi.org/10.31234/osf.io/qdxw3

Cui, L., & Liu, J. (2022). Recognition of Studied Words in Perceptual Disfluent Sans Forgetica Font. *Vision*, *6*(3), Article 3. https://doi.org/10.3390/vision6030052

Cui, L., Liu, J., & Liu, Z. (2021). *Sans Forgetica font may help memory for words but not for numbers, nor does it help analytical thinking*. PsyArXiv. https://doi.org/10.31234/osf.io/a7n2r

Cushing, C., & Bodner, G. E. (2022). Reading aloud improves proofreading (but using Sans Forgetica font does not). *Journal of Applied Research in Memory and Cognition*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/mac0000011

Darren George, & Mallery, P. (2003). *SPSS for Windows Step by Step: A Simple Guide and Reference, 11.0 Update*. Allyn and Bacon. https://books.google.ru/books?id=AghHAAAAMAAJ

Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2011). Fortune favors the Bold (and the Italicized): Effects of disfluency on educational outcomes. *Cognition*, *118*(1), 111–115. https://doi.org/10.1016/j.cognition.2010.09.012

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, October 8). *Q&A: Designing a font to help students remember key information* [Interview]. https://www.teachermagazine.com/au_en/articles/qa-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

Eitel, A., Kühl, T., Scheiter, K., & Gerjets, P. (2014). Disfluency Meets Cognitive Load in Multimedia Learning: Does Harder-to-Read Mean Better-to-Understand?: Disfluency in multimedia learning. *Applied Cognitive Psychology*, *28*(4), 488–501. https://doi.org/10.1002/acp.3004

Eskenazi, M. A., & Nix, B. (2021). Individual differences in the desirable difficulty effect during lexical acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(1), 45–52. https://doi.org/10.1037/xlm0000809

Franzen, L., Stark, Z., & Johnson, A. (2019). AB004. The dyslexia font OpenDyslexic facilitates visual processing of text and improves reading comprehension in adult dyslexia. *Annals of Eye Science*, *4*, AB004–AB004. https://doi.org/10.21037/aes.2019.AB004

French, M. M. J., Blood, A., Bright, N. D., Futak, D., Grohmann, M. J., Hasthorpe, A., Heritage, J., Poland, R. L., Reece, S., & Tabor, J. (2013). Changing Fonts in Education: How the Benefits Vary with Ability and Dyslexia. *The Journal of*

Educational Research, *106*(4), 301–304. https://doi.org/10.1080/00220671.2012.736430

Geller, J., Davis, S. D., & Peterson, D. J. (2020). Sans Forgetica is not desirable for learning. *Memory*, *28*(8), 957–967. https://doi.org/10.1080/09658211.2020.1797096

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*, *47*(12), 1924–1938. https://doi.org/10.1037/xlm0001042

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 & Cognition*, *46*(7), 1109–1126. https://doi.org/10.3758/s13421-018-0824-6

Huff, M. J., Maxwell, N. P., & Mitchell, A. (2022). Distinctive Sans Forgetica font does not benefit memory accuracy in the DRM paradigm. *Cognitive Research: Principles and Implications*, *7*(1), 102. https://doi.org/10.1186/s41235-022-00448-9

Ishikawa, T. (2023). Individual Differences and Skill Training in Cognitive Mapping: How and Why People Differ. *Topics in Cognitive Science*, *15*(1), 163–186. https://doi.org/10.1111/tops.12605

JASP Team. (2024). *JASP (Version 0.19.0)[Computer software]* (Version 0.19.0) [Computer software]. https://jasp-stats.org/

Katzir, T., Hershko, S., & Halamish, V. (2013). The Effect of Font Size on Reading Comprehension on Second and Fifth Grade Children: Bigger Is Not Always Better. *PLoS ONE*, *8*(9), e74061. https://doi.org/10.1371/journal.pone.0074061

Kühl, T., Eitel, A., Scheiter, K., & Gerjets, P. (2014). A Call for an Unbiased Search for Moderators in Disfluency Research: Reply to Oppenheimer and Alter (2014). *Applied Cognitive Psychology*, *28*(5), 805–806. https://doi.org/10.1002/acp.3030

Lakens, D. (2017). Equivalence Tests. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

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

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

Maslennikova, E. P., Feklicheva, I. V., Esipenko, E. A., Sharafieva, K. R., Ismatullina, V. I., Golovin, G. V., Miklashevskiy, A. A., Chipeeva, N. A., & Soldatova, E. L. (2017). Vocabulary Size as a Verbal Intelligence Index: Application of the Express Methods of an Estimating Vocabulary Size. *Bulletin of the South Ural State University Series 'Psychology'*, *10*(3), 63–69. https://doi.org/10.14529/psy170306

Meyer, A., Frederick, S., Burnham, T. C., Guevara Pinto, J. D., Boyer, T. W., Ball, L. J., Pennycook, G., Ackerman, R., Thompson, V. A., & Schuldt, J. P. (2015). Disfluent fonts don't help people solve math problems. *Journal of Experimental Psychology: General*, *144*(2), e16–e30. https://doi.org/10.1037/xge0000049

Moran, T. P. (2016). Anxiety and working memory capacity: A meta-analysis and narrative review. *Psychological Bulletin*, *142*(8), 831–864. https://doi.org/10.1037/bul0000051

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

Murphy, S. A., Fisher, P. A., & Robie, C. (2021). International comparison of gender differences in the five-factor model of personality: An investigation across 105 countries. *Journal of Research in Personality*, *90*, 104047. https://doi.org/10.1016/j.jrp.2020.104047

Oppenheimer, D. M., & Alter, A. L. (2014). The Search for Moderators in Disfluency Research: Moderators of disfluency. *Applied Cognitive Psychology*, *28*(4), 502–504. https://doi.org/10.1002/acp.3023

Petersen, J. (2018). Gender Difference in Verbal Performance: A Meta-analysis of United States State Performance Assessments. *Educational Psychology Review*, *30*(4), 1269–1281. https://doi.org/10.1007/s10648-018-9450-x

Pieger, E., Mengelkamp, C., & Bannert, M. (2016). Metacognitive judgments and disfluency – Does disfluency lead to more accurate judgments, better control, and better performance? *Learning and Instruction*, *44*, 31–40. https://doi.org/10.1016/j.learninstruc.2016.01.012

Pieger, E., Mengelkamp, C., & Bannert, M. (2018). Disfluency as a Desirable Difficulty—The Effects of Letter Deletion on Monitoring and Performance. *Frontiers in Education*, *3*, 101. https://doi.org/10.3389/feduc.2018.00101

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(2), 514–528. https://doi.org/10.1037/a0020990

Reilly, D., Neumann, D. L., & Andrews, G. (2019). Gender differences in reading and writing achievement: Evidence from the National Assessment of Educational Progress (NAEP). *American Psychologist*, *74*, 445–458. https://doi.org/10.1037/amp0000356

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

RStudio Team. (2020). [Computer software]. http://www.rstudio.com/

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

Saito, Y., Garza, T. J., & Horwitz, E. K. (1999). Foreign Language Reading Anxiety. *The Modern Language Journal*, *83*(2), 202–218. https://doi.org/10.1111/0026-7902.00016

Sanchez, C. A., & Naylor, J. S. (2018). Disfluent presentations lead to the creation of more false memories. *PLOS ONE*, *13*(1), e0191735. https://doi.org/10.1371/journal.pone.0191735

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

Sparks, R. L., Patton, J., Ganschow, L., & Humbach, N. (2009). Long-Term Crosslinguistic Transfer of Skills From L1 to L2. *Language Learning*, *59*(1), 203–243. https://doi.org/10.1111/j.1467-9922.2009.00504.x

Sparks, R. L., Patton, J., & Luebbers, J. (2019). Individual differences in L2 achievement mirror individual differences in L1 skills and L2 aptitude: Crosslinguistic transfer of L1 to L2 skills. *Foreign Language Annals*, *52*(2), 255–283. https://doi.org/10.1111/flan.12390

Steindorf, L., Pink, S., Rummel, J., & Smallwood, J. (2023). When there is noise on Sherlock Holmes: Mind wandering increases with perceptual processing difficulty during reading and listening. *Cognitive Research: Principles and Implications*, *8*(1), 31. https://doi.org/10.1186/s41235-023-00483-0

Stoet, G., & Geary, D. C. (2013). Sex differences in mathematics and reading achievement are inversely related: Within-and across-nation assessment of 10 years of PISA data. *PloS One*, *8*(3), e57988.

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

Sung, B., Vanman, E. J., & Hartley, N. (2022). Revisiting (dis)fluency: Metacognitive difficulty as a novelty cue that evokes feeling-of-interest. *Psychology & Marketing*, *39*(8), 1451–1466. https://doi.org/10.1002/mar.21664

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. 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*, *28*(7), 850–857. https://doi.org/10.1080/09658211.2020.1758726

Teimouri, Y., Goetze, J., & Plonsky, L. (2019). Second language anxiety and achievement: A meta-analysis. *Studies in Second Language Acquisition*, *41*(2), 363–387. https://doi.org/10.1017/S0272263118000311

Toppino, T. C., & Cohen, M. S. (2009). The Testing Effect and the Retention Interval: Questions and Answers. *Experimental Psychology*, *56*(4), 252–257. https://doi.org/10.1027/1618-3169.56.4.252

von der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders*, *227*, 483–493. https://doi.org/10.1016/j.jad.2017.11.048

Weissgerber, S. C., Brunmair, M., & Rummer, R. (2021). Null and Void? Errors in Meta-analysis on Perceptual Disfluency and Recommendations to Improve Meta-analytical Reproducibility. *Educational Psychology Review*, *33*(3), 1221–1247. https://doi.org/10.1007/s10648-020-09579-1

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

Wenzel, K., & Reinhard, M.-A. (2019). Relatively unintelligent individuals do not benefit from intentionally hindered learning: The role of desirable difficulties. *Intelligence*, *77*, 101405. https://doi.org/10.1016/j.intell.2019.101405

Wery, J. J., & Diliberto, J. A. (2017). The effect of a specialized dyslexia font, OpenDyslexic, on reading rate and accuracy. *Annals of Dyslexia*, *67*(2), 114–127. https://doi.org/10.1007/s11881-016-0127-1

Wetzler, E. L., Pyke, A. A., & Werner, A. (2021). Sans Forgetica is Not the “Font” of Knowledge: Disfluent Fonts are Not Always Desirable Difficulties. *SAGE Open*, *11*(4), 21582440211056624. https://doi.org/10.1177/21582440211056624

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

Yim, S. Y. (2014). An anxiety model for EFL young learners: A path analysis. *System*, *42*, 344–354. https://doi.org/10.1016/j.system.2013.12.022

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 & Cognition*, *41*(2), 229–241. https://doi.org/10.3758/s13421-012-0255-8

Yue, C. L., Storm, B. C., Kornell, N., & Bjork, E. L. (2015). Highlighting and Its Relation to Distributed Study and Students' Metacognitive Beliefs. *Educational Psychology Review*, *27*(1), 69–78. https://doi.org/10.1007/s10648-014-9277-z

Zaretsky, E. (2014). The role of L1 and L2 reading on L1 preservation and positive cross-linguistic transfer among sequential bilinguals. *Written Language & Literacy*, *17*(1), 139–164. https://doi.org/10.1075/wll.17.1.07zar

Tables

Table 1. Experimental conditions and numbers of participants in each group

Font	Short-delay test, N	Long-delay test, N
Control group		
Fluent Arial 14-point font *	75	66
Lorem ipsum dolor sit amet, consectetur adipiscing elit.		
Control group 2		
Dyslexia-friendly OpenDyslexic * 12-point	75	63
Lorem ipsum dolor sit amet, consectetur adipiscing elit.		
Experimental group Dysfluent Sans Forgetica 12-point *	89	64

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

---

* Different sizes of typeface were chosen to visually equalize sizes of words in different fonts; non-parametric Kruskal-Wallis One-way ANOVA demonstrated no difference in L2 proficiency between experimental groups.

Table 2. Descriptive statistics (means and standard deviations) of objective memory performance for each group

Condition		Measure	Mean	SD
		Recognition	5.95	2.57
	Arial	Exact recall	3.71	2.89
		Synonym recall	4.56	3.26
		Recognition	5.45	2.46
Short-delay	OpenDyslexic	Exact recall	3.45	3.04
		Synonym recall	4.12	3.3
		Recognition	5.51	2.56
	Sans Forgetica	Exact recall	3.39	2.45
		Synonym recall	4.3	2.81
		Recognition	4.86	2.6
	Arial	Exact recall	2.79	2.94
		Synonym recall	3.71	3.29
		Recognition	4.41	2.05
Long-delay	OpenDyslexic	Exact recall	2.64	2.78
		Synonym recall	3.64	3.33
		Recognition	4.56	2.59
	Sans Forgetica	Exact recall	3.25	3.15
		Synonym recall	4.23	3.54

Note: minimum and maximum scores for all measures varied from 0 to 10; Skewness and kurtosis of all measures varied within acceptable range (i.e. below cut-off of 2 as recommended by (George & Mallery, 2003).

1

Figures

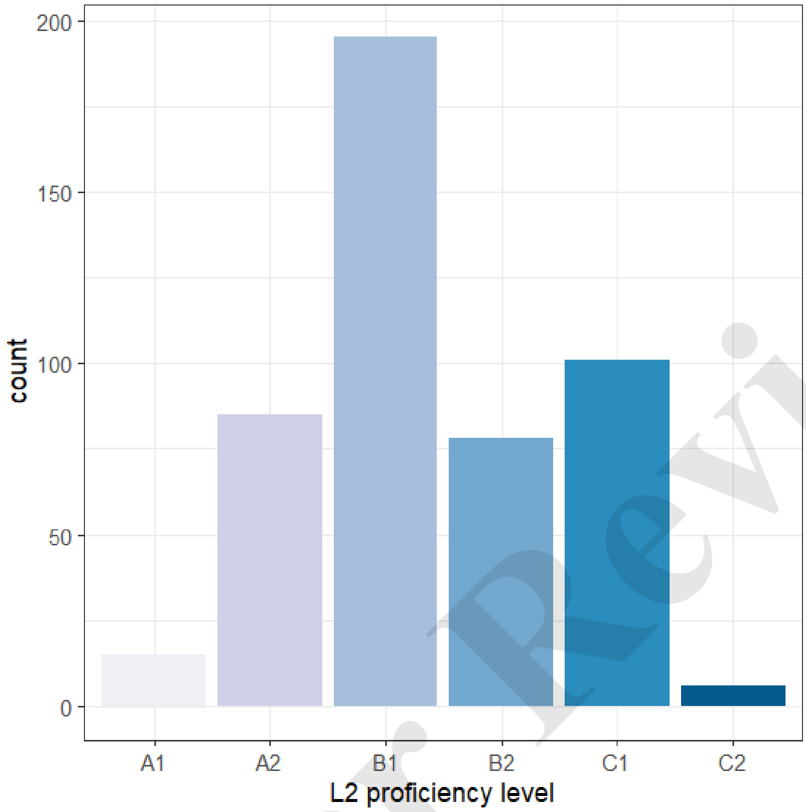

Figure 1. Participants' L2 proficiency levels

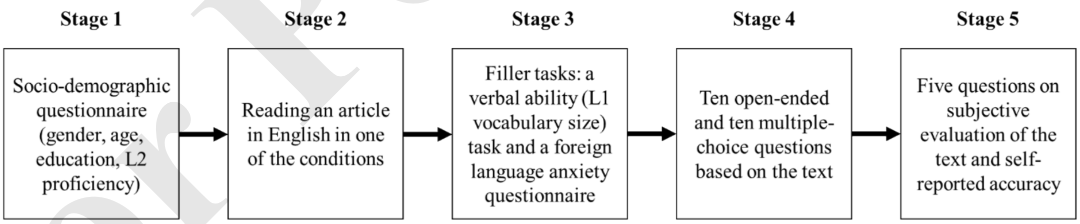

Figure 2. Schema of experimental procedure

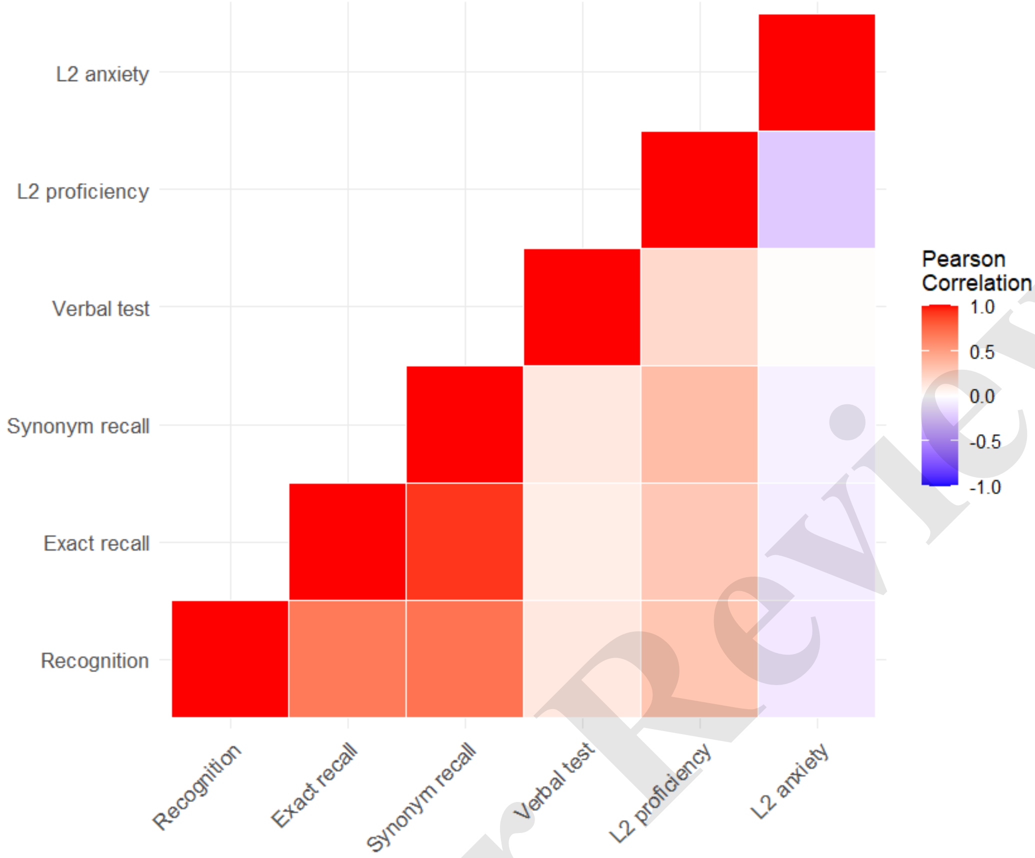

Figure 3. Correlations between three types of memory tests and possible moderators

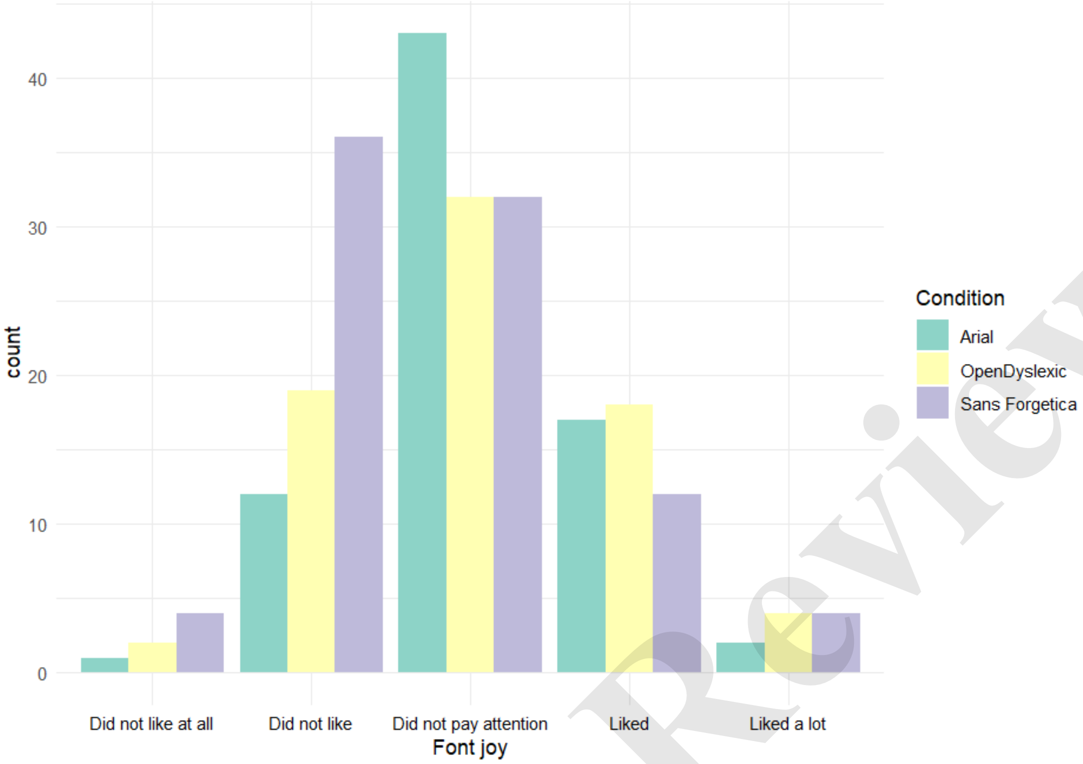

Figure 4. Participants' estimates of font enjoyment in three conditions

Figure captions

Figure 1. Participants' L2 proficiency levels.

Figure 2. Schema of experimental procedure

Figure 3. Correlations between three types of memory tests and possible moderators

Figure 4. Participants' estimates of font enjoyment in three conditions

Supplementary materials

Full list of Knowledge test items.

Recognition questions

1. __________ is a measure of how well the spaces are connected.

a) Porosity

b) Permeability

c) Evaporation

d) Seeping

2. Water seeping down from the land surface adds to the ground water and is called _______ water.

a) recharge

b) table

c) ground

d) leaking

3. The spaces in a gravel aquifer are called__________.

a) cracks

b) wells

c) pollutants

d) pores

4. The spaces in a fractured rock aquifer are called __________.

a) cracks

b) pores

c) fractures

d) wells

5. Water leaving an aquifer is called __________ water.

a) recharge

b) discharge

c) ground

d) leaking

6. If a material contains pores that are not connected, it is said to be __________.

a) unusable

b) shallow

c) leaky

d) impermeable

7. More than __________ percent of the people in the United States use ground water for drinking and other household uses.

a) 50

b) 20

c) 90

d) 70

8. The water table under a hillside is usually __________.

a) shallow

b) deep

c) unusable

d) impermeable

9. A well is usually a __________ in the ground that fills with ground water.

a) pore

b) pump

c) pipe

d) crack

10. Heavy rains or melting snow may increase the recharge and cause the water table to ____.

a) fall

b) leak

c) rise

d) flow

Exact and Synonym Recall questions

1. The top of the water in the soil, sand, or rocks is called the ___________________.

2. The amount of spaces is the ___________________.

3. Ground water can be obtained by drilling or digging ___________________.

4. The water that fills the empty spaces and cracks is called ___________________.

5. ___________________ is the name given to the underground soil or rock through which groundwater can easily move.

6. Some wells, called ___________________ wells, do not need a pump.

7. An extended period of dry weather may decrease recharge and cause the water table to ___________________.

8. Ground water can be brought to the land surface by a ___________________.

9. Ground water is also used in some way by about ___________________ percent of cities and by many factories.

10. Ground water can become unusable if it becomes ___________________and is no longer safe to drink.

Subjective evaluation of the text and Self-reported accuracy (Judgements of learning) items

Subjective evaluation of the text

1. How difficult was it to read the text? (1 – Very easy; 5 – Very difficult).

2. How much did you like the font the text were written in? (1 – Did not like at all; 5 – Liked very much).

3. How much did you like the text? (1 – Did not like at all; 5 – Liked very much).

Self-reported accuracy (Judgements of learning)

4. How many questions did you manage to answer correctly? (1 – None; 5 – All of them).

5. How well did you memorize the text? (1 – Memorized nothing; 5 – Memorized everything).

Figure S1. Raincloud plots for performance on three tasks in two conditions in long and short delay groups

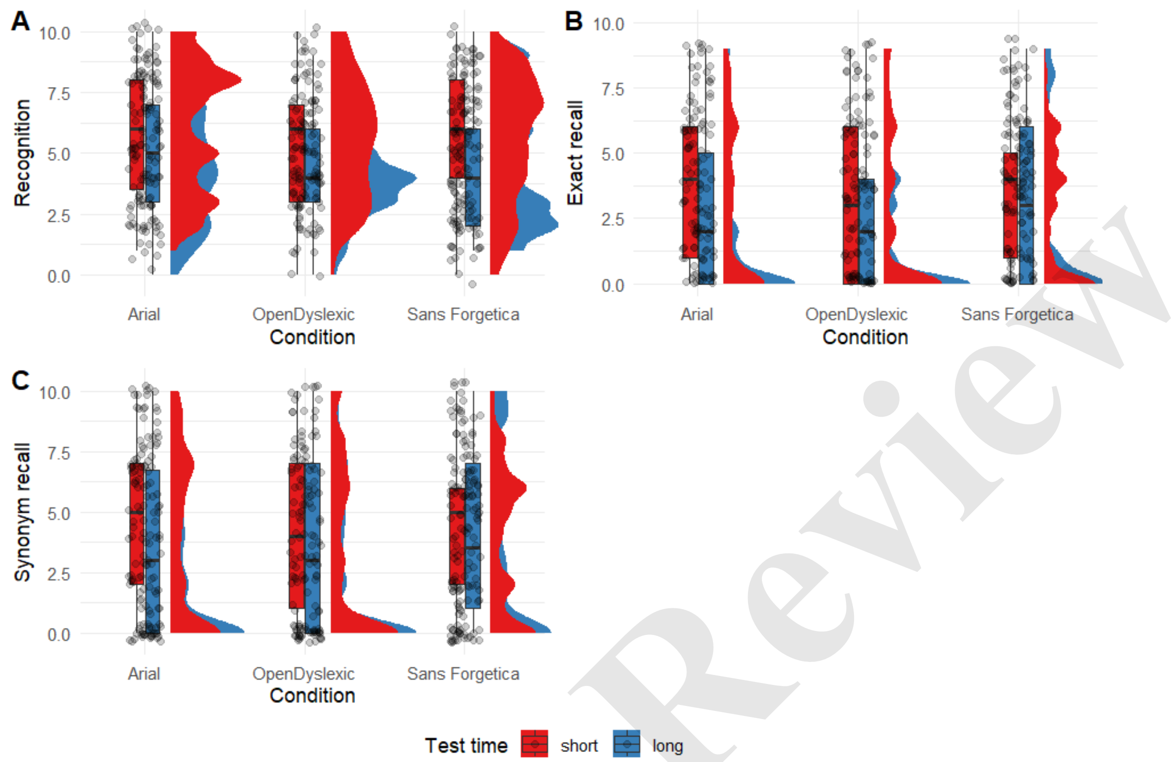

Table S1. Repeated Measures Bayesian ANOVA model comparison (BF01)

Models	P(M)	P(M|data)	BF_M	BF_{01}	error %
Null model (incl. subject and random slopes)	0.053	7.044×10^{-66}	1.268×10^{-64}	1.000	
QT + TT	0.053	0.432	13.707	1.629×10^{-65}	1.877

QT + TT + QT ✻ TT	0.053	0.374	10.74	1.885×10^{-65}	2.153
QT	0.053	0.09	1.785	7.809×10^{-65}	1.831
QT + D + TT	0.053	0.044	0.833	1.593×10^{-64}	7.296
QT + D + TT + QT ✻ TT	0.053	0.035	0.655	2.007×10^{-64}	11.668
QT + D	0.053	0.009	0.165	7.750×10^{-64}	7.432
QT + D + TT + D ✻ TT	0.053	0.007	0.131	9.761×10^{-64}	10.787
QT + D + TT + QT ✻ TT + D ✻ TT	0.053	0.007	0.126	1.016×10^{-63}	15.92
QT + D + TT + QT ✻ D	0.053	6.664×10^{-4}	0.012	1.057×10^{-62}	20.288

Note. All models include subject, and random slopes for all repeated measures factors. The table shows the best 10 out of 19 models. QT - Question type; TT - Test time; D - Dysfluency

Table S2. Repeated Measures Bayesian ANOVA model comparison (BF10)

Models	P(M)	P(M|data)	BF_M	BF_{01}	error %
Null model (incl. subject and random slopes)	0.053	6.781×10^{-66}	1.221×10^{-64}	1.000	
QT + TT	0.053	0.422	13.133	$6.221 \times 10^{+64}$	1.889

QT + TT + QT ✻ TT	0.053	0.384	11.214	$5.661\times10^{+64}$	4.044
QT	0.053	0.087	1.716	$1.284\times10^{+64}$	0.974
QT + D + TT	0.053	0.049	0.92	$7.169\times10^{+63}$	8.354
QT + D + TT + QT ✻ TT	0.053	0.035	0.659	$5.208\times10^{+63}$	10.33
QT + D	0.053	0.009	0.157	$1.272\times10^{+63}$	6.839
QT + D + TT + D ✻ TT	0.053	0.008	0.138	$1.121\times10^{+63}$	11.02
QT + D + TT + QT ✻ TT + D ✻ TT	0.053	0.006	0.104	$8.494\times10^{+62}$	20.625
QT + D + TT + QT ✻ D + QT ✻ TT	0.053	5.934×10^{-4}	0.011	$8.752\times10^{+61}$	32.731

Note. All models include subject, and random slopes for all repeated measures factors. The table shows the best 10 out of 19 models. QT - Question type; TT - Test time; D - Dysfluency

Table S3. Bayesian ANCOVA for Recognition model comparison

Models	P(M)	P(M|data)	BF_M	BF_{01}	error %
Null model	0.031	9.489×10^{-5}	0.003	1.000	
L2 proficiency	0.031	0.461	26.465	2.060×10^{-4}	2.251×10^{-4}
Gender + L2 proficiency	0.031	0.148	5.398	6.398×10^{-4}	0.002
L2 proficiency + Verbal ability	0.031	0.116	4.057	8.199×10^{-4}	0.002
L2 proficiency + L2 anxiety	0.031	0.088	2.976	0.001	0.002

Gender + L2 proficiency + Verbal ability	0.031	0.048	1.568	0.002	0.011
Gender + L2 proficiency + L2 anxiety	0.031	0.041	1.332	0.002	0.01
L2 proficiency + L2 anxiety + Verbal ability	0.031	0.028	0.9	0.003	0.01
L2 proficiency + D	0.031	0.026	0.828	0.004	1.433
Gender + L2 proficiency + L2 anxiety + Verbal ability	0.031	0.016	0.512	0.006	2.633×10^{-4}

Note. Showing the best 10 out of 32 models.

Table S4. Bayesian ANCOVA for Exact recall model comparison

Models	P(M)	P(M|data)	BF_M	BF_{01}	error %
Null model	0.031	1.639×10^{-4}	0.005	1.000	
L2 proficiency	0.031	0.437	24.035	3.754×10^{-4}	3.781×10^{-4}
Gender + L2 proficiency	0.031	0.21	8.236	7.810×10^{-4}	0.002
L2 proficiency + Verbal ability	0.031	0.085	2.897	0.002	0.001
L2 proficiency + L2 anxiety	0.031	0.079	2.658	0.002	0.001
Gender + L2 proficiency + L2 ANXIETY	0.031	0.055	1.792	0.003	0.01
Gender + L2 proficiency + Verbal ability	0.031	0.052	1.708	0.003	0.01
D + L2 proficiency	0.031	0.021	0.656	0.008	1.453

L2 proficiency + L2 anxiety + Verbal ability	0.031	0.02	0.619	0.008	0.01
Gender + L2 proficiency + L2 anxiety + Verbal ability	0.031	0.016	0.51	0.01	3.375×10^{-4}

Note. Showing the best 10 out of 32 models.

Table S5. Bayesian ANCOVA for Synonym recall model comparison

Models	P(M)	P(M|data)	BF_M	BF_{01}	error %
Null model	0.031	1.480×10^{-6}	4.587×10^{-5}	1.000	
L2 proficiency	0.031	0.414	21.909	3.573×10^{-6}	5.783×10^{-4}
Gender + L2 proficiency	0.031	0.235	9.514	6.301×10^{-6}	0.003
L2 proficiency + Verbal ability	0.031	0.087	2.961	1.697×10^{-5}	0.003
L2 proficiency + L2 anxiety	0.031	0.07	2.34	2.108×10^{-5}	0.003
Gender + L2 proficiency + Verbal ability	0.031	0.064	2.13	2.302×10^{-5}	0.006
Gender + L2 proficiency + L2 anxiety	0.031	0.046	1.501	3.203×10^{-5}	0.007
Condition + L2 proficiency	0.031	0.022	0.687	6.827×10^{-5}	1.235
L2 proficiency + L2 anxiety + Verbal ability	0.031	0.018	0.577	8.094×10^{-5}	0.008
Gender + L2 proficiency + L2 anxiety + Verbal ability	0.031	0.015	0.464	1.003×10^{-4}	0.006

Note. Showing the best 10 out of 32 models.

Dear editor,

We would like to express our sincere gratitude to you and the reviewers for thoughtful evaluation of our manuscript titled "Why dysfluent Font does not aid Second Language Learning". These insightful comments and suggestions have been immensely valuable in improving the quality and clarity of our work. We have carefully considered each comment and have made the necessary revisions, which are outlined below. Reviewers' comments are in *italics*, and the changes made in the manuscript are highlighted in yellow in the revised manuscript.

Reviewer 1:

Overall, this manuscript is very well written. However, on page 14, the researchers changed the spelling of the word "Dysfluency" to "Disfluency." There needs to be consistency.

We are grateful to the reviewer for bringing this to our attention. The spelling of the word "dysfluency" has been corrected on page 14 and throughout the text.

Reviewer 2:

No estimates of internal consistency are reported for the reading tests (i.e., Recognition, Exact recall, and Synonym recall). Please calculate and report these. If these estimates are large, they may be obscuring the effects of interest, so be sure to check for this in the revised MS.

We are grateful to the reviewer for this observation. The internal consistency of the reading tests was evaluated using Cronbach's alpha and McDonald's omega. These estimates are now presented in the corresponding subsections of the Methods section on pages 10–11.

"Internal consistency of the recognition test was moderate (Cronbach's α = .67; McDonald's ω = .73)."

"Internal consistency of the exact recall test was high (Cronbach's α = .77; McDonald's ω = .84)."

"Internal consistency of the synonym recall test was high (Cronbach's α = .81; McDonald's ω = .84)."

Given high resulting estimates, a sentence was also incorporated into the Limitations and future directions section on p.18:

Thirdly, the Recognition, Exact Recall and Synonym Recall tests had high internal consistency. Some studies have shown that higher internal consistency for a test in a particular group can lead to higher scores in that group (Ishikawa, 2023). However, our data showed that the tasks were sufficiently complicated, as indicated by absence of ceiling effect (Table 2).

The authors might consider, in subsequent studies, employing a measure of anxiety that is specific to L2 reading (L2 reading anxiety).

We are grateful to the reviewer for this recommendation. In the present study, we decided to utilize a more comprehensive foreign language anxiety questionnaire to evaluate the participants' general concerns regarding L2 learning. Nevertheless, a more specific questionnaire on L2 reading may potentially demonstrate more intricate correlations with dysfluency in the text. A sentence on this issue has been added to the Limitations and Future Directions section on p. 18:

"Fourthly, we tested general L2 anxiety as a moderator of the dysfluency effect with null result. More specific measures, such as L2 reading anxiety scale (e.g. FLRAS; Saito et al., 1999), may demonstrate more complex relations to dysfluency in the text."

Please insert the tables into the main body of the text. It's disruptive to have to go back and forth to the end of the MS to see and interpret the tables while reading.

After careful consideration of the reviewer's comment and a thorough consultation with the journal submission guidelines, we have decided to retain the tables in their current positions. The most pertinent tables will be included in the main text of the final publication (now marked with e.g., [Table 1 near here]). We have also given careful consideration to the option of moving some tables from the supplementary materials to the main body of the manuscript. However, we believe that the specific information they contain may not be of interest to the majority of readers.

Thank you once again for the opportunity to revise our manuscript, we believe that the revisions adequately address the concerns raised by the reviewers, and we hope that it can now be published in The Journal of Experimental Education.

Sincerely,

Authors