Metacognition and Learning
https://doi.org/10.1007/s11409-021-09284-6

Perceptually fluent features of study words do not inflate judgements of learning: evidence from font size, highlights, and Sans Forgetica font type

Nicholas P. Maxwell[1] · **Trevor Perry**[1] · **Mark J. Huff**[1]

Received: 15 April 2021 / Accepted: 20 October 2021

Abstract
Judgments of learning (JOL) are often used to assess memory monitoring at encoding. Participants study a cue-target word pair (e.g., mouse-cheese) and are asked to rate the probability of correctly recalling the target in the presence of the cue at test (e.g., mouse -?). Prior research has shown that JOL accuracy is sensitive to perceptual cues. These cues can produce metamemory illusions in which JOLs overestimate memory, such as the *font-size effect* (Rhodes & Castel, 2008), which occurs when participants inflate JOLs for pairs presented in large versus small fonts without a concomitant increase to recall. The present study further tests the font-size effect and examines whether other perceptual manipulations can affect the correspondence between JOLs and recall. Experiments 1A and 1B were designed to replicate the font-size effect and test whether the effect extended to highlighted pairs that were related or unrelated in the same study list. Experiment 2A and 2B examined font size and highlighting effects on JOLs using only unrelated pairs. Finally, Experiment 3 tested whether Sans Forgetica—a perceptually disfluent font designed to improve memory—would result in inflated JOLs and/or recall. Large fonts similarly increased both JOLs and recall relative to small fonts, highlights had no effect on JOLs or recall, and Sans Forgetica font yielded a memory cost (though no effect on JOLs). Collectively, perceptually fluent and disfluent study pairs do not appear to inflate JOLs relative to subsequent recall.

Keywords Judgments of learning · Font-size effect · Perceptual fluency · Sans Forgetica · Cued-recall

The ability for individuals to accurately monitor their learning progress is important for successfully encoding new information. Successful monitoring allows individuals to maximize retention by adjusting their study strategies and can inform what strategies are used in future study tasks (Nelson & Narens, 1990). Metamemory judgments (i.e., having individuals judge aspects of their memorial abilities) are commonly used by researchers to obtain information about the learning process. While researchers use several types of

✉ Mark J. Huff
mark.huff@usm.edu

[1] School of Psychology, The University of Southern Mississippi, 118 College Dr. #5025, Hattiesburg, MS 39406, USA

judgments to assess metacognitive processes, judgments of learning (JOLs) are commonly used. When making JOLs, participants typically study sets of cue-target word pairs (e.g., mouse-cheese) and are asked to estimate the likelihood of correctly retrieving a target word in the presence of a cue (e.g., mouse -?). While JOLs can be made using several measurement scales (e.g., Likert or binary "yes-no" responses; Hanczakowski et al., 2013), they are commonly elicited using a continuous 0 to 100 scale representing the percent likelihood of the target item being successfully recalled at test (e.g., 100% = definitely would remember; 0% = definitely would not remember). The use of a 100-point scale is beneficial because it allows for a direct comparison between predicted recall (via JOLs) and the proportion of items that are correctly recalled at test.

Although JOLs are often predictive of future recall (e.g., Nelson & Dunlosky, 1991), certain situations can produce metamemory illusions in which JOLs underpredict or overpredict subsequent memory. For example, relatedness cues such as the associative direction between cue-target pairs have repeatedly been shown to induce an *illusion of competence* in which JOLs overpredict later recall for certain types of paired associates (Koriat & Bjork, 2005). Specifically, forward associates, in which the cue is highly predictive of the target (e.g., lamp-shade), tend to produce JOLs that are well-calibrated with later recall. However, backward associates, in which the cue does not readily converge upon the target (e.g., shade-lamp), display a marked overconfidence effect such that JOLs greatly overestimate subsequent memory. Castel et al. (2007) have reported an illusion of competence pattern on identical pairs and, more recently, Maxwell and Huff (2021) have extended this pattern to symmetrical associates (e.g., king-queen), in which the forward and backward relations between pairs are matched. Like Koriat and Bjork, Maxwell and Huff found that JOL ratings were generally well-calibrated for forward pairs, but produced an illusion of competence pattern for backward, symmetrical, and unrelated word pairs. Additionally, the illusion of competence was robust and persisted across a variety of experimental manipulations designed to improve the correspondence between JOLs and recall, such as JOL timing (e.g., concurrent, immediate, or delayed JOLs) and pacing (e.g., self-paced vs. experimenter paced). Thus, although JOLs can accurately predict later recall, predictions are best when cues are related to targets in the forward direction.

In addition to relatedness cues, other factors have been shown to influence judgments. For example, perceptual cues have been shown to affect a variety of judgment tasks, including affective judgments (e.g., judging a target item's beauty, Reber et al., 1998), veridicality judgments (e.g., truthfulness of statements; Reber & Schwarz, 1999), and JOLs (Rhodes & Castel, 2008). Typically, studies investigating the effects of perceptual cues on judgment making do so by varying the ease with which participants can encode stimuli (see Schwarz, 2004, for a review). These ease-of-processing manipulations typically occur by changing some aspect of the stimuli (e.g., size, clarity, etc.) such that certain items are made more difficult to encode relative to others. For example, Reber et al. reported that participants judge perceptually fluent items as being more affectively pleasing versus disfluent items. Additionally, Reber and Schwarz (1999) showed that participants are more likely to judge perceptually fluent statements (e.g., black ink against a white background) as being true compared to perceptually disfluent statements (e.g., yellow ink against a white background).

Importantly, ease-of-processing-type effects have been shown to extend to JOLs. For example, Rhodes and Castel (2008) tested participants on word pairs that were studied in either large (48-pt.) or small (18-pt.) font sizes. A *font-size effect* was found in which JOLs were greater when pairs were presented in large versus small font. However, this increase in JOLs did not translate to recall as both font types were recalled at equivalent rates.

Subsequent experiments indicated that the font-size effect was largely driven by the additional ease-of-processing afforded by large-font pairs. For example, the font-size effect was largely diminished when ease of reading was manipulated such that words were presented by alternating between uppercase and lowercase letters (e.g., HoUsE) and, furthermore, the effect was moderated by pair relatedness, as the effect was reduced when participants studied related versus unrelated pairs (Rhodes & Castel).

The font-size effect has been reported across several studies. Kornell et al. (2011) replicated the font-size effect and showed that this pattern holds when pairs are studied repeatedly. More recently, Hu et al. (2015) divided participants into groups that either studied or observed the participants who had been assigned to the study group. Participants in the study group made JOLs for pairs presented in either large or small fonts, while participants in the observer group were asked to guess the JOLs that participants in the study group would make and were only made aware of the font size of the pair that was being viewed, not the pair itself. Participants in both groups provided higher JOLs for large- than small-font pairs. Finally, Price and Harrison (2017) examined whether the font-size effect influenced the magnitude of pre-study JOLs. Overall, they showed that participants tended to assign higher JOLs for items presented in a large than small font, regardless of whether the JOL was provided pre- or post-study.

Although the font-size effect has been reported under several conditions, the underlying factors driving the effect remain unclear. Two accounts have been proposed for the font-size effect—the fluency account and the beliefs account. The fluency account suggests that larger words are more perceptually fluent than smaller words. Due to enhanced fluency, participants process larger words more efficiently and/or effectively, leading to greater JOLs relative to smaller words. In a test of the fluency account, Undorf et al. (2017) presented participants with images of objects, faces, and words which were initially too small to perceive and incrementally increased the size of the stimuli. Participants were asked to make a JOL once they could recognize the stimulus, with the recognition latency recorded. Overall, JOLs were found to be inversely related to the recognition latency, indicating that items judged as more memorable were processed more quickly. Relatedly, Yang et al. (2018) tested the fluency account by comparing the results of a continuous identification (CID) task to the results of lexical decision tasks. The CID task tested the relationship between perceptual fluency and JOLs by alternating between a word and a corresponding mask (e.g., switching between the word "ball" and "####"). The speed in which alternations occurred was gradually decreased over time such that the word was made visible on the screen for longer durations (e.g., 20 ms in the first cycle, 40 ms in the second cycle, etc.). The goal of the CID was to slowly increase fluency by gradually making the word less obscure. Like Undorf et al., JOLs were greater for words that could be identified faster (i.e., those with a higher perceptual fluency).

Whereas the fluency account is based on the ease of processing items at study, the beliefs account posits that participants' extra-experimental expectations regarding an item's memorability contributes to JOLs. Regarding the font-size effect, participants may assign higher JOLs to large items because they hold the belief that large pairs are easier to learn than small pairs. To test the beliefs account, Mueller et al. (2014) had participants first complete a lexical decision task for a set of large versus small items. Unlike Yang et al. (2018), latencies on the lexical-decision task did not differ as a function of font size, suggesting that the perceptually fluent large font did not facilitate latencies as predicted by a fluency account. Importantly, however, reported beliefs about the memorability of large versus small fonts and pre-study JOLs indicated that participants did indeed hold the belief that large-font items will be better remembered.

While the fluency and beliefs accounts are often framed as competing explanations of the font-size effect, these two accounts are not mutually exclusive. Jemstedt et al. (2018) found that while perceptual fluency affected the magnitude of ease-of-learning judgments, the effect was largely moderated by beliefs. Specifically, participants studied words presented using either a constant or alternating case (e.g., BASKET vs. bAsKeT). While ease-of-learning judgments were higher when pairs were presented using the more fluent constant case, judgments were greatest for participants who self-reported that pairs presented in the constant condition were easier to learn than the alternating condition compared to participants who believed there was no difference. Thus, while fluency and beliefs have each been shown to separately influence the magnitude of metacognitive judgements, these judgments often reflect a combination of fluency and belief processes.

Finally, perceptual manipulations aside from font sizes have been shown to affect JOLs. Ball et al. (2014) tested how bolding word pairs affected JOLs and subsequent memory. Compared to non-bolded pairs, bolded pairs received higher JOLs; however, like the font-size effect, no differences in recall performance were detected between the two pair types. Additionally, Besken (2016) had participants complete a memory task in which images were presented either intact or with sections removed (i.e., fluent vs. disfluent) and had participants provide JOLs at encoding. Consistent with a fluency-based process, intact images received higher JOLs relative to incomplete images.

The present study provided a further test of the font-size effect while extending it to include other perceptual manipulations designed to affect JOLs. Specifically, Experiments 1A and 2A sought to replicate the font-size effect using related and unrelated pairs. Next, Experiments 1B and 2B tested whether highlighting word pairs (vs. not highlighting) would affect JOLs as font size. Finally, Experiment 3 tested whether JOLs and recall rates would be affected by Sans Forgetica font—a disfluent font that is more perceptually difficult to process relative to a standard font such as Arial. Sans Forgetica allowed for a potential evaluation of fluency effects on JOLs by testing whether a perceptually disfluent presentation can reduce JOLs—the opposite pattern to what is generally found for fluent presentations.

Finally, we expand upon previous work (e.g., Rhodes & Castel, 2008) by including a pure-control group comparison in which only the standard perceptual condition is used (i.e., all pairs presented in a standard font size), rather than a mix of perceptually fluent/disfluent pairs. These control groups were included because encoding manipulations have been shown to spill over into other encoding tasks when encoding is manipulated within-subjects (Bodner et al., 2014; Huff et al., 2021). Thus, our inclusion of these control groups allowed us to gauge perceptual effects on JOLs more accurately relative to a baseline condition.

Experiment 1A: Font-size effects on related and unrelated pairs

The goal of Experiment 1A was to replicate the font-size effect using a set of related and unrelated word pairs. Overall, we expected that because large-font pairs are more perceptually fluent or because participants possess a belief that large pairs are more memorable, JOLs would be greater for large than for small pairs (cf. Rhodes & Castel, 2008). We also included comparisons to a control group who viewed pairs presented

using a standard, 32-pt. Arial font to assess whether any effects of perceptually fluent pairs hold when compared to a pure list of standard pairs. Font-size effects were tested using a mixed list of forward, backward, and symmetrical paired associates and unrelated word pairs.

Method

Participants

Eighty participants were recruited from Prolific (www.prolific.co), an online academic crowdsourcing platform, and completed the study at rate of $4.00 per half hour. Participants were required to be native English speakers and have completed at least a high school education or equivalent. Participants were randomly assigned to the font-size group ($n = 41$) which studied large and small font pairs or the control group that studied pairs in a standard font size ($n = 39$). We modeled our sample size for each group after Rhodes and Castel (2008), who found evidence for the font-size effect using a sample of 20 participants and Maxwell and Huff (2021), who found reliable illusion of competence patterns for backward, symmetrical, and unrelated pairs using a sample of 30 participants. Due to an anticipated increase in performance variability from an online sample, we recruited more participants than these previous studies to improve reliability of the data set. Cued-recall performance was used as a compliance check (participants would be omitted for correct recall rates <5%, which suggested that instructions were not properly followed), however all exceeded this threshold and therefore no participants were omitted. Across groups, participants reported a mean age of 29.43 ($SD = 14.19$), and all participants reported normal or corrected-to-normal vision. Full demographics for each encoding group are reported in the Appendix (Table 1).

Materials

One-hundred-eighty word pairs taken from Maxwell and Huff (2021) served as study materials. These pairs included 40 forward pairs (e.g., bounce-ball), 40 backward pairs (e.g., ball-bounce), 40 symmetrical pairs (e.g., off-on), 40 unrelated pairs (e.g., pencil-fence), and 20 weakly related buffer pairs that were not tested to control for primacy and recency effects. The University of South Florida Free Association Norms (Nelson et al., 2004) were used to equate the related pairs in associative strength and to ensure that symmetrical pairs were equivalent in associative strength in the forward and backward direction. Pair types were also equated on lexical and semantic properties including word length, SUBTLEX frequency (Brysbaert & New, 2009), and concreteness, as reported in the English Lexicon Project (Balota et al., 2007). All pairs were evenly distributed into two study lists which contained 20 forward, backward, symmetrical, and unrelated pairs, and 10 buffer pairs. Both lists were matched on the above lexical and semantic properties. Study materials for all experiments have been made available at https://osf.io/3xwdr/. Associative strength, lexical, and semantic properties are listed in the Appendix (Tables 2 and 3).

All participants studied both lists which were evenly divided into two study/test blocks, the order of which was counterbalanced across participants. Each list was organized such

that five buffer pairs were presented at the beginning and end of each study list with the remaining pairs presented in a newly randomized order for each participant. Counterbalanced versions were produced from each study list that reversed the order of the pair lists (i.e., A-B pairs become B-A pairs), which allowed for greater control of item differences across pair types.

Participants in the font-size group saw lists in which half of the pairs were presented in a small 12-pt. font and the other half of pairs were presented in a large 54-pt. font, which was counterbalanced across pair types. All pairs were presented in Arial font style. In the control group, pairs were presented in 32-pt. Arial font.

For the cued-recall test, participants were presented with all 80 cue items from the initial study list (buffers were not tested). The cue was presented alongside a question mark (e.g., bounce-?), and participants were instructed to retrieve the target from memory. Test items were newly randomized for each participant. Test instructions did not mention font size. The cue word in all groups was presented in a standard 32-pt. Arial font.

Procedure

All participants were tested online via *Collector*, an open-source program for presenting web-based psychological experiments (Garcia & Kornell, 2015). Participants were informed that they would study a series of cue-target pairs in which the cue would be presented on the left, and the target on the right. They were further instructed that their memory for the target item would be tested following study, with only the cue word presented at test. In addition to studying the pairs, participants were instructed to provide a JOL by rating the likelihood they would be able to correctly recall the target if only presented with the cue. JOLs were provided using a scale ranging from 0 to 100. A rating of 0 indicated that the participant had no confidence in their ability to recall the word at test; a rating of 100 indicated complete certainty that they would recall the target. Participants were encouraged to utilize the full range of the scale and to avoid anchoring on extremes and mid points when providing ratings (i.e., 0, 50, or 100 ratings). Following instructions, participants then studied the first block of word pairs and provided JOLs concurrently with study such that ratings were provided while the word pair was displayed on the screen. Participants advanced to the next study pair after entering in their JOL into a dialog box and clicking a labeled "next" button on the screen. Upon completion of the first study list, participants completed a filler task where they had to list the 50 U.S. states in alphabetical order for 2 min, which was immediately followed by the cued-recall test. Participants were presented with the cue word paired with a question mark (e.g., credit -?) and were asked to retrieve the correct target word by typing it into a dialog box. This test was untimed, and participants were instructed to press the enter key to advance to the next test item. If participants were unable to retrieve the target, they were instructed to advance to the next item without providing a guess. Participants entered their memory responses and/or advanced to the next item by clicking a labeled "next" button. Following the first cued-recall test, participants completed a second study list, filler task, and second cued-recall test which only tested pairs from the second study list. Following the second cued-recall test, participants were debriefed and provided with compensation. The duration of the experiment was less than 30 min across groups.

Results

All JOL responses were initially screened for missing responses and outliers (i.e., JOLs outside the 0–100% range). This screening process removed fewer than 0.5% of total responses. All missing recall responses were coded as incorrect. A liberal scoring criterion was used such that misspellings or pluralizations were scored as correct.

A $p < .05$ significance level was used for all analyses. Effect size estimates using partial-eta squared (η_p^2) and Cohen's *d* were computed for all significant analyses of variance (ANOVAs) and *t*-tests, respectively. To supplement standard null-hypothesis significance testing, we include a Bayesian estimate of the strength of evidence supporting the null hypothesis (Masson, 2011; Wagenmakers, 2007). This analysis compares a model that assumes a significant effect to one that assumes a null effect. A probability estimate is computed termed p_{BIC} (Bayesian Information Criterion) which indicates the likelihood that the null hypothesis is retained. Thus, null effects are supplemented with a p_{BIC} estimate. Figure 1 plots mean JOL and cue-recall percentages for the large-font, small-font, and control groups as a function of forward, backward, symmetrical, and unrelated pair types. For completeness, cell means are reported in Appendix Table 4. Finally, though our analyses focus on calibration, for completeness we report Goodman-Kruskal gammas correlations (Appendix Table 5) as a metric of resolution, and analyses of gammas across experiments are available in our Supplemental Materials (https://osf.io/xymez/).

In our analyses, we first compare JOL and recall percentages across pair types in the font-size group and then compare between the within large- and small-font pairs and the control group. We then test for an illusion of competence pattern (i.e., JOLs overestimating recall, often as a function of pair direction; Koriat & Bjork, 2005; Maxwell & Huff, 2021), as previous research has only assessed the font-size effect using forward associates and unrelated pairs.

Starting with the font-size group, a 2(Measure: JOL vs. Recall) × 2(Font Size: Large vs. Small) × 4(Pair Type: Forward vs. Backward vs. Symmetrical vs. Unrelated) within-subject ANOVA yielded an effect of measure, $F(1, 40) = 10.11$, $MSE = 1258.14$, $\eta_p^2 = .20$, in which JOLs exceeded recall percentages (50.34 vs. 41.53, for JOLs and recall rates, respectively). An effect of pair type was also found, $F(3, 120) = 414.56$, $MSE = 218.49$, $\eta_p^2 = .91$, in which JOL/recall percentages were greatest for forward pairs (64.28), followed by symmetrical pairs (61.25), backward pairs (45.23), and unrelated pairs (12.99), with all pairs differing from each other, $ts > 3.27$, $ds > 0.24$. An effect of font size was also found, $F(1, 40) = 12.20$, $MSE = 66.26$, $\eta_p^2 = .23$, indicating that JOL/recall percentages overall were greater for large than small font pairs (47.05 vs. 44.83). Importantly, all interactions with font size, including the three-way interaction, were not reliable, $Fs < 1.63$, $ps > .18$, $p_{BIC}s > .99$, indicating that the large font did not differentially inflate JOLs relative to recall across pair types (cf. Rhodes & Castel, 2008).

Regarding the illusion of competence, a measure × pair type interaction was found, $F(3, 120) = 45.27$, $MSE = 146.88$, $\eta_p^2 = .53$, which indicated that JOLs overestimated recall for some pair types. Consistent with previous research (e.g., Koriat & Bjork, 2005; Maxwell & Huff, 2021), JOLs were well-calibrated to recall on forward pairs (61.84 vs. 66.72), $t(40) = 1.41$, $SEM = 3.46$, $p = .17$, $p_{BIC} = .70$. However, for the more deceptive backward pairs, JOLs greatly exceeded later recall (58.20 vs. 32.26), $t(40) = 6.95$, $SEM = 3.73$, $d = 1.48$, a pattern which extended to symmetrical pairs (64.75 vs. 57.75), $t(40) = 2.28$, $SEM = 3.08$, $d = 0.40$, and unrelated pairs (16.56 vs. 9.42), $t(40) = 2.91$, $SEM = 2.46$, $d = 0.64$.

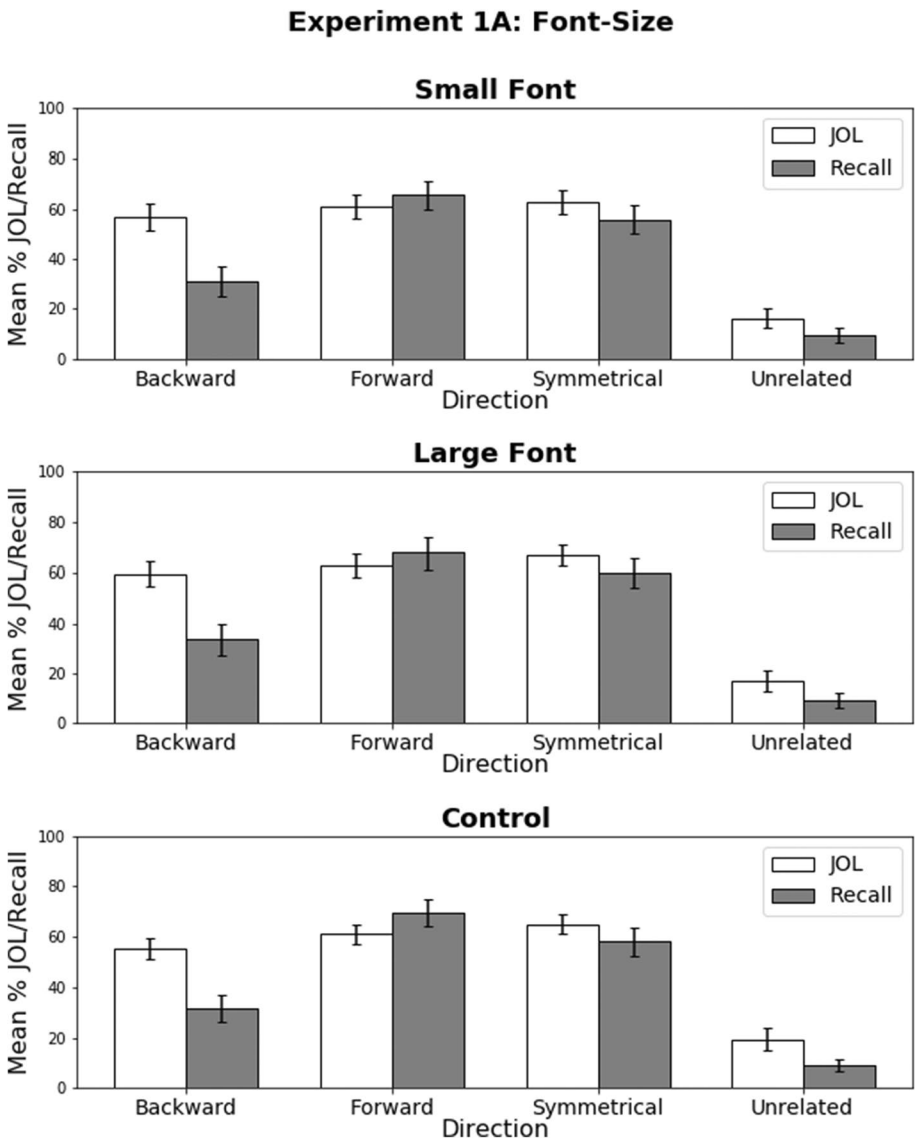

Fig. 1 Mean JOL and recall rates as a function of pair direction for pairs presented in small font (top panel), large font (middle panel), and the control group (bottom panel) in Experiment 1A. Bars represent 95% confidence intervals

We then compared changes in the magnitude of JOLs/recall for large and small font pairs relative to the control group to evaluate font size effects compared to a pure group that encoded all pairs in a single font size. The control group similarly showed robust pair type differences on JOLs/recall percentages, $F(3, 114) = 421.14$, $MSE = 100.03$, $\eta_p^2 = .92$, and the same illusion of competence pattern found for both large- and small-font pairs, $F(3, 114) = 68.49$, $MSE = 49.12$, $\eta_p^2 = .64$; however, no main effects or interactions were found

when comparing large and small font size pairs relative to the control group, all Fs < 1.47, ps > .22, p_{BIC}s > .99. Collectively, increasing font size increased both JOLs and recall percentages equally relative to small-font sizes. However, JOLs and recall for large-font sizes did not differ relative to the control group.

Experiment 1B: Highlighting effects on related and unrelated pairs

Experiment 1B was a replication of Experiment 1A but used a highlight perceptual manipulation in which half of the pairs were presented using a yellow-highlight format and the other half were presented in a standard, non-highlight format. All pairs were presented using the same font size with the only perceptual difference being the difference in highlight presentation. We selected this manipulation, as under some conditions, the use of highlighting can be beneficial to comprehension and learning, as highlighting makes text distinguishable from non-highlighted material (Fowler & Barker, 1974; Yue et al., 2015). By making text more perceptually distinguishable (and thus possibly perceptually fluent), we expected that highlighting would operate similarly to other manipulations that enhance both distinctiveness and fluency (e.g., bolded vs. unbolded pairs; Ball et al., 2014, font-color; Wehr & Wippich, 2004, etc.; but see Price, McElory, & Martin, 2016, who showed that bolded items received lower JOLs and were recalled at lower rates than unbolded pairs).

Like Experiment 1A, we expected that highlighting pairs would increase perceptual fluency and thus increase the likelihood that participants would provide elevated JOL ratings relative to non-highlighted pairs, a pattern consistent with large font-size effects reported by Rhodes and Castel (2008). However, given that large-font pairs were only found to produce a small and equivalent increase to both JOLs and recall relative to small-font pairs in Experiment 1A, it is possible that highlighting pairs would also increase both JOL and recall percentages equally. We also included comparisons to the control group used in Experiment 1A (non-highlighted pairs of the same font size) to gauge whether any highlighting benefits would hold when compared to a pure list of non-highlighted pairs. Again, highlighting effects were compared across forward, backward, symmetrical, and unrelated pair types.

Method

Participants

An additional 41 participants were recruited from Prolific to complete the study using the same recruitment criteria as Experiment 1A and were compensated at rate of $4.00 per half hour. Participants reported a mean age of 32.24 (SD = 16.74), and all were native English speakers who reported normal or corrected-to-normal vision. Full demographic information is available in Appendix Table 1.

Materials and procedure

The same materials and general procedure in Experiment 1A were again used in Experiment 1B, with the only difference being the highlight versus no highlight presentation of

word pairs. All pairs were presented in a 32-pt. Arial font type, and half of the pairs were presented in a bright yellow highlighted format, whereas the other half were presented in a standard non-highlighted format. The cued recall test was identical to Experiment 1A, and all test pairs were presented in a randomized order using a non-highlighted format. The control group from Experiment 1A was also used.

Results

Experiment 1B followed the same data screening procedure as Experiment 1A; less than 0.5% of the total JOL trials were removed. Figure 2 plots mean JOL and cued-recall percentages for highlight and no-highlight pairs across the four pair types. As in Experiment 1A, we first compare JOL/recall percentages across highlight and no-highlight pair types and then compare the within-subject highlight pairs relative to the control group. First, a 2(Measure) × 2(Highlight) × 4(Pair Type) within-subject ANOVA yielded an effect of measure, $F(1, 40) = 7.69$, $MSE = 1346.04$, $\eta_p^2 = .16$, in which overall, JOLs exceeded recall rates (50.65 vs. 42.70). An effect of pair type, $F(3, 120) = 410.75$, $MSE = 197.25$, $\eta_p^2 = .91$, indicated that JOL/recall percentages were greatest for forward pairs (64.87), followed by symmetrical pairs (60.88), backward pairs (45.06), and unrelated pairs (15.90). All pair types differed from each other, ts > 3.10, ds > 0.34. Unlike Experiment 1A, however, the fluent highlighting factor did not result in a main effect, $F < 1$, $p_{BIC} = .83$, nor were any interactions with this factor reliable including the three-way interaction, all Fs < 1, ps > .72, p_{BIC}s > .99.

The measure × pair type interaction was again significant, $F(3, 120) = 56.96$, $MSE = 114.88$, $\eta_p^2 = .59$, indicating an illusion of competence pattern across highlight pairs. For forward pairs, JOLs were lower than recall (61.64 vs. 68.10), $t(40) = 2.17$, $SEM = 2.94$, $d = 0.46$, however, illusion of competence patterns were found in which JOLs exceeded recall rates for symmetrical pairs (64.85 vs. 56.90), $t(40) = 2.51$, $SEM = 3.16$, $d = 0.49$, and backward pairs (57.21 vs. 32.91), $t(40) = 6.89$, $SEM = 3.53$, $d = 1.55$, but were only marginally greater than recall on unrelated pairs (18.91 vs. 12.90), $t(40) = 1.90$, $SEM = 3.16$, $p = .06$, $p_{BIC} = .52$.

We then used a set of mixed ANOVAs to compare JOLs/recall percentages on the within-subject highlight and no-highlight pairs relative to control group pairs. Consistent with Experiment 1A, no effects or interactions were found when comparing the control-group pairs to either of the highlight pairs, all Fs < 1.56, ps > .19, p_{BIC}s > .99. Collectively, highlighting pairs had no effect on JOLs or recall rates when compared to either no-highlight pairs in a mixed list or when compared to the pure list control of non-highlighted pairs.

Discussion

First, Experiment 1A did not show evidence consistent with font-size effect as originally reported by Rhodes and Castel (2008) as the expected interaction was not observed. Instead, relative to the small font, the large font increased both JOLs and correct recall. Furthermore, while large font did increase JOLs, the magnitude of this effect was smaller than reported by Rhodes and Castel. Finally, no differences were detected between either large or small font pairs and the control group. Thus, although the large

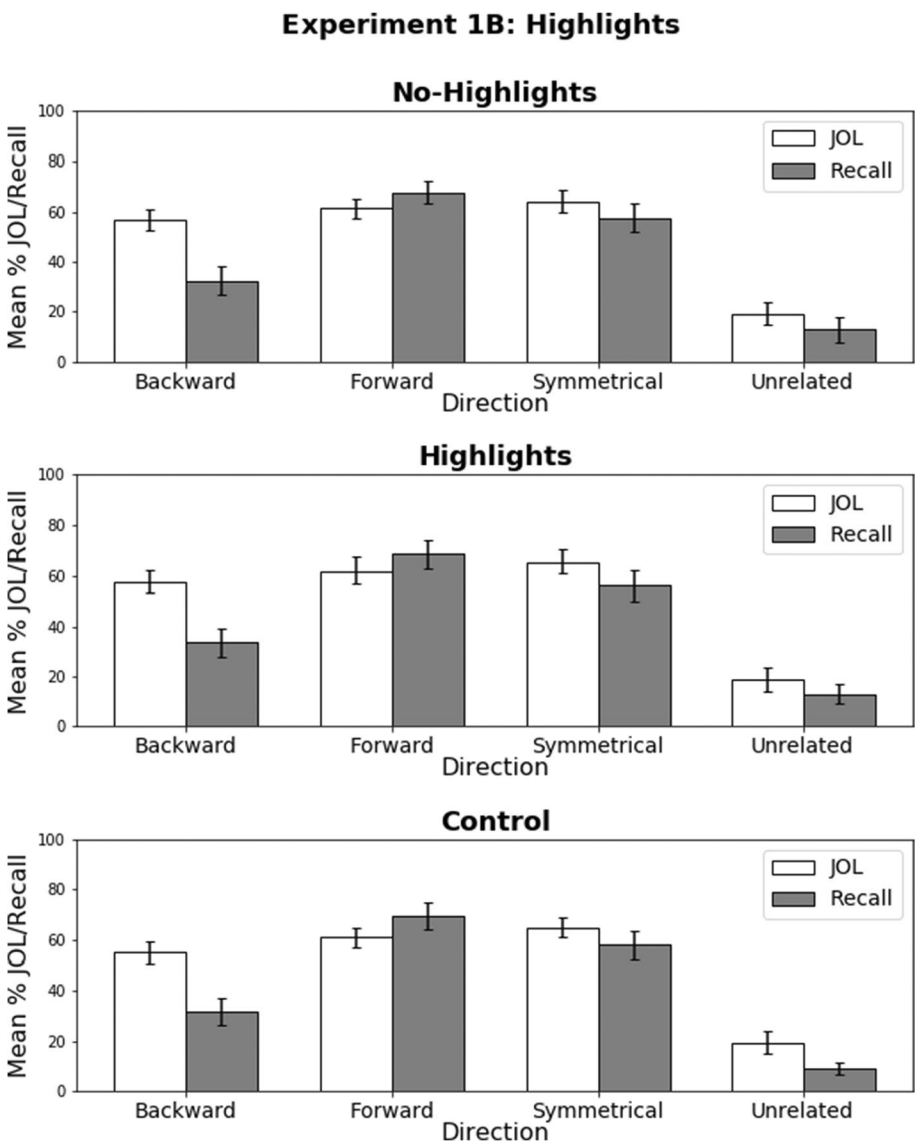

Fig. 2 Mean JOL and recall rates as a function of pair direction for non-highlighted pairs presented in mixed lists (top panel), highlighted pairs presented in mixed lists (middle panel), and non-highlighted pairs presented in the control group (bottom panel) in Experiment 1B. Bars represent 95% confidence intervals

font resulted in an increase to JOLs relative to the small font, it did not selectively increase JOLs without also affecting recall as previously reported. Second, to test the effects of other types of perceptual manipulations on JOLs and recall, we introduced a highlighting manipulation in Experiment 1B. However, the presence of highlighting did not affect JOLs or recall relative to non-highlighted pairs or the control group.

The discrepancies regarding the font-size effect in Experiment 1A may have resulted from our inclusion of both related and unrelated stimuli pairs within a mixed list. While Rhodes and Castel (2008) found a font-size effect on both related and unrelated pairs (Experiment 3), they noted that the effect was stronger when participants studied only unrelated pairs. Thus, our inclusion of related pairs may have negated potential fluency effects on JOLs. To test this possibility, Experiments 2A and 2B followed the same methods as Experiments 1A and 1B but included only unrelated pairs.

Experiment 2A: Font-size effects on pure unrelated lists

Because Rhodes and Castel (2008) found that the font-size effect was dampened when participants studied related pairs, Experiment 2A sought to replicate the font-size effect using only unrelated pairs. Overall, our predictions followed Experiment 1A. Specifically, we anticipated that pairs presented using a large font would have inflated JOLs relative to small font pairs. We again expected that there would be no differences in recall as a function of font-size, with only unrelated pairs. Finally, as in Experiment 1, we included comparisons to a control group who studied a pure list of pairs presented in a standard 32-pt. Arial font.

Method

Participants

Sixty-five participants were recruited from Prolific and completed the study at a rate of $4.00 per half hour. Prolific recruitment followed the same guidelines used in Experiment 1 such that participants were required to be native English speakers and have obtained at least a high school education or equivalent. An additional 12 undergraduates were recruited from The University of Southern Mississippi's psychology research pool and completed the study in exchange for course credit. Participants were randomly assigned to either the font-size or control group. Our sample size was based on Experiment 1. Data from 9 participants were excluded due to low recall rates (e.g., correct recall rates <5%), which suggested that experiment instructions were not properly followed. This resulted in 36 participants in the font-size group and 32 participants in the control group. Across groups, participants reported a mean age of 24.63 ($SD = 10.19$). All participants were native English speakers who reported normal or corrected-to-normal vision. Full demographics for Experiment 2 are reported in Appendix Table 1.

Materials and procedure

Experiment 2A followed the same procedure used in Experiment 1A with the exception that participants studied only unrelated pairs rather than a mixed list of related and unrelated pairs. To ensure a sizeable list of unrelated pairs, unrelated pairs from Experiment 1A were combined with a new set of unrelated pairs, leading to a total of 160 unrelated study pairs (80 pairs per block; see Appendix Table 6 for lexical characteristics). All other materials, including buffer pairs and the procedure, were identical to Experiment 1A.

Results

Figure 3 plots mean JOL and cued-recall percentages for large- and small-font pairs in the mixed group and pairs in the control group. For completeness, cell means are reported in Appendix Table 7 (see Appendix Table 5 for gammas). We first compared font size differences in the mixed group using a 2(Measure)×2(Font Size) within-subject ANOVA. Across font sizes, JOLs were not greater than recall rates (27.99 vs. 23.30), $F(1, 35) = 2.15$, $MSE = 369.08$, $p = .15$, $p_{BIC} = .67$, but collapsed across measure, mean JOLs/recall rates were greater for large than small fonts respectively (27.00 vs. 24.30), $F(1, 35) = 19.10$, $MSE = 13.76$, $\eta_p^2 = .35$. Importantly, however, font size did not affect JOLs and recall rates differently, $F < 1$, $p_{BIC} = .85$.

Next, using a pair of mixed ANOVAs, we compared JOLs and recall of large font and small font to the control group. Overall, relative to control pairs, JOLs exceeded recall rates—an illusion of competence pattern—both when compared to large-font pairs (27.15 vs. 22.04), $F(1, 66) = 4.43$, $MSE = 202.28$, $\eta_p^2 = .06$, and when compared to small-font pairs (25.80 vs. 20.53), $F(1, 66) = 5.75$, $MSE = 164.00$, $\eta_p^2 = .08$. JOLs/recall rates were marginally greater for large-font pairs than control pairs (27.00 vs. 21.89), $F(1, 66) = 3.54$, $MSE = 249.20$, $p = .06$, $\eta_p^2 = .05$, $p_{BIC} = .58$, but no difference occurred between small-font pairs and control pairs (24.30 vs. 21.89), $F < 1$, $p_{BIC} = .85$. Like the large- and small-font pair comparison above, font size did not differentially affect JOLs from recall rates relative to control pairs, Fs < 1, p_{BIC}s $> .88$.

Experiment 2B: Highlighting effects on pure unrelated lists

Experiment 2B provided a replication of Experiment 1B using only unrelated item pairs. Our predictions followed Experiment 1B and were in line with Rhodes and Castel's (2008) font-size effect. We expected that highlighted pairs would be more perceptually fluent and would receive inflated JOLs relative to non-highlighted pairs. Recall was again not expected to differ as a function of highlighting. Consistent with the previous experiments, a control-group comparison was included. Thus, both JOLs and recall for highlighted and non-highlighted pairs were compared to the pure-control group from Experiment 2A.

Fig. 3 Mean JOL and recall rates as function of pair type in Experiment 2A. Bars represent 95% confidence intervals. All study pairs were unrelated

Method

Participants

An additional 40 participants from Prolific completed Experiment 2B. Data for three participants were omitted using the same exclusion criteria as Experiment 2A (recall <5%). Participants completed the study at a rate of $4.00 per half hour. Participants reported a mean age of 24.35 years ($SD = 10.15$). All were native English speakers who reported normal or corrected-to-normal vision. Full demographic information is presented in Appendix Table 1.

Materials and procedure

The same unrelated pairs from Experiment 2A were again used in Experiment 2B. All procedures were identical with the exception that instead of large/small font sizes, half of the pairs were presented in a highlighted modality as in Experiment 1B, and the other half were presented in a non-highlighted modality. The font size for pairs in the highlight group was also identical to Experiment 1B, which matched the font size of the pairs in the control group.

Results

Figure 4 plots mean JOL and cued-recall percentages for highlight and no-highlight pairs and control-group pairs. Highlight differences were first compared using a 2(Measure) × 2(Highlight) within-subject ANOVA. Consistent with an illusion of competence pattern, a main effect of Measure was detected, such that overall JOLs exceeded later recall rates (29.17 vs. 20.54), $F(1, 36) = 6.26$, $MSE = 440.59$, $\eta_p^2 = .15$; and like Experiment 1B, the highlight main effect was not reliable, $F(1, 36) = 2.82$, $MSE = 18.23$, $p = .10$, $p_{BIC} = .60$, nor was the interaction, $F < 1$, $p_{BIC} = .86$.

We then compared highlight and no-highlight pairs to the control group. A pair of mixed ANOVAs revealed that JOLs exceeded recall rates both in the highlight/control comparison (27.40 vs. 20.20), $F(1, 67) = 8.58$, $MSE = 201.54$, $\eta_p^2 = .11$, and in the no-highlight/

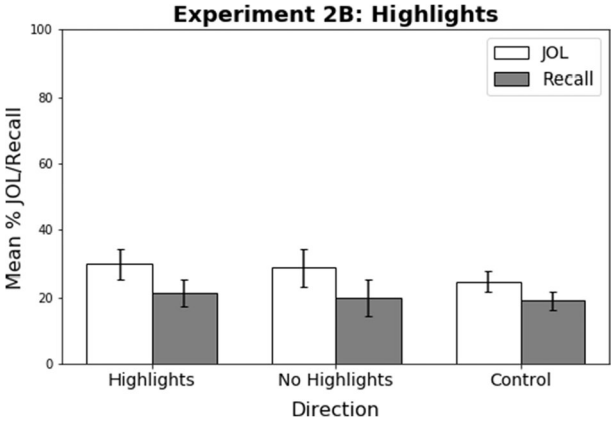

Fig. 4 Mean JOL and recall rates as function of pair type in Experiment 2B. Bars represent 95% confidence intervals. All study pairs were unrelated

control comparison (26.85 vs. 19.47), $F(1, 67) = 8.73$, $MSE = 80.57$, $\eta_p^2 = .12$. There were no differences when comparing either highlight or no-highlight pairs relative to the control group, Fs < 1, p_{BIC}s > .85. The interactions were also not reliable, Fs < 1, p_{BIC}s > .87.

Discussion

Findings from Experiments 2A and 2B are quite clear. First, in Experiment 2A, the font-size effect was not in evidence, even after removing related pairs and presenting participants with only unrelated pairs at study. While large font was again found to increase JOLs, it also produced an equivalent increase in recall, replicating the pattern observed in Experiment 1A. Finally, font-size did not affect JOLs or recall relative to the control group. Next, Experiment 2B replicated the pattern of results observed in Experiment 1B. Specifically, JOLs and recall did not differ between highlighted and non-highlighted pairs, and no differences were detected for either pair type relative to the control group. Thus, while highlighted pairs are likely to be perceptually fluent relative to non-highlighted pairs, neither memory predictions nor recall are affected.

Because font-size only produced a small effect on JOLs in Experiments 1A and 2A compared to previous studies (e.g., Mueller et al., 2014; Rhodes & Castel, 2008, etc.) and no highlighting effect was observed in Experiments 1B or 2B, we tested whether perceptually disfluent information would affect JOL estimations. Specifically, Experiment 3 tested whether Sans Forgetica, a font designed to benefit retention, would affect JOLs and recall relative to Arial font.

Sans Forgetica is a specialized font that was developed by researchers at Royal Melbourne Institute of Technology to aide with retention (Earp, 2018). This font was purposely designed to be disfluent and uses an italicized, back-slanted, and hashed style (see Fig. 5 for examples), which has been suggested to facilitate encoding due to its perceptual difficulty (i.e., desirable difficulties; Bjork & Bjork, 2011; Bjork, 1994). This specific font was selected by the Sans Forgetica research team based on the results of an in-lab study in which undergraduates were presented with four fonts (slightly, moderately, and extremely disfluent, plus a fluent control font). Overall, recall was highest for pairs presented using the moderately disfluent font (69%) versus the slightly disfluent (61%) and the extremely disfluent (61%) pair types. Based on these results, the moderately disfluent font was selected as the one most likely to induce desirable difficulties and was eventually branded Sans Forgetica. We note, however, that this memory boost was small, as 68% of pairs in the fluent control group were correctly recalled (Earp). Additionally, as this study was unpublished and its findings not made publicly available, it remains unclear whether recall differences between Sans Forgetica and the control font reached conventional significance.

Recently, several studies have sought to replicate findings from the Sans Forgetica team. Results from these studies, however, suggest that Sans Forgetica does not benefit

Fig. 5 Examples of unrelated word pairs presented in Experiment 3 using Sans Forgetica font (left) and Arial font (right)

memory as originally claimed. Geller et al. (2020) showed that unlike generation (a task which consistently boosts memory via desirable difficulties; see Slamecka & Graf, 1978), Sans Forgetica did not yield an improvement versus a standard Arial font (Experiment 1). Subsequent experiments found that Sans Forgetica was less effective at improving memory than highlighting (Experiment 2) and did not improve performance on an old/new recognition task (Experiment 3). Additionally, research by Taylor et al. (2020) found that although subjects rated Sans Forgetica as more challenging to encode relative to Arial, Sans Forgetica consistently failed to yield a memory benefit and, occasionally, resulted in a memory cost. Taken together, Sans Forgetica does not appear to aid memory.

Despite that the lack of benefits for Sans Forgetica font, an important question is whether the disfluent nature of the font type might affect participants' JOLs at the time of study. In particular, given the findings that perceptually fluent fonts increase JOLs (e.g., Experiments 1A and 2A; Hu et al., 2015; Mueller et al., 2014; Rhodes & Castel, 2008; etc.), it is possible that Sans Forgetica may actually reduce JOLs relative to a standard font type. Consistent with this possibility, Sungkhassettee et al. (2011) presented participants with words that were either presented inverted (i.e., disfluent) or presented in a standard upright position. They found that while recall was enhanced for the more perceptually difficult inverted words (consistent with desirable difficulties), participants reported lower JOLs for disfluent items over multiple study/test cycles. Thus, consistent with a fluency-based account, when compared to a standard presentation condition, disfluent study pairs via Sans Forgetica may produce lower JOLs which may be consistent with the recent evidence suggesting that Sans Forgetica fonts may induce a memory cost (Taylor et al., 2020).

Experiment 3: Unrelated word pairs in Sans Forgetica font

The goal of Experiment 3 was therefore to compare JOLs that are provided on pairs studied in Sans Forgetica font relative to a standard Arial font and evaluate font-type effects on subsequent recall. Though Sans Forgetica is purported to increase recall via desirable difficulties (Earp, 2018), perceptual disfluency might also induce a reduction in JOLs relative to control conditions (Sungkhassettee et al., 2011). Thus, a desirable difficulties account would predict that the disfluency of Sans Forgetica would reduce JOLs at encoding. Separately, however, because Sans Forgetica was developed with the purpose of enhancing memory, an alternative possibility is that participants may hold a belief that Sans Forgetica is beneficial to memory and therefore increase their JOLs at study. While Experiment 3 does not inform participants about the background of Sans Forgetica in advance, participants may hold the belief that Sans Forgetica is beneficial despite clear perceptions of disfluency. While not designed to test a fluency versus beliefs account of JOLs, an increase in JOLs for Sans Forgetica font may indicate that participants beliefs can affect JOLs even when study information is disfluent.

In addition to directly comparing JOLs and recall rates for word pairs presented in Sans Forgetica and Arial font types, we also compare these mixed pairs directly to a pure control group as in Experiments 1 and 2. Specifically, we compare a pure Arial control group to evaluate whether differences in JOLs/recall rates are due to the within-subject context.

Method

Participants

A total of 88 participants completed Experiment 3. Of these participants, 33 were recruited via Prolific using the same recruitment criteria as the previous experiments and were compensated at a rate of $4.00 per half hour. The remaining 55 participants were undergraduate students recruited from The University of Southern Mississippi's psychology research pool who completed the study in exchange for partial course credit. Cued-recall performance was again used as a compliance check. Data from 8 participants were excluded (< 5% correct recall), resulting in 39 participants in the Sans Forgetica/Arial mixed group and 41 in the Arial-only control group, consistent with Experiments 1 and 2. Across both groups, participants reported a mean age of 21.78 ($SD = 6.23$). Full demographics are available in Appendix Table 1. All participants were native English speakers reporting normal or corrected-to-normal vision.

Materials and procedure

Experiment 3 used the same set of unrelated pairs in Experiments 2A and 2B and followed the same general procedure with the following exception. Participants were randomly assigned to either the Sans Forgetica or control groups. Participants in the Sans Forgetica group studied mixed lists in which half of the word pairs were presented in 32-pt. Sans Forgetica font while the other half were presented in a standard, 32-pt. Arial font. For participants assigned to the control group, all pairs were presented 32-pt. Arial font (as in previous experiments). In both groups, participants made JOL ratings concurrently with study. All other materials and procedures were identical to those used in Experiments 2A and 2B.

Results

Figure 6 plots mean JOL and cued-recall percentages for Sans Forgetica and Arial font types in the mixed group as well as JOL/recall rates for the control group. Appendix Table 8 reports all cell means for completeness; Appendix Table 5 reports mean gammas.

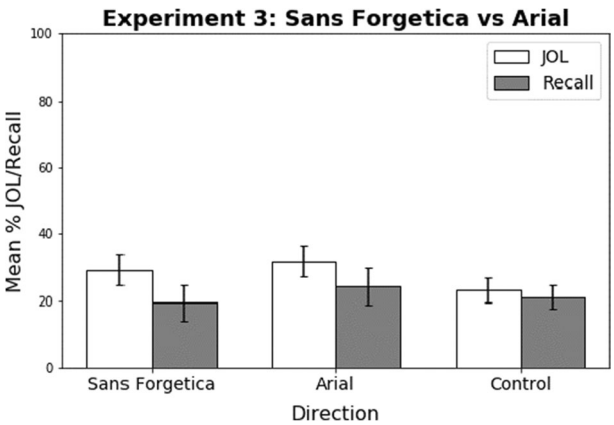

Fig. 6 Mean JOL and recall rates as function of pair type in Experiment 3. Bars represent 95% confidence intervals. All study pairs were unrelated

We first evaluated Sans Forgetica font effects using a 2(Measure: JOL vs. Recall)×2(Font: Sans Forgetica vs. Arial) within-subject ANOVA. Consistent with Experiment 1, an effect of measure was found, $F(1, 38) = 7.69$, $MSE = 383.54$, $\eta_p^2 = .17$, in which JOL rates exceeded correct recall (30.49 vs. 21.79). An effect of font was also found, $F(1, 38) = 17.77$, $MSE = 28.66$, $\eta_p^2 = .32$, in which Sans Forgetica produced *lower* JOL/recall rates relative to Arial font (24.24 vs. 27.95). The interaction was not reliable, $F(1, 38) = 1.98$, $MSE = 25.19$, $p = .17$, $p_{BIC} = .70$.

We then separately compared Sans Forgetica and Arial pairs in the mixed group to the control group. Starting with the comparison between Sans Forgetica pairs and the control group, a 2(Pair Type: Sans Forgetica vs. Control)×2 (Measure: JOL vs. Recall) mixed measures ANOVA yielded an effect of measure, $F(1, 78) = 8.46$, $MSE = 166.33$, $\eta_p^2 = .10$, in which JOLs exceeded recall rates (26.12 vs. 20.82). No difference was found on JOLs/recall rates between Sans Forgetica and control pairs, $F < 1$, $p_{BIC} = .86$, but a marginal interaction was found, $F(1, 78) = 3.64$, $MSE = 166.33$, $p = .06$, $\eta_p^2 = .05$, $p_{BIC} = .59$. Follow-up comparisons indicated that this interaction was due to an illusion of competence pattern for Sans Forgetica pairs, but not control pairs. Specifically, for Sans Forgetica pairs, JOLs exceeded recall rates (29.25 vs. 19.42), $t(38) = 3.06$, $SEM = 3.21$, $d = 0.62$, but for control pairs JOLs and recall rates were well-calibrated (23.14 vs. 21.10), $t < 1$, $p_{BIC} = .82$.

Turning to Arial pairs, a 2(Pair Type: Arial vs. Control)×2 (Measure: JOL vs. Recall) mixed ANOVA again found a significant effect of measure, $F(1, 78) = 5.43$, $MSE = 169.94$, $\eta_p^2 = .07$, in which JOLs exceeded recall (31.73 vs. 24.17). JOLs/recall rates were greater for Arial font pairs than the control pairs (27.95 vs. 22.12), $F(1, 78) = 5.01$, $MSE = 271.12$, $\eta_p^2 = .06$, indicating that although Arial and control pairs were perceptually identical (same font type and size), mixed Arial pairs presented in the same context as Sans Forgetica pairs were rated as more likely to be remembered than control pairs presented without a Sans Forgetica context. The interaction was not reliable, $F(1, 78) = 1.73$, $MSE = 169.94$, $p = .18$, $p_{BIC} = .78$.

Discussion

Experiment 3 tested whether a Sans Forgetica font type would affect JOLs and recall rates relative to a standard Arial font. We expected that if JOLs are sensitive to fluency, then participants should assign lower JOL ratings to the disfluent Sans Forgetica than Arial pairs, but recall of Sans Forgetica pairs should be higher than Arial pairs due to the benefits of desirable difficulties on learning (Bjork & Bjork, 2011). In contrast, if participants hold a belief that Sans Forgetica aids memory, then Sans Forgetica should produce an increase in JOLs relative to Arial pairs. Overall, JOLs were lower for pairs presented using the less fluent Sans Forgetica font, an observation that was both in line with the fluency-based account and suggested that participants did not believe Sans Forgetica would aid retention. Additionally, Sans Forgetica produced a cost to recall when compared to Arial pairs that had been presented within the same study list, suggesting that Sans Forgetica does not operate as a desirable difficulty and is instead costly to memory. Furthermore, Arial pairs in the mixed list received higher JOLs and were recalled at a greater rate relative to the control group, which presented pairs using the same font and size. These findings suggest that participants favor more fluent fonts when placed in the same context as disfluent fonts, and our inclusion of a control group allowed us to test this context effect.

General discussion

The primary goal of our study was to evaluate the effects of perceptual fluency on JOLs and on the subsequent recall of word pairs. We based our study on Rhodes and Castel (2008), who reported a font-size effect in which JOLs were inflated for pairs presented in a perceptually large font relative to small font but had no effect on later recall. The present study similarly evaluated the font-size effect in addition to testing highlighting (vs. not highlighting) and Sans Forgetica (vs. a standard Arial) font type on JOLs and recall rates.

Experiment 1A examined the font-size effect using a set of related (i.e., forward, backward, and symmetrical paired associates) and unrelated word pairs presented in the same study list. We expected that pairs presented using a large font, which are more perceptually fluent and thus perceived as easier to encode, would have inflated JOLs relative to small-font pairs without affecting recall. We also compared these mixed-list font sizes to a pure-control group in which all pairs were presented using a standard font size. The control comparison allowed us to evaluate large and small-font size effects relative to a baseline font size. While the large-font size increased both JOLs and recall rates similarly relative to small-font pairs, JOLs for large-font word pairs did not increase at a greater rate than recall as predicted by the font-size effect (e.g., Rhodes & Castel, 2008). Finally, neither JOLs nor recall differed as a function of font-size when compared to the control group.

Experiment 1B used the same set of related and unrelated stimuli pairs to test whether the font-size patterns extended to highlighted pairs. We similarly expected highlighted pairs would be more perceptually fluent, leading to an increase in JOLs relative to non-highlighted pairs but have no effect on recall rates, a pattern consistent with font-size effect as originally reported by Rhodes and Castel (2008). Overall, highlighted pairs affected neither JOLs nor recall rates when compared to non-highlighted pairs in either the mixed or the pure control group.

Because perceptual manipulations did not differentially affect JOLs versus recall rates in Experiment 1 when using a mixed list of related and unrelated pairs, Experiments 2A and 2B provided an additional test of potential font-size and highlighting effects on JOLs using lists that only contained unrelated pairs. Experiment 2A again found that large-font sizes significantly increased both JOLs and recall rates relative to small-font pairs and marginally relative to control pairs, but again, the standard font-size effect was not found, as JOLs and recall both increased as a function of font-size. In Experiment 2B, the highlight manipulation again produced no effect on either JOLs or recall rates. Thus, font size and highlighting effects were not due to differences in the related versus unrelated pairs presented within the study list.

Finally, Experiment 3 evaluated the effects of Sans Forgetica font on JOLs. We selected this font because, although it is perceptually disfluent, it is suggested to enhance recall through desirable difficulties (Earp, 2018). While several recent studies have indicated that presenting study materials using Sans Forgetica does not benefit memory (e.g., Geller et al., 2020; Taylor et al., 2020), JOLs may be sensitive to the perceptually disfluent nature of Sans Forgetica. However, we also posited that given that Sans Forgetica was developed with the purpose of enhancing memory, participants may expect Sans Forgetica to benefit memory which could increase JOLs. Consistent with a fluency-based account, Sans Forgetica was found to decrease JOLs relative to the Arial font, but only when Sans Forgetica was compared to the Arial pairs in the mixed list and not the pure Arial pairs in the control group. Recall was also impacted by font type, as recall of Sans Forgetica pairs was lower relative to mixed Arial pairs but not the pure Arial pairs—a Sans Forgetica cost.

Interestingly, mixed Arial pairs produced greater JOLs and recall rates than Arial pairs in the pure group, suggesting that the mixed list context increased both participants' JOLs and the encoding of Arial pairs. Thus, the disfluent nature of Sans Forgetica results in lower JOLs relative to a standard Arial control pairs while producing no recall benefit.

In addition to our use of other perceptual manipulations beyond font-size, an important distinction between the current study and Rhodes and Castel (2008) is that each of our experiments included a pure-control group, a novel comparison that has not been included in previous font-size experiments. This control task involved participants studying only one type of word pair rather than both perceptually fluent and disfluent word pairs like in the experimental groups. The inclusion of this task allowed us to assess the effects of list presentation (e.g., mixed vs pure lists) and control for potential carryover effects (e.g., Huff et al., 2021). Relative to the control group, Experiments 1 and 2 showed no significant effect of font-size or highlighting on JOLs or recall. Experiment 3, however, showed an increase to both JOLs and recall for Arial pairs in the mixed list relative to the same pairs in the control group. Thus, our inclusion of pure list control groups allowed us to assess the effects of context on fluency effects, providing a more complete assessment of how these processes affect both JOLs and recall.

Finally, while not a primary focal point of the current study, illusion of competence patterns (Koriat & Bjork, 2005; Maxwell & Huff, 2021) consistently emerged across experiments. In Experiment 1, JOLs overpredicted recall for backward, symmetrical, and unrelated pairs, regardless of font-size, highlights, or control group pairs. For forward associates, however, JOLs and recall were well calibrated. This replicated findings by Maxwell and Huff (2021), who showed that JOLs consistently overpredicted correct recall for study pairs in which the cue was not predictive of the target. Additionally, the illusion of competence pattern extended to unrelated pairs in Experiment 2 and 3, such that JOLs again overpredicted recall, regardless of perceptual fluency or encoding group (e.g., mixed lists or control).

Taken together, our experiments showed that memory predictions were largely unaffected by manipulations designed to affect perceptual fluency. This finding was surprising given that the font-size effect has been shown to be robust and has a replicable pattern (see Halamish et al., 2018). However, even though we used large sample sizes and tested across a variety of modalities, the expected increase in JOLs while having no effect on recall did not occur. Although the same font-size pattern as reported by Rhodes and Castel (2008) did not emerge, we note that Experiments 1A and 2A did show that large font-size increased JOLs. Our effect of large font on JOLs was smaller however than that reported by Rhodes and Castel, despite a greater size difference between pairs in our experiment (12-pt. vs. 54-pt.) than those used previously (18-pt. vs. 48-pt.). Regardless, however, the increase in both JOLs and recall rates for larger pairs is consistent with other studies (e.g., Miele et al., 2011; Susser et al., 2013; Undorf et al., 2017; Yang et al., 2018).

While font-size produced a small benefit to both JOLs and recall, neither highlighting nor Sans Forgetica font type produced a similar increase to JOLs or recall. Indeed, Sans Forgetica produced a mixed-list cost to memory relative to Arial font. This finding lends support to a growing body of literature suggesting that Sans Forgetica may not be an effective tool for improving retention (e.g., Geller et al., 2020; Taylor et al., 2020). Although Sans Forgetica was designed to improve retention through desirable difficulties, which have been suggested to contribute to other memory benefits (e.g., spacing and retrieval-practice effects; see Maddox, 2016, and Rowland, 2014, for reviews), it is not always clear what constitutes a sufficient level of difficulty to promote memory (e.g., McDaniel & Butler, 2010). Regarding the desirable difficulty

of Sans Forgetica, recent work by Eskenazi and Nix (2021) has shown that within the context of learning, any benefits of this font may be moderated by individual differences, such as spelling and reading level. However, a trend is emerging in which Sans Forgetica may be more appropriately termed "Sans Remembrica" due to its potential memory costs.

Though the expected font-size effect patterns were not found and the increases to JOLs and recall observed for large font in Experiments 1A and 2A did not extend to the other perceptual manipulations, we note a few methodological departures within our study that merit further discussion. First, across experiments, participants studied cue-target pairs followed by a cue-recall test. Rhodes and Castel (2008) initially found the font-size effect by having participants study individual words (vs. word pairs) followed by free-recall testing—a pattern that has been reported by others (e.g., Mueller et al., 2014; Yang et al., 2018). However, we also note that the font-size effect has been found on cued-recall of word pairs (Price et al., 2016; Rhodes & Castel, 2008; Experiment 3), indicating that the presence versus absence of a font-size effect is likely not due to differences in study and test formats. A second possibility is that whereas Price et al. and Rhodes and Castel used experimenter-paced encoding, our participants self-paced their study. But again, the font-size effect has also been found using participant-paced encoding (Su et al., 2018), suggesting that the font-size effect cannot be chalked up to differences in encoding type. Finally, when generating related and unrelated study materials in our study, we were careful to match all pair types on various lexical and semantic characteristics which may affect cued-recall rates (see Appendix Table 3). Previous font-size effect studies that found the font size memory enhancement for related but not unrelated materials only controlled for a subset of these characteristics (e.g., word frequency), and it was unclear whether these characteristics were matched across counterbalances. It is possible that lexical and semantic differences may have covaried with the relatedness between the study materials contributing to the font-size effect. Of course, evaluating the interaction between item characteristics and the memory benefits of large fonts is outside the scope of our study, but we highlight this discrepancy to be focal for future research.

Conclusion

The present study investigated the effects of perceptual fluency on JOLs and recall by (1) evaluating the font-size effect and (2) testing whether highlighted pairs and perceptually disfluent Sans Forgetica font would affect JOLs and subsequent cued-recall. While font-size increased both JOLs and recall rates similarly for lists with both related and unrelated pairs and in lists with only unrelated pairs, JOLs were not overly inflated relative to recall. The highlighting manipulation, however, produced no effect on JOLs or recall, regardless of list type. Furthermore, Experiment 3 showed that Sans Forgetica font, which was designed to improve retention, can induce a memory cost under certain circumstances. Finally, our inclusion of control groups within each experiment provided a baseline comparison group given possible mixed-list carryover effects that have previously been unaccounted for within this context. Collectively, this set of experiments provides a greater understanding of how perceptual features influence JOLs and recall, particularly within the context of cued-recall testing.

Appendix

Table 1 Demographic information for participants in each encoding group in Experiments 1, 2, and 3

Variable/Group	Control	Font-Size	Highlight	Sans Forgetica
Experiment 1				
n	39	41	41	–
Age (years)	26.49 (10.29)	32.24 (16.74)	25.55 (13.81)	–
% Female	64%	63%	61%	–
% White	87%	78%	68%	–
% College	69%	54%	54%	–
Experiment 2				
n	32	36	37	–
Age (years)	25.28 (9.72)	24.06 (10.68)	24.35 (10.15)	–
% Female	88%	69%	70%	–
% White	66%	69%	57%	–
% College	69%	34%	51%	–
Experiment 3				
n	41	–	–	39
Age (years)	21.15 (4.82)	–	–	22.46 (7.44)
% Female	71%	–	–	82%
% White	68%	–	–	72%
% College	76%	–	–	74%

Age rows report mean age in years. Parentheses denote *SDs*. % College rows denote percentage of respondents indicating that they have completed partial college coursework or greater. Participants were required to be native English speakers who had achieved at least a high school education or equivalent

Table 2 Mean associative strength summary statistics forward, backward, and symmetrical pairs in Experiment 1A and 1B

Condition	Variable	*M*	*SD*	*Min.*	*Max.*
Forward	FAS	.37	.21	.05	.81
	BAS	.00	.00	.00	.00
Backward	FAS	.00	.00	.00	.00
	BAS	.37	.21	.05	.81
Symmetrical	FAS	.19	.13	.01	.46
	BAS	.19	.13	.02	.52

FAS (forward associative strength) and BAS (backward associative strength) values for unrelated pairs as these items share zero associative overlap

Table 3 Summary statistics for cue and target concreteness, length, and frequency item properties as a function of pair type in Experiments 1A and 1B

Pair Type	Position	Variable	*M*	*SD*	*Min.*	*Max.*
Forward	Cue	Concreteness	4.97	1.22	2.61	6.53
		Length	6.20	1.86	3	10
		Frequency	3.74	0.67	2.44	5.29
	Target	Concreteness	4.96	1.14	2.40	6.67
		Length	4.46	1.27	3	8
		Frequency	2.49	0.63	1.59	4.86
Backward	Cue	Concreteness	4.96	1.14	2.40	6.67
		Length	4.46	1.27	3	8
		Frequency	2.49	0.63	1.59	4.86
	Target	Concreteness	4.97	1.22	2.61	6.53
		Length	6.20	1.86	3	10
		Frequency	3.74	0.67	2.44	5.29
Symmetrical	Cue	Concreteness	4.93	1.36	2.35	6.86
		Length	5.05	1.62	3	10
		Frequency	3.27	0.61	1.66	4.37
	Target	Concreteness	4.44	1.37	2.05	6.53
		Length	5.38	2.23	3	13
		Frequency	3.18	0.73	1.68	5.50
Unrelated	Cue	Concreteness	4.59	1.40	2.18	6.93
		Length	5.13	1.56	3	11
		Frequency	3.20	0.80	1.28	4.76
	Target	Concreteness	4.67	1.15	2.88	6.63
		Length	5.30	1.49	3	8
		Frequency	3.18	0.90	0.95	4.96

Frequency is measured using SUBTLEX word frequency measure (Brysbaert & New, 2009). Concreteness and length were taken from the English Lexicon Project (Balota et al., 2007)

Table 4 Mean (± 95% CI) JOL ratings and correct recall percentages as a function of associative direction (forward, backward, symmetrical, and unrelated) for the control, highlight, and font size groups in Experiments 1A and 1B

Pair Type/Group	Forward	Backward	Symmetrical	Unrelated	Overall
JOL Ratings					
Control Group	60.87 (3.85)	55.18 (4.07)	64.84 (3.74)	19.43 (4.76)	50.08 (3.50)
Font Size Group					
Large Items	62.76 (4.68)	59.59 (4.33)	66.74 (4.30)	16.81 (4.18)	51.47 (3.82)
Small Items	60.93 (4.83)	56.81 (5.41)	62.76 (4.56)	16.31 (3.92)	49.21 (3.77)
Highlight Group					
Highlight Items	61.95 (5.02)	57.86 (4.33)	65.53 (4.43)	18.55 (4.76)	50.97 (3.70)
No Highlight Items	61.32 (4.08)	56.55 (4.24)	64.17 (4.40)	19.26 (4.49)	50.33 (3.51)
Correct Recall %					
Control Group	69.29 (5.39)	31.67 (5.29)	57.76 (5.62)	8.85 (2.50)	41.89 (3.50)
Font Size Group					
Large Items	67.76 (6.33)	33.47 (6.47)	59.81 (5.64)	9.43 (3.00)	42.62 (4.47)
Small Items	65.67 (5.72)	31.06 (5.79)	55.67 (5.58)	9.40 (3.06)	40.65 (4.23)
Highlight Group					
Highlight Items	68.51 (5.20)	33.51 (5.93)	56.27 (6.39)	12.90 (3.94)	42.80 (4.24)
No Highlight Items	67.69 (4.36)	32.32 (5.65)	57.53 (5.76)	12.91 (5.28)	42.61 (4.21)

Right-most column reports mean JOLs/Recall percentages collapsed across pair direction

Table 5 Mean (± 95% CI) Goodman-Kruskal Gamma correlations between JOLs and Recall for each pair type/encoding group as a function of associative direction in Experiments 1A and 1B and for unrelated pairs in Experiments 2A, 2B, and 3

Pair Type/Group	Forward	Backward	Symmetrical	Unrelated	Overall
Experiment 1					
Control Group	.25 (.08)	.19 (.08)	.20 (.07)	−.01 (.17)	.17 (.06)
Font Size Group					
Large Items	.21 (.11)	.17 (.14)	.29 (.11)	.27 (.20)	.23 (.07)
Small Items	.16 (.13)	.23 (.12)	.29 (.11)	.12 (.22)	.20 (.08)
Highlight Group					
Highlight Items	.26 (.14)	.25 (.13)	.21 (.13)	.26 (.16)	.25 (.07)
No Highlight Items	.26 (.10)	.32 (.12)	.24 (.10)	−.16 (.16)	.20 (.07)
Experiment 2					
Control Group	–	–	–	.38 (.08)	–
Font Size Group					
Large Items	–	–	–	.31 (.11)	–
Small Items	–	–	–	.37 (.09)	–
Highlight Group					
Highlight Items	–	–	–	.12 (.15)	–
No Highlight Items	–	–	–	.28 (.13)	–
Experiment 3					
Control Group	–	–	–	.34 (.07)	–
Sans Forgetica Group					
Sans Forgetica Font	–	–	–	.33 (.12)	–
Standard Font	–	–	–	.28 (.11)	–

All study/test items were unrelated in Experiments 2 and 3. Right-most column denotes mean gamma collapsed across associative direction in Experiment 1. Gamma analyses are available at https://osf.io/xymez/

Table 6 Summary statistics for cue and target concreteness, length, and frequency item properties for unrelated pairs in Experiments 2A, 2B, and 3

Position	Variable	*M*	*SD*	*Min.*	*Max.*
Cue	Concreteness	4.55	1.24	2.18	6.93
	Length	5.16	1.50	3	11
	Frequency	3.04	0.84	0.95	4.96
Target	Concreteness	4.20	1.42	1.56	6.78
	Length	5.10	1.36	3	8
	Frequency	3.13	0.76	1.18	5.47

Frequency is measured using SUBTLEX word frequency measure (Brysbaert & New, 2009). Concreteness and length were taken from the English Lexicon Project (Balota et al., 2007)

Table 7 Mean (± 95% CI) JOL ratings and correct recall percentages for the control, highlight, and font size groups in Experiments 2A and 2B

Group	JOL Rating	Correct Recall %
Control	24.76 (5.05)	19.09 (4.95)
Font Size		
Large	29.27 (4.89)	24.72 (5.25)
Small	26.72 (4.42)	21.86 (4.94)
Highlight		
Highlight	30.42 (5.40)	21.60 (5.05)
No Highlight	29.39 (5.57)	20.35 (5.63)

All study/test items were unrelated

Table 8 Mean (± 95% CI) JOL ratings and correct recall percentages for the control, and Sans Forgetica groups in Experiment 3

Group	JOL Rating	Correct Recall %
Control Group	23.14 (3.56)	21.10 (4.59)
Sans Forgetica Group		
Sans Forgetica Font	29.25 (4.59)	19.42 (5.31)
Standard Font	31.73 (4.64)	24.17 (5.51)

All study/test items were unrelated

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11409-021-09284-6.

Funding There is no funding source.

Data availability Study materials and analyzed data are available via OSF (https://osf.io/3xwdr/). This study was completed as part of the Honors Thesis requirements for Trevor Perry.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All experiments reported were approved by the authors institutional review board.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

Ball, B. H., Klein, K. N., & Brewer, G. A. (2014). Processing fluency mediates the influence of perceptual information on monitoring learning of educationally relevant materials. *Journal of Experimental Psychology: Applied, 20*(4), 336.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*(3), 445–459.

Besken, M. (2016). Picture-perfect is not perfect for metamemory: Testing the perceptual fluency hypothesis with degraded images. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(9), 1417.

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).

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society, 2*(59–68).

Bodner, G. E., Taikh, A., & Fawcett, J. M. (2014). Assessing the costs and benefits of production in recognition. *Psychonomic Bulletin & Review, 21*(1), 149–154.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*, 977–990.

Castel, A. D., McCabe, D. P., & Roediger, H. L. (2007). Illusions of competence and overestimation of associative memory for identical items: Evidence from judgments of learning. *Psychonomic Bulletin & Review, 14*(1), 107–111.

Earp, J. (2018). Q&A: Designing a font to help students remember key information.

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.

Fowler, R. L., & Barker, A. S. (1974). Effectiveness of highlighting for retention of text material. *Journal of Applied Psychology, 59*(3), 358.

Garcia, M. & Kornell, N. (2015). Collector [computer software]. Retrieved from https://github.com/gikeymarica/Collector. Accessed 3 April 2020.

Geller, J., Davis, S. D., & Peterson, D. J. (2020). Sans forgetica is not desirable for learning. *Memory, 28*(8), 957–967.

Halamish, V., Nachman, H., & Katzir, T. (2018). The effect of font size on children's memory and metamemory. *Frontiers in Psychology, 9*, 1577.

Hanczakowski, M., Zawadzka, K., Pasek, T., & Higham, P. A. (2013). Calibration of metacognitive judgments: Insights from the underconfidence-with-practice effect. *Journal of Memory and Language, 69*(3), 429–444. https://doi.org/10.1016/j.jml.2013.05.003

Hu, X., Li, T., Zheng, J., Su, N., Liu, Z., & Luo, L. (2015). How much do metamemory beliefs contribute to the font-size effect in judgments of learning? *PLoS One, 10*(11), e0142351.

Huff, M. J., Bodner, G. E., & Gretz, M. R. (2021). Distinctive encoding of a subset of DRM lists yields not only benefits, but also costs and spillovers. *Psychological Research, 85*, 280–290.

Jemstedt, A., Schwartz, B. L., & Jönsson, F. U. (2018). Ease-of-learning judgments are based on both processing fluency and beliefs. *Memory, 26*(6), 807–815.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 187–194.

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science, 22*(6), 787–794.

Maddox, G. B. (2016). Understanding the underlying mechanism of the spacing effect in verbal learning: A case for encoding variability and study-phase retrieval. *Journal of Cognitive Psychology, 28*(6), 684–706.

Masson, M. E. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43*(3), 679–690.

Maxwell, N. P., & Huff, M. J. (2021). The deceptive nature of associative word pairs: Effects of associative direction on judgments of learning. *Psychological Research, 85*(4), 1757–1775.

McDaniel, M. A., & Butler, A. C. (2010). A contextual framework for understanding when difficulties are desirable. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A Festschrift in honor of Robert a. Bjork* (pp. 175–198). Psychology Press.

Miele, D. B., Finn, B., & Molden, D. C. (2011). Does easily learned mean easily remembered?: It depends on your beliefs about intelligence. *Psychological Science, 22*(3), 320–324.
Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language, 70*, 1–12.
Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The delayed-JOL effect. *Psychological Science, 2*, 267–270.
Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation, 26*, 125–173.
Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36*(3), 402–407.
Price, J., & Harrison, A. (2017). Examining what prestudy and immediate judgments of learning reveal about the bases of metamemory judgments. *Journal of Memory and Language, 94*, 177–194.
Price, J., McElroy, K., & Martin, N. J. (2016). The role of font size and font style in younger and older adults' predicted and actual recall performance. *Aging, Neuropsychology, and Cognition, 23*(3), 366–388.
Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition, 8*(3), 338–342.
Reber, R., Winkielman, P., & Schwarz, N. (1998). Effects of perceptual fluency on affective judgments. *Psychological Science, 9*(1), 45–48.
Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137*(4), 615–625.
Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432.
Schwarz, N. (2004). Metacognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology, 14*(4), 332–348.
Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 592–604.
Soderstrom, N. C., Clark, C. T., Halamish, V., & Bjork, E. L. (2015). Judgments of learning as memory modifiers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(2), 553–558.
Su, N., Tongtong, L., Zheng, J., Hu, X., Fan, T., & Luo, L. (2018). How font size affects judgments of learning: Simultaneous mediating effect of item-specific beliefs about fluency and moderating effect of beliefs about font size and memory. *PLoS One, 13*, e0200888:1–e020088814.
Sungkhassettee, 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*, 973–978.
Susser, J. A., Mulligan, N. W., & Besken, M. (2013). The effects of list composition and perceptual fluency on judgments of learning (JOLs). *Memory & Cognition, 41*, 1000–1011.
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.
Undorf, M., Zimdahl, M. F., & Bernstein, D. M. (2017). Perceptual fluency contributes to effects of stimulus size on judgments of learning. *Journal of Memory and Language, 92*, 293–304.
Wagenmakers, E. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14*(5), 779–804.
Wehr, T., & Wippich, W. (2004). Typography and color: Effects of salience and fluency on conscious recollective experience. *Psychological Research, 69*, 138–146.
Yang, C., Huang, T. S. T., & Shanks, D. R. (2018). Perceptual fluency affects judgments of learning: The font size effect. *Journal of Memory and Language, 99*, 99–110.
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.