Perceptually Distinctive Features of Study Words Do Not Inflate Judgements of Learning: Evidence from Font Size, Highlights, and Sans Forgetica Font Type

Nicholas P. Maxwell, Trevor Perry, & Mark J. Huff

The University of Southern Mississippi

**Author Note**

Correspondence concerning this article can be addressed to Mark. J. Huff, School of Psychology, The University of Southern Mississippi, 118 College Dr. #5025, Hattiesburg, MS 39406, United States. Study materials and analyzed data are available via OSF [LINK]. This study was completed as part of the Honors Thesis requirements for Trevor Perry. Email: mark.huff@usm.edu

Abstract

Effective monitoring is important for successful learning, and the judgment of learning task (JOL) is often used to assess monitoring at encoding. In the JOL task, participants study a cue-target word pair (e.g., mouse-cheese) and are asked to rate the probability of correcting recalling the target word (e.g., cheese) if shown only the cue word (e.g., mouse). Prior research has shown that JOL accuracy is sensitive to perceptual cues. These cues can produce metacognitive 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 font relative to small font. The present study provides an additional test of the font-size effect and tests whether other perceptual manipulations can affect the correspondence between JOLs and recall. First, Experiment 1 was designed to replicate the font-size effect and test whether the effect extended to highlighting while using a set of related and unrelated word pairs. Experiment 2 provided an additional test of font size and highlighting effects on JOLs using only unrelated pairs. Finally, Experiment 3 tested whether Sans Forgetica—a perceptually distinctive font specifically designed to improve memory—would result in inflated JOLs. Across experiments, the perceptually distinctive conditions did not result in an overestimation of later recall relative to non-distinctive conditions, and Sans Forgetica font in Experiment 3 yielded a memory cost (though no effect on JOLs). Collectively, perceptually distinctive study items do not appear to inflate JOLs at study.

Word Count: 245

*Keywords:* XX; XX; XX; XX

Perceptually Distinctive Features of Study Words Do Not Inflate Judgements of Learning: Evidence from Font Size, Highlights, and Sans Forgetica Font Type

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). Metacognitive 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 can use several types of judgments to assess metacognitive processes, the judgment of learning task (JOL) is commonly used. In a standard JOL task, participants are presented with a cue-target item pair (e.g., mouse-cheese) and are asked to estimate their likelihood of correctly retrieving the target word (e.g., cheese) at test if prompted only by the cue (e.g., mouse). While JOL ratings can be made using a variety of measurement scales (e.g., Likert Scales or binary “yes-no” responses; Hanczakowski, Zawadzka, Pasek, & Higham, 2013), JOLs 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 an easy 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 result in metacognitive illusions in which JOLs either underpredict or overpredict later recall. For example, the associative direction between cue-target pairs has repeatedly been shown to induce an *illusion of competence* in which JOLs overpredict later recall for certain pair types (Koriat & Bjork, 2005). Specifically, forward associates in which the cue is highly predictive of the target (e.g., base-ball) 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., ball-base) display a marked overconfidence effect in which JOLs greatly overestimate future memory performance. Subsequent research by Castel, McCabe, and Roediger (2007) extended this finding to include identical word pairs. More recently, Maxwell and Huff (in press) further examined the relationship between associative direction and JOL magnitude by including symmetrical associates (e.g., king-queen), in which the forward and backward relations between pairs are matched. Similar to findings reported by Koriat & Bjork (2005), Maxwell and Huff (in press) found that JOL ratings were generally well calibrated for forward pairs but produced an illusion of competence that overestimated actual recall performance for backward, symmetrical, and unrelated word pairs. Additionally, they showed that the illusion of competence was highly robust and persisted across a variety of experimental manipulations designed to improve the correspondence between JOLs and recall. Thus, JOLs can be accurate predictors of recall in some situations but can also produce illusions of competence under different situations and manipulations.

In addition to associative pair direction, other factors have been shown to influence the magnitude of judgments. For example, perceptual manipulations have been shown to be highly effective at influencing judgment magnitude. Commonly, 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, research by Reber, Winkielman, & Schwarz (1998) has shown that participants judge perceptually fluent items as being more affectively pleasing relative to disfluent items. Furthermore, Winkielman and Cacioppi (2002) found that more easily processed pictures elicited a positive emotional response, indicating that the ease of processing of a stimulus produces positive affect. Finally, Reber and Shwarz (1999) showed that participants are more likely to judge perceptually fluent statements as being true compared to less perceptually fluent statements.

Previous research has shown that the ease-of-processing effect extends to JOLs and influences both the magnitude and accuracy of these judgments. For example, Rhodes and Castel (2008), presented participants with a mix of stimuli consisting of large and small font words. Overall, they demonstrated a font-size effect in which JOLs were inflated when stimuli were presented in a large font relative to a small font. Critically, recall performance did not change as a function of font-size, indicating that perceptually fluency primarily affected judgment making rather than memory performance. Subsequent experiments performed by Rhodes and Castel (2008) revealed that this effect was largely driven by the additional ease-of-processing afforded by the large font items. For example, the font-size effect was largely diminished when ease of reading was manipulated, and the effect was moderated by pair relatedness (Rhodes & Castel, 2008).

The font size effect has been replicated numerous times since initially being reported by Rhodes and Castel. For example, Kornell, Rhodes, Castel, and Tauber (2011) replicated the font size effect and showed that effect occurs when multiple presentations of words are used. More recently, Hu, Li, Zheng, Su, Liu, & Luo (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 words 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 word that was being viewed. Again, the font-size effect replicated as JOLs were higher for pairs presented in large font, regardless of whether participants were assigned to the study or observer group. 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 font higher relative to small font, regardless of whether the JOL was made pre- or post-study. Thus, the font-size effect appears to be highly robust and replicates across a variety of judgment manipulations.

Although the font-size effect has been extensively replicated under a variety of conditions, the underlying factors driving the effect remain unclear. Two theories have been put forth to explain the font-size effect. First, the fluency account states that larger words are more perceptually fluent than smaller words. Due to the greater perceptual fluency of larger words, participants predict that they will be more likely to recall those words in the future compared to smaller, less perceptually fluent words. To test the fluency account, Undorf, Zimdahl, and Bernstein (2017) presented participants with images of objects, faces, and words and incrementally increased the size in which stimuli were presented. Participants were asked to make a JOL once they could recognize the stimulus, with the time that it took participants to reach recognition recorded as a measure of perceptual fluency. The results showed that JOLs increased as the time it took to identify the stimulus decreased, supporting the hypothesis that the perceptual fluency affects the magnitude of JOLs. Similarly, Yang, Huang, and Shanks (2018) tested the fluency account by comparing the results of a continuous identification task (CID) in which [EXPLAIN HERE] and lexical decision tasks. Like Undorf et al. (2017), they found higher JOL ratings for words that could be identified faster and were thus more perceptually fluent in the CID task which provides evidence that the perceptual fluency differences in the font size effect accounts for the JOL effects.

Whereas the fluency account is based solely on how the perceptual fluency of study items influences JOLs, the beliefs account posits that participants beliefs about item fluency (rather than fluency itself) are the primary factor influencing JOL magnitude. Within the context of the font-size effect, participants assign higher JOLs to large items because they hold the belief that large pairs are easier to learn relative to small items. To test the beliefs account, Mueller, Dunlosky, Tauber, and Rhodes (2014) had participants first complete a lexical decision task for a set of large vs small items. Overall, they showed that response latencies on the lexical decision task did not differ as a function of font-size. Additionally, in their second experiment, they showed that study times did not differ between font sizes. Because response times did not differ based on font-size in either experiment 1 or 2, Mueller et al., 2014 surmised that the font-size effect was not driven by fluency, instead advocating for a beliefs-based account. To test this hypothesis, experiment 3 had participants complete a questionnaire that had them predict their recall performance for small and large words, while experiment 4 had participants make pre-study JOLs about the to-be-studied words, removing potential fluency effects. Overall, responses on the questionnaires and pre-study JOLs each suggested that participants generally hold the belief that large items (e.g., higher fluency) are easier to recall, a finding that is consistent with the beliefs account.

While heavy emphasis has been placed on the font-size effect, other perceptual manipulations have also been shown to similarly effect JOLs. For example, Ball, Klein, and Brewer (2014) tested how bolding word pairs affected JOL ratings and memory performance relative to non-bolded items. Overall, they showed that bolded items were recieved significantly higher JOLs relative to the non-bolded items. However, like the font-size effect, no differences in recall performance were detected between the two conditions. Additionally, Besken (2016) had participants complete a memory task in which images were presented either intact images or with sections removed (i.e., fluent vs disfluent) and had participants complete a JOL task at encoding. Overall, intact images received higher JOLs relative to incomplete images, indicating that other fluency manipulations beyond font-size can influence the magnitude of JOLs.

The present study provided a further test of the font size effect while extending it to include other perceptual manipulations designed to similarly affect JOLs. Specifically, Experiments 1A and 2A sought to replicate the font size effect using. Next, Experiments 1B and 2B tested whether highlighting word pairs (vs not highlighting) would similarly affect JOLs as font size. Finally, Experiment 3 tested whether this effect would emerge for pairs presented in Sans Forgetica font a highly distinctive font that is more perceptually difficult to process relative to a standard font such as Arial.

Finally, we expand upon previous work (e.g., Rhodes & Castel, 2008) by including within each set of experiments a pure control group comprised of participants who only studied one type of pair (i.e., all pairs presented in a standard font size), rather than a mix of perceptually fluent and disfluent pairs. These additional groups were included because encoding manipulations have been shown to spill over into other encoding tasks when encoding is manipulated within-subjects (Huff, Bodner, & Gretz, in press). Thus, our inclusion of the control groups allowed us to more accurately control for these carry-over effects.

**Experiment 1A: Font-Size Effects on Related and Unrelated Pairs**

The goal of Experiment 1A was 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 distinct, that participants would provide elevated JOL ratings relative to small font pairs, a pattern consistent with previous research (Rhodes & Castel, 2008). We also included comparisons to the control group used in Experiment 1 (items presented using a standard, 32 pt Arial font)) to assess whether any benefits of perceptually distinct pairs hold when compared to a pure list of non-distinct items. Font-size effects were testing for using a mixed list of forward, backward symmetrical and symmetrical paired associates and unrelated pair types.

**Method**

**Participants**

Eighty participants were recruited from Prolific ([www.prolific.co](http://www.prolific.co)), an online academic crowdsourcing platform, and completed the study at rate of $4.00 per half hour. Participants were randomly assigned to the font-size group (*n* = 41) or the control group that completed a standard JOL task at encoding (*n* = 39). Participants reported a mean age of 32.65 (*SD* = 15.29), and all participants were native English speakers who reported normal or corrected-to-normal vision. A sensitivity analysis conducted with *G\*Power* (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that our sample size provided adequate power (.80) to detect a medium effect size (Cohen’s *d* = 0.50) or larger.

**Materials**

One-hundred-eighty word pairs taken from Maxwell and Huff (in press) 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. All 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. Study materials for all experiments have been made available at [OSF LINK]. Associative strength, lexical, and semantic properties are listed in the Appendix (Tables A1 and A2).

All participants studied both lists which were divided into two study/test blocks, the order of which was counterbalanced across participants. Each list was organized in which five buffer pairs were presented at the beginning and end of each study list with the remainder of the pairs presented in a newly randomized order for each participant. Counterbalanced versions were produced from each study list that reversed the order of the word pair lists (cue-target pairs become target-cue and target-cue pairs become cue-target), 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.

The cued-recall test the followed the study list in each block contained all 80 cues from the initial 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. Instructions did not mention font size at encoding in the font-size group.

**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 rating 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 0-100 points. 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 full complete certainty that they would recall the target. Participants were encouraged to utilize the range of the scale and not to anchor on extremes when providing ratings (i.e., 0 or 100 ratings). Following instructions, participants then studied the first block of word pairs and provided JOL concurrently with study such that ratings were provided while the word pair was displayed 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 type the correct target word. The cued-recall test was identical across all groups, and test items were displayed in a standard font (32 pt. Arial font). If participants were unable to retrieve the target, they were asked to press the enter key to advance to the next test item. 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 cue recall test, participants were debriefed and provided with compensation. The duration of the experiment was less than 30 min across all groups.

**Results**

All JOL responses were initially screened for missing responses and outliers (i.e., JOLs made outside the instructed 0-100% range). These 26 responses were removed, resulting in 12774 observations in the final dataset. All missing recall responses were coded as incorrect.

A *p* < .05 significance level was used for all analyses. Effect size estimates using partial-eta squared (*η*p2) 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.

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 control groups. In 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, *η*p2 = .20, in which JOLs exceeded recall percentages (50.34 vs. 41.53). An effect of pair type was also found, *F*(3, 120) = 414.56, *MSE* = 218.49, *η*p2 = .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, *t*s >3.27, *d*s > 0.24. An effect of font size was also found, *F*(1, 40) = 12.20, *MSE* = 66.26, *η*p2 = .23, indicating that JOL/recall percentages overall were greater for large font than small font pairs (47.05 vs. 44.83). Importantly, all interactions with font size, including the three-way interaction, were not reliable, *F*s < 1.63, *p*s > .18, *p*BICs > .99, indicating that the large font size did not differentially inflate JOLs relative to recall rates across pair types (cf. Rhodes & Castel, 2008).

A measure × pair type interaction was found, *F*(3, 120) = 45.27, *MSE* = 146.88, *η*p2 = .53, which indicated the presence of an illusion of competence pattern for some pair types. Specifically, JOLs were well-calibrated to recall on forward pairs (61.84 vs. 66.72, for JOLs and recall percentages, respectively), *t*(40 = 2.28, *SEM* = 3.46, *p* = .17, *p*BIC = .70, but JOLs exceeded recall on symmetrical pairs (64.75 vs. 57.75), *t*(40) = 2.28, *SEM* = 3.08, *d* = 0.40, unrelated pairs (16.56 vs. 9.42), *t*(40) = 2.91, *SEM* = 2.46, *d* = 0.64, and especially on deceptive backward pairs (58.20 vs. 32.26), *t*(40) = 6.95, *SEM* = 3.73, *d* = 1.48.

We then compared JOLs/recall percentages on large and small font pairs relative to the control group to evaluate font size effects relative to a pure group that presented 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, *η*p2 = .92, and the same illusion of competence pattern found in large and small font pairs, *F*(3, 114) = 68.49, *MSE* = 49.12, *η*p2 = .64, no main effects or interactions were found when comparing large and small font size pairs relative to the control group, all *F*s < 1.47, *p*s > .22, *p*BICs> .99. Collectively, increasing font size increased both JOLs and recall percentages equally relative to small font sizes and increased font sizes had no effect when compared to a standard font size control.

**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 in 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. Like Experiment 1A, we expected that highlighting pairs would make pairs perceptually distinctive and increase the likelihood that participants would provide elevated JOL ratings relative to non-highlighted pairs, 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 in both JOLs and recall rates relative to small font pairs in Experiment 1A, it is possible that highlighting items would also increase both JOL and recall percentages equally. We also included comparisons to the control group used in Experiment 1A (non-highlighted items of the same font size) to gauge whether any highlighting benefits would hold when compared to a pure list of non-highlighted items. Again, highlighting effects were compared on forward, symmetrical, backward, and unrelated pair types.

**Method**

**Participants**

An additional 41 participants were recruited from Prolific to complete the study and were compensated at rate of $4.00 per half hour. Participants reported a mean age of 32.68 (*SD* = 15.27), and all were native English speakers who reported normal or corrected-to-normal vision.

**Materials and Procedure**

The same materials and general procedure in Experiment 1A was again used in Experiment 1B with the only difference being the 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 also similar to Experiment 1B and all test items were presented in a non-highlight format.

**Results**

Experiment 1B followed the same data screening procedure as Experiment 1A, and less than 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. A 2(Measure) × 2(Highlight) × 4(Pair Type) within-subject ANOVA yielded an effect of measure, *F*(1, 40) = 7.69, *MSE* = 1346.04, *η*p2 = .16, in which overall, JOLs exceeded recall (50.65 vs. 42.70). An effect of pair type, *F*(3, 120) = 410.75, *MSE* = 197.25, *η*p2 = .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, *t*s > 3.10, *d*s > 0.34. Unlike Experiment 1A, the distinctive 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 *F*s < 1, *p*s > .72, *p*BICs > .99.

The measure × pair type interaction was again significant, *F*(3, 120) = 56.96, *MSE* = 114.88, *η*p2 = .59, indicating an illusion of competence pattern. 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 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. JOLs 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 compared JOLs/recall percentages on the within-subject highlight and no-highlight pairs relative to control group pairs. Like Experiment 1A, no effect or interactions were found when comparing the control group pairs to either of the highlight pairs, all *F*s < 1.56, *p*s > .19, *p*BICs > .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 pure non-highlighted pairs.

**Discussion**

[Discussion paragraph here recapping Ex 1A and B and transitioning to Ex 2]

[Explain why we switched to all unrelated pairs]

**Experiment 2A: Font-Size Effects on Pure Unrelated Lists**

[We’ll need a paragraph that provides an overview of Ex 2]

**Method**

**Participants**

65 participants were recruited from Prolific and completed the study at a rate of $4.00 per half hour. 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 the Data for 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, leading to a total of 68 participants included in the analyses below. Participants reported a mean age of 25.41 (*SD* = 11.24). All participants were native English speakers who reported normal or corrected-to-normal vision.

**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 mix of related and unrelated study pairs. To ensure a sizeable list of unrelated pairs, the 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 Table A3 for lexical characteristics). All other materials, including buffer items, 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. We first compared font size differences in the mixed group using a 2(Measure) × 2(Font Size) within-subject ANOVA. Across pair types, JOLs were not statistically greater than recall rates (27.99 vs. 23.30, *F*(1, 35) = 2.15, *MSE* = 369.08, *p* = .15, *p*BIC = .67, but JOLs/recall rates were greater for large than small fonts (27.00 vs. 24.30), *F*(1, 35) = 19.10, *MSE* = 13.76, *η*p2 = .35. Importantly however, just as in Experiment 1A, font size did not affect JOLs and recall rates differently, *F* < 1, *p*BIC = .85.

When compared to the control group pairs, JOLs were greater than 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, *η*p2 = .06, and when compared to small-font pairs (25.80 vs. 20.53), *F*(1, 66) = 5.75, *MSE* = 164.00, *η*p2 = .08. JOLs/recall rates were marginally greater for large-font pairs compared to control pairs (27.00 vs. 21.89), *F*(1, 66) = 3.54, *MSE* = 249.20, *p* = .06, *η*p2 = .05, *p*BIC = .58, , but difference occurred between small-font pairs and control pairs (24.30 vs. 21.890, *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, *F*s < 1, *p*BICs > .88.

**Experiment 2B: Highlighting Effects on Pure Unrelated Lists**

**Method**

**Participants**

We recruited an additional 37 participants via Prolific to complete Experiment 2B. All participants completed the study at a rate of $4.00 per half hour. Participants reported a mean age of 25.00 (*SD* = 10.39). All participants were native English speakers who reported normal or corrected-to-normal vision.

**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, overall JOLs exceeded later recall rates (29.17 vs. 20.54), *F*(1, 36) = 6.26, *MSE* = 440.59, *η*p2 = .15; however, 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 separately to the control group. Again, JOLs exceeded recall rates both in the highlight /control analysis (27.40 vs. 20.20), *F*(1, 67) = 8.58, *MSE* 201.54, *η*p2 = .11, and in the no-highlight/control analysis (26.85 vs. 19.47), *F*(1, 67) = 8.73, *MSE* = 80.57, *η*p2 = .12. There were no differences when comparing either highlight or no-highlight pairs relative to the control group, *F*s < 1, *p*BICs > .85. The interactions were also not reliable, *F*s < 1, *p*BICs > .87.

**Discussion**

[WORDS HERE]

No effect, so we looked at a different type of perceptual manipulation

**Experiment 3: Unrelated Word Pairs in Sans Forgetica Font**

[Overview of EX 3 here]

**Method**

**Participants**

A total of 86 participants completed the study. Of these participants, 33 were recruited via Prolific and compensated at a rate of $4.00 per half hour and 53 recruited from The University of Southern Mississippi’s psychology research pool and completed the study in exchange for partial course credit. Data from 6 participants were excluded using the same criteria as Experiment 2, resulting in 41 in the control group and 39 in the Sans Forgetica group. Participants reported a mean age of 22.34 (*SD* = 7.33), and all were native English speakers with normal or corrected-to-normal vision.

**Materials and Procedure**

Experiment 3 used the same set of unrelated pairs used in Experiment 2 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 Experiments 1 and 2). In both groups, participants made JOL ratings concurrently with study. All other materials and procedures were identical to those used in Experiment 1.

**Results**

Figure 5 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. We first evaluated Sans Forgetica font effects using a 2(Measure) × 2(Font: Sans Forgetica vs. Arial) within-subject ANOVA. Consistent with Experiment 2, an effect of measure, *F*(1, 38) = 7.69, *MSE* = 383.54, *η*p2 = .17, in which JOL rates exceed correct recall (30.49 vs. 21.79)—an illusion of competence pattern. An effect of font was also found, *F*(1, 38) = 17.77, *MSE* = 28.66, *η*p2 = .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 Sans Forgetica pairs, an effect of measure was found, *F*(1, 78) = 8.46, *MSE* = 166.33, *η*p2 = .10, in which JOLs exceeded recall (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, *η*p2 = .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, an effect of measure was again found, *F*(1, 78) = 5.43, *MSE* = 169.94, *η*p2 = .07, in which JOLs exceeded recall (27.33 vs. 22.59). JOLs/recall rates were greater for Arial font than the control pairs (27.95 vs. 22.12), *F*(1, 78) = 5.01, *MSE* = 271.12, *η*p2 = .06, indicating that although Arial and control pairs were perceptually identical (same font type and size), Arial pairs that were in the same context as Sans Forgetica pairs were rated as more likely to be remembered and remembered more frequently 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**

[WORDS HERE]

**General Discussion**

[WORDS HERE]

**Conclusion**

[WORDS HERE]

References

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.

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.

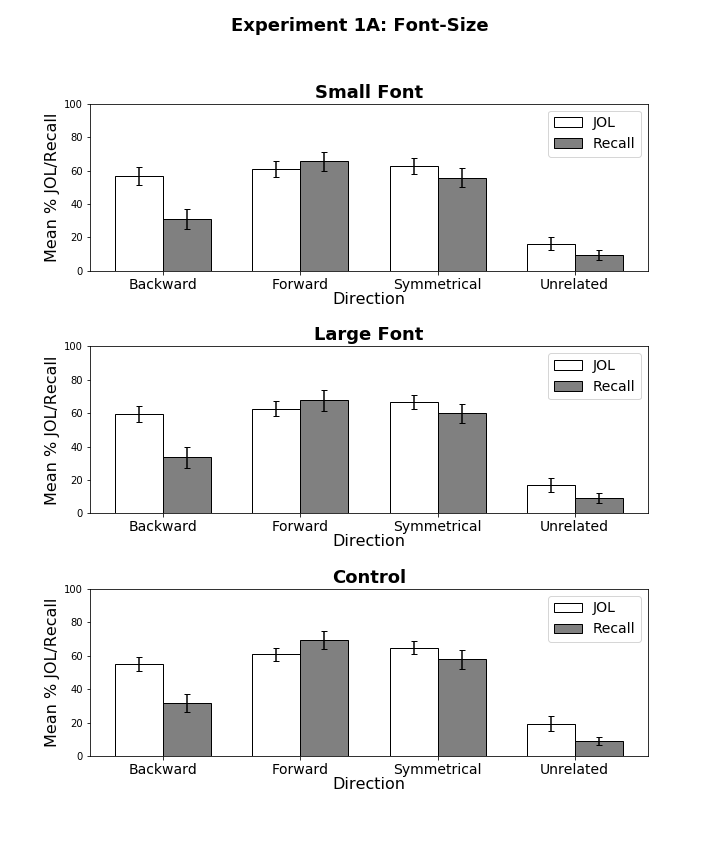
Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G\*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior* *Research Methods*, *39* (2), 175–191.

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

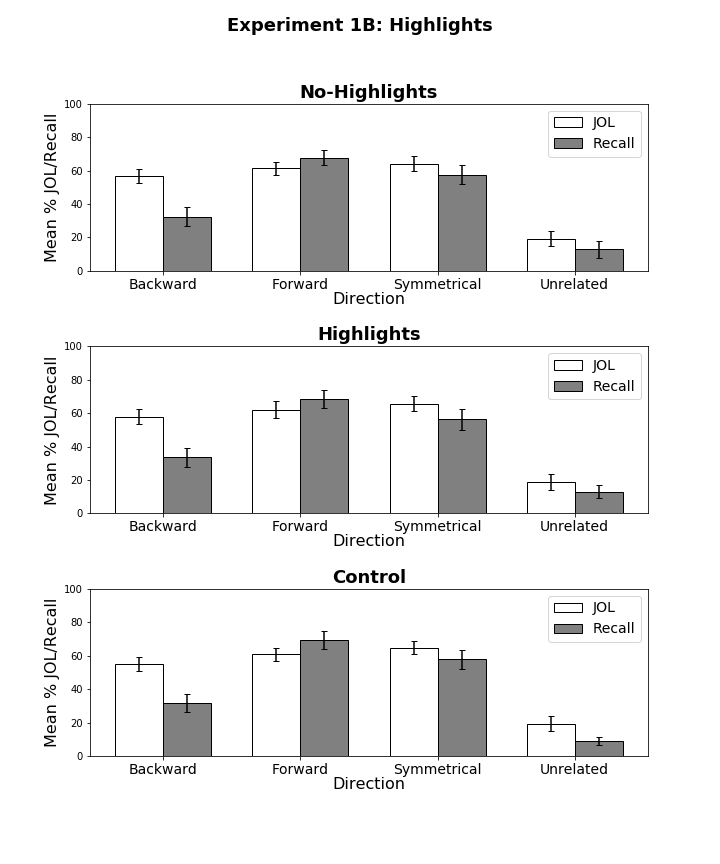
Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one’s knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31* (2), 187–194.

Maxwell, N. P., & Huff, M. J. (in press). The deceptive nature of associative word pairs: Effects of associative direction on judgments of learning. *Psychological Research*, 1-19.

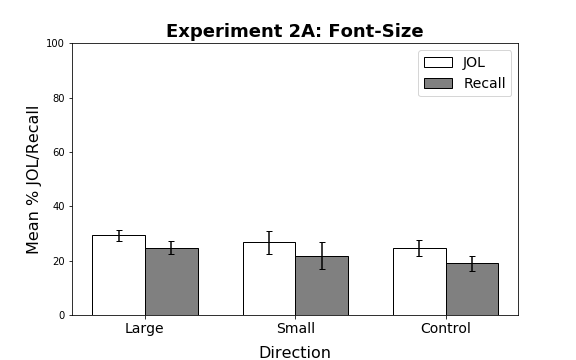
Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods,* *Instruments, & Computers*, *36* (3), 402–407.



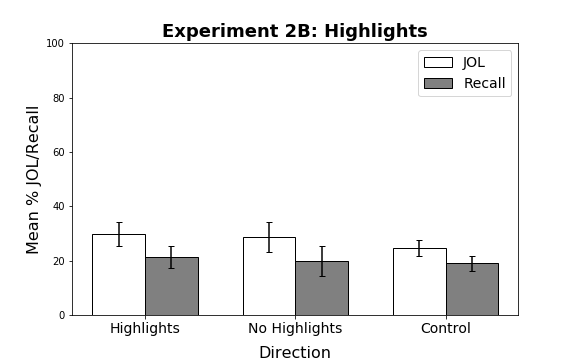
*Figure 1*. Mean JOL and recall rates as a function of pair type for pairs presented in small font (top panel), large font (middle panel), and the control group (bottom panel) in Experiment 1A. Bars represent 95% confidence intervals.



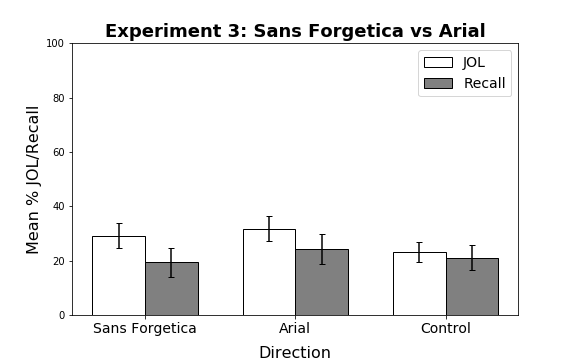
*Figure 2*. Mean JOL and recall rates as a function of pair type for highlighted pairs presented in mixed lists (top panel), non-highlighted pairs presented in mixed lists (middle panel), and non-highlighted pairs presented the control group (bottom panel) in Experiment 1B. Bars represent 95% confidence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



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



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



*Figure 5*. Mean JOL and recall rates as function of pair type in Experiment 3. Bars represent 95% confidence intervals. All study pairs were unrelated.

Table 1

*Mean (± 95% CI) JOL Ratings and Correct Recall Percentages as a Function of Pair Type (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

JOL Ratings

Control Group 60.87 (3.85) 55.18 (4.07) 64.84 (3.74) 19.43 (4.76)

Font Size Group

Large Items 62.76 (4.68) 59.59 (4.33) 66.74 (4.30) 16.81 (4.18)

Small Items 60.93 (4.83) 56.81 (5.41) 62.76 (4.56) 16.31 (3.92)

Highlight Group

Highlight Items 61.95 (5.02) 57.86 (4.33) 65.53 (4.43) 18.55 (4.76)

No Highlight Items 61.32 (4.08) 56.55 (4.24) 64.17 (4.40) 19.26 (4.49)

Correct Recall %

Control Group 69.29 (5.39) 31.67 (5.29) 57.76 (5.62) 8.85 (2.50)

Font Size Group

Large Items 67.76 (6.33) 33.47 (6.47) 59.81 (5.64) 9.43 (3.00)

Small Items 65.67 (5.72) 31.06 (5.79) 55.67 (5.58) 9.40 (3.06)

Highlight Group

Highlight Items 68.51 (5.20) 33.51 (5.93) 56.27 (6.39) 12.90 (3.94)

No Highlight Items 67.69 (4.36) 32.32 (5.65) 57.53 (5.76) 12.91 (5.28)

Table 2

*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 24.26 (4.87) 16.69 (4.69)

Font Size

Large 27.70 (4.80) 22.31 (5.24)

Small 25.53 (4.33) 19.81 (4.84)

Highlight

Highlight 30.34 (5.23) 19.69 (4.89)

No Highlight 29.64 (5.46) 18.56 (5.34)

*Note.* All study/test items were unrelated.

Table 3

*Mean (± 95% CI) JOL Ratings and Correct Recall Percentages for the Control, Highlight, and Font Size Groups in Experiment 3.*

Group JOL Rating Correct Recall %

Control Group 23.14 (3.56) 21.10 (3.56)

Sans Forgetica Group

Sans Forgetica Font 29.25 (4.59) 19.42 (5.31)

Standard Font 31.73 (4.64) 24.17 (5.51)

*Note.* All study/test items were unrelated.

**Appendix**

Table A1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 |

*Mean Associative Strength Summary Statistics Forward, Backward, and Symmetrical Pairs in Experiment 1A and 1B .*

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

Table A2

*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* |
| Forward | Cue | Concreteness | 4.97 | 1.22 |
|  |  | Length | 6.20 | 1.86 |
|  |  | Frequency | 3.74 | 0.67 |
|  | Target | Concreteness | 4.96 | 1.14 |
|  |  | Length | 4.46 | 1.27 |
|  |  | Frequency | 2.49 | 0.63 |
| Backward | Cue | Concreteness | 4.96 | 1.14 |
|  |  | Length | 4.46 | 1.27 |
|  |  | Frequency | 2.49 | 0.63 |
|  | Target | Concreteness | 4.97 | 1.22 |
|  |  | Length | 6.20 | 1.86 |
|  |  | Frequency | 3.74 | 0.67 |
| Symmetrical | Cue | Concreteness | 4.93 | 1.36 |
|  |  | Length | 5.05 | 1.62 |
|  |  | Frequency | 3.27 | 0.61 |
|  | Target | Concreteness | 4.44 | 1.37 |
|  |  | Length | 5.38 | 2.23 |
|  |  | Frequency | 3.18 | 0.73 |
| Unrelated | Cue | Concreteness | 4.59 | 1.40 |
|  |  | Length | 5.13 | 1.56 |
|  |  | Frequency | 3.20 | 0.80 |
|  | Target | Concreteness | 4.67 | 1.15 |
|  |  | Length | 5.30 | 1.49 |
|  |  | Frequency | 3.18 | 0.90 |

*Notes.* 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 A3

*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* |
| Cue | Concreteness | 4.55 | 1.24 |
|  | Length | 5.16 | 1.50 |
|  | Frequency | 3.04 | 0.84 |
| Target | Concreteness | 4.20 | 1.42 |
|  | Length | 5.10 | 1.36 |
|  | Frequency | 3.13 | 0.76 |

*Notes.* 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).