The Deceptive Nature of Associative Word Pairs:

Effects of Associative Direction on Judgments of Learning

Nicholas P. Maxwell & Mark J. Huff

The University of Southern Mississippi

Author Note

Correspondence concerning this article should be addressed to Nicholas P. Maxwell, 118 College Dr, Hattiesburg, MS, 39406. E-mail: nicholas.maxwell@usm.edu. Code used for data screening and analyses as well as all applicable stimuli and data files have been made available on our OSF page (https://osf.io/hvdma/). All code is embedded inline within the manuscript in an R markdown document written with the *papaja* package (Aust & Barth, 2018).

Abstract

The accuracy of judgments of learning (JOLs) in forecasting later recall of cue-target pairs is sensitive to the associative direction. JOLs are generally well-calibrated for forward associative pairs (e.g., credit-card), but often overestimate later recall accuracy for backward pairs (e.g., card-credit). The present study further examines the effect of associative direction on JOL accuracy by comparing forward and backward pairs to symmetrical associates (e.g., salt-pepper), and unrelated pairs. The correspondence between initial JOLs and recall accuracy was examined when study was either self-paced (Experiment 1), when study and JOLs were made under a 5 s deadline (Experiment 2), or when JOLs were made after a delay (Experiment 3). Across experiments, JOLs accurately estimated correct recall for forward pairs, but overestimated recall for symmetrical, backward, and unrelated pairs—an overestimation that was particularly robust for backward pairs. Calibration plots depicting JOL ratings against their corresponding recall accuracy indicated overestimations occurred for all pair types, though overestimations only occurred at high JOL ratings for symmetrical and forward pairs, a qualitative difference that was not captured in standard analyses of mean JOL and recall rates.

Word Count: 182

Keywords: Judgments of Learning; Paired-Associative Learning; Cued-Recall; Calibration Plots; Overestimation

The Deceptive Nature of Associative Word Pairs:
Effects of Associative Direction on Judgments of Learning

Metacognitive judgments are important for successful learning. At study, individuals must accurately monitor their own ability to learn new information so they can modify study strategies to maximize retention (Nelson & Narens, 1990). A common method for gauging metacognitive judgments is through the Judgment of Learning (JOL) paradigm, in which individuals estimate their likelihood of accurately retrieving a target word when prompted by a cue word on a later test (e.g., 100% = definitely would remember; 0% = definitely would not remember). While JOL ratings tend to be relatively accurate, certain factors have been shown to produce inconsistencies between predicted and actual performance. For instance, JOL accuracy has been shown to be sensitive towards perceptual information such as font size (Rhodes & Castel, 2008), the presence versus absence of retrieval practice (Miller & Geraci, 2014), and importantly, the associative direction of cue-target pairs (e.g., root-plant vs. plant-root) and their magnitude of association (Koriat & Bjork, 2005). Our study contributes to this area by further examining the relationship between JOLs and cued-recall accuracy by directly comparing four different types of word pairs (forward, backward, symmetrical, or unrelated). Further, we compare these pairs under self-paced study and JOL ratings, when study/JOLs are timed, and when JOLs are delayed as a means to improve JOL accuracy.

Interest in the correspondence between memory predictions at study and later recall accuracy for word pairs is not new. In an early demonstration, Arbuckle and Cuddy (1969) reported a relationship between word-pair association and recall performance in which participants generally perceived strong (vs. weak) associates as being more easily remembered. More recently, Koriat and Bjork (2005) showed that the direction of association in cue-target

pairs can similarly affect the correspondence between memory predictions using JOLs and later recall performance. In particular, two directions of associations were suggested to affect the correspondence between JOLs and later recall: A priori and a posteriori. A priori associations correspond to associations in the forward direction (e.g., door-open) and refer to the likelihood that a cue word will elicit a target word. A posteriori associations refer to the perceived association between a cue and target that is only apparent when both are viewed simultaneously. A posteriori associations include weakly associated pairs (e.g., door-stop) and strongly associated pairs, but in the backward direction (e.g., knob-door; see too Koriat, 1981). Koriat and Bjork reported that initial JOL ratings were generally well predictive of later recall but showed an *illusion of competence* on a posteriori pairs in which initial JOLs often exceeded later recall rates. Additional experiments indicated that the illusion of competence on a posteriori pairs was likely dependent upon the direction of the association rather than the strength of association, as JOLs were well-calibrated to later recall for weakly related pairs—a pattern replicated by other researchers (Castel, McCabe, & Roediger, 2007). The illusion of competence is consistent with Koriat's (1997) cue-utilization model in which intrinsic, extrinsic, and mnemonic cues that facilitate ease of processing (including associative relations between the cue and target in a posteriori pairs) can affect JOL accuracy (Dunlosky & Matvey, 2001; Tiede & Leboe, 2009).

When examining the role of cue-target associations, differences in direction and magnitude are often indexed through free-association norms. Such norms are collected using a free-association task in which participants are shown a cue word and are instructed to respond with the first word that comes to mind. From these norms, the probability of responding to word A with word B (forward-associative strength, FAS) can then be computed as an approximate measure of the forward associative overlap shared between pairs. Similarly, backward-

associative strength (BAS), or the probability of responding to word B with A in an A-B pair, can be also computed (see Nelson, McEvoy, & Dennis, 2000). Thus, norms are useful for assessing the associative strength and direction of cue-target pairs when evaluating their effects on JOLs and recall accuracy.

Using the Nelson, McEvoy, and Schrieber (2004) free association norms, Castel et al. (2007) further evaluated the correspondence between JOLs and recall accuracy by comparing pairs that were strongly or weakly related in the forward direction or were unrelated. Additionally, their study also contained a set of identical pairs to evaluate the effects of pair similarity given evidence that identical word pairs are poorly remembered (e.g., Tulving, 1974). The authors reasoned that for identical pairs, participants may rely upon item similarity given the items are perceptually and semantically identical when making JOL ratings. As a result, JOL ratings for identical pairs would be high though their recall would be low, producing an illusion of competence. Indeed, an illusion of competence was found for identical pairs (but not for weakly and strongly related pairs in the forward direction) and this pattern was found both when study duration was limited to 4 seconds and when study was self-paced. The illusion of competence pattern found for identical pairs is particularly intriguing given identical pairs provided participants with a cue word that was perfectly predictive of the target. Nevertheless, recall rates were worse than those for strongly related forward pairs, contributing to the illusion of competence.

Although previous work has demonstrated that semantic relations can induce an illusion of competence for identical pairs, an important question is whether the illusion of competence for identical pairs was due to identical pairs being a perfect match perceptually and semantically, or because identical pairs are *symmetrical* in association. Symmetrical in this case refers to cue-

target pairs that are strongly associated to each other in both directions (e.g., On-Off) according to word association norms. Based on Koriat (1997), we suggest that participants prioritize semantic relatedness over consideration of retrieval conditions when providing JOLs at study. Therefore, strongly associated pairs would likely encourage JOL ratings that would exceed later recall when the retrieval target was ambiguous such as backwards pairs. A similar pattern would likely also emerge for symmetrical pairs given the cue word does not directly converge upon on obvious target (as in forward pairs). Therefore, evaluating the relationship between JOLs and later recall for symmetrical pairs is important for determining how association affects JOLs outside of when cue-target pairs are perceptually and semantically identical.

In the present study, we further evaluated the illusion of competence between JOLs and recall accuracy by examining the direction of association. Specifically, we examined differences in JOL ratings and recall performance (and the calibration between the two) for forward, backward, and symmetrical paired associates relative to unrelated word pairs. To date, no study has investigated illusions of competence for JOLs on symmetrical associates. Unlike the identical word pairs used by Castel et al. (2007), symmetrical associates contain forward and backward associations that are equivalent, without having to rely on the use of repeated words. Because symmetrical pairs are semantically related, they can be directly compared to forward and backward associates when using word association norms. This allowed us to equate study lists on levels of FAS and BAS as we could then match pairs within each list on levels of associative overlap norms across pair direction. Doing so allowed us to control for the effects of association strength on recall, as cued-recall performance has been shown to improve as the associative strength between item pairs increases (Nelson, et al., 2004).

Finally, to further investigate the relationship between JOL estimations and later cued-recall across experiments, we constructed calibration plots in which JOLs ratings are plotted against their corresponding recall accuracy (Nelson & Dunlosky, 1991; see too Roediger, Wixted, & Desoto, 2012; Sauer, Brewer, Zweck, & Weber, 2010 for use of calibration plots with confidence ratings). The use of these plots provides the advantage of pinpointing the level of JOL rating at which the illusion of competence emerges (i.e., low vs. high JOL ratings) and in doing so, more accurately characterizes the effects of associative direction on JOLs.

Comparisons of different associative pairs and their respective calibration plots were conducted across three experiments to test the reliability of the illusion of competence. We first evaluate the correspondence between JOLs and recall accuracy when study and JOL ratings for word pairs were self-paced (Experiment 1), when study and JOL ratings were restricted through the use of a study deadline designed to reduce excessively long study durations thereby equating study/JOL ratings across pair types (Experiment 2), and when JOL ratings were delayed and were provided after study as a means to increase JOL accuracy (Experiment 3; see Dunlosky & Nelson, 1992; Rhodes & Tauber, 2011). To preview, illusions of competence were found for backward, symmetrical, and unrelated pairs, but not for forward pairs, and these patterns were found consistently across experiments. Thus, the effects of associative direction on JOLs and recall appear to be consistent across different methodologies.

Experiment 1: Self-Paced Study Instructions

In Experiment 1, we followed a similar design to Koriat and Bjork (2005) and Castel et al. (2007) to evaluate the effects of associative direction on JOL ratings and recall. The goal of this experiment was to replicate illusion of competence findings for backward associates and compare this pattern to forward and symmetrical associates and unrelated pairs. We expected

that an illusion of competence would be related to the effectiveness of the cue word to elicit the target word. For backward pairs, although the cue and target word are ostensibly related, the associative direction between the items makes the cue a poor predictor of the target word and therefore, we expected a robust illusion of competence. We similarly expected that unrelated pairs would show an illusion of competence given that these cues are also a poor predictor of an unrelated target. We note, however, that Castel et al. reported that JOL ratings accurately predicted later cued-recall rates on unrelated pairs, which suggests that participants are perhaps better able to adjust their JOL ratings in response to pairs that have no association. Our experiment provides another test of this pair type. We further expected that an illusion of competence would emerge for symmetrical pairs, though to a lesser degree, as symmetrical pairs have an association in the forward direction. Finally, we expected that JOLs would be well calibrated to later recall for forward pairs, as the cue would be an accurate predictor of the target.

Methods

Participants

Thirty-one University of Southern Mississippi undergraduates participated in this study for partial course credit. As described below, three participants were excluded for failure to report 10% or more of their JOL responses, leaving 28 participants available for analysis. All participants were native English speakers and had normal or corrected-to-normal color 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 small effect size (Cohen's d = 0.27) or larger.

Materials

One-hundred-eighty associative word pairs were taken from the University of South Florida Free Association Norms (Nelson et al., 2004). These pairs consisted of 40 asymmetric forward pairs in which association only occurred in the forward direction (e.g., bounce-ball), 40 asymmetric backward pairs in which association only occurred in the backward direction (e.g., ball-bounce), 40 symmetric pairs in which forward and backward strength were equivalent (e.g., on-off), 40 unrelated pairs (e.g., building-cat), and 20 non-tested buffer items to control for primacy and recency effects. Pairs were equally distributed across two study lists, each consisting of 20 symmetrical, forward, backward, and unrelated paired associates and 10 buffer items. All participants were presented with both study lists which were organized into two study-test blocks, the order of which was counterbalanced across participants. Both study lists were organized such that five buffer words were presented at the beginning and end of each list, with the remaining pairs randomized anew for each participant. Thus, each study block contained 90 pairs (80 tested, 10 buffer). Additionally, pair types were equated on associative strength (i.e., FAS and BAS) using the Nelson et al. (2004) free association norms and lexical and semantic properties including word length, SUBTLEX frequency (Brysbaert & New, 2009), and concreteness values derived from the English Lexicon Project (Balota et al., 2007). Associative strength, semantic, and lexical properties of the pair types are reported in the Appendix (Tables A1-A2). Furthermore, all study blocks were matched on these properties so that mean associative overlap and lexical/semantic properties were equivalent between direction types and across study lists. For all pair types, counterbalanced versions of the study lists were created that switched the order of the word pairs (i.e., forest-tree vs. tree-forest). This allowed for greater control of item differences, particularly on forward and backward pairs, as the same items were used in both the

forward and backward directions across counterbalances. Pair order was similarly flipped and counterbalanced across the unrelated and symmetrical pairs.

The cued-recall test in each block consisted of all 80 cues from the original study items (minus buffers). The cue was presented next to a blank space that was to be completed with the studied target word. The order of test items was randomized anew for each participant.

Procedure

All participants were tested individually via computers running *E-Prime 3* software (Psychology Software Tools, Pittsburgh, PA). Following informed consent, participants were instructed that they would view a series of cue-target word pairs in which the cue was always presented on the left and the target on the right and that their memory for the target word would be tested. In addition to studying the word pairs, participants were further instructed to provide a JOL rating for each pair. Specifically, they were instructed to rate the likelihood that they would be able to remember the target word which was presented with the cue word at test using a 0 to 100 scale. They were informed that a rating of 0 indicated that they would be unable to correctly remember the target word, while a response of 100 indicated that they were certain they would correctly remember target word. Participants were encouraged to use the full range of the scale when making their judgments to limit anchoring on scale extremes (i.e., judgments of 0 and 100). Following instruction, participants were presented with the first study list. The study phase was self-paced with participants viewing an item pair and typing a JOL rating before proceeding to the next pair. Participants made their JOL ratings while the pair was displayed.

Following the first study list, participants completed an arithmetic filler task for two minutes which was immediately followed by a cued-recall test in which participants were presented with the first word from each study pair and asked to type the target word from

memory. If participants were unable to retrieve the target word, they could skip to the next test cue by pressing the enter key. After completing the first cued-recall test, participants began the second study/test block which used the same instructions as the first block. After completion of the second study/test block, participants were fully debriefed. Each experimental session lasted approximately 30 minutes.

Results

Prior to conducting analyses, all data was screened for missing responses and outliers (i.e., JOLs that were outside of the 0-100 range) which were subsequently removed. For participants with fewer than 10% missing JOL responses, these missing responses were imputed in *R* using the *mice* package (Van Buuren & Groothuis-Oudshoorn, 2011). Data imputation was used to minimize the total number of JOL trials excluded in the analyses[1]. Three participants were missing greater than 10% of their total JOL responses and were removed, leaving 28 participants for analysis. For these remaining participants, less than 1% of their total JOL trials were imputed, which were randomly distributed across different pair types. Recall was scored such that missing recall responses were scored as incorrect (likely skipped by participants), but misspellings of correct items were counted as correct.

A $p < .05$ significance level was used for all analyses unless noted otherwise. Partial-eta squared (η_p^2) and Cohen's *d* effect size indices were included for significant Analyses of Variance (ANOVAs) and *t*-tests, respectively. Figure 1 (top panel) plots mean JOL ratings and cued-recall rates for each word pair type. For completeness, individual comparisons of JOLs and correct recall proportions across pair types (including effect size estimates) are reported in the Appendix for all experiments (Table A3).

A 2 (Measure: JOL vs. Recall) × 4 (Pair Type: Forward vs. Backward vs. Symmetrical vs. Unrelated) within-subject ANOVA was conducted to test for differences between mean JOL ratings and recall rates across the four associative pair types. A significant effect of measure was found, $F(1, 27) = 21.49$, $MSE = 616.80$, $\eta_p^2 = .24$, which indicated that across pair types, JOL ratings exceeded later recall rates (57.97 vs. 42.59). An effect of pair type was also found, $F(3, 81) = 266.52$, $MSE = 108.88$ $\eta_p^2 = .67$, in which JOL ratings/recall rates were greatest for symmetrical pairs (67.80), followed by forward pairs (65.24), backward pairs (49.79), and unrelated pairs (18.29). All comparisons across pair types were statistically different, ts $\geq$ 13.89, ds $\geq$ 1.97, except for symmetrical and forward pairs, which was marginal, $t(27) = 1.87$, $SEM = 1.44$, $p = .07$, $d = 0.28$. Critically, a significant interaction was also found, $F(3, 81) = 29.41$, $MSE = 81.89$, $\eta_p^2 = .14$. A series of follow up t-tests confirmed a robust illusion of competence for backward pairs in which JOLs exceeded later recall accuracy (66.09 vs. 33.48), $t(27) = 8.74$, $SEM = 4.67$, $d = 2.17$. An illusion of competence was also found for symmetrical pairs (71.64 vs. 58.84), $t(27) = 3.04$, $SEM = 4.42$, $d = 0.41$, and unrelated pairs (25.96 vs. 10.63), $t(27) = 4.86$, $SEM = 3.31$, $d = 0.90$, though at a lesser magnitude. For forward pairs however, JOL ratings did not differ from later recall (68.21 vs. 67.41), $t < 1$.

We next assessed the correspondence between the JOLs provided at study and correct recall for each of the pair types using a series of calibration plots. In these plots, JOLs were first rounded to the nearest 10% increment which were then plotted against the proportion of correct recall for items that were rated at that increment. For instance, the 0% JOL increment contains the proportion of correct recall for items given an initial judgment of 0%, the 10% increment contains the proportion of correct recall for items given an initial judgment of 10%, and so on.

Calibration plots for each of the four pair types are reported in Figure 2. Each plot includes a calibration line which reflects perfect correspondence between JOL ratings and correct recall (e.g., 30% JOL and 30% correct recall). Overestimations (i.e., data points that fall below the calibration line) were found to emerge at different JOL ratings for each pair type. For unrelated pairs, JOL overestimations occurred across nearly all JOL ratings (JOLs > 20%), however overestimations emerged later for associative pairs. For backward pairs, overestimations occurred at JOLs greater than 60%, for symmetrical pairs, overestimations occurred at JOLs greater than 80%, and for forward pairs, overestimations were only found at the highest JOL ratings (90-100%). These patterns were confirmed by effects of Pair Type, $F(3, 81) = 71.70$, $MSE = 1471.60$, $\eta_p^2 = .73$, JOL Increment, $F(10, 270) = 6.35$, $MSE = 1204.60$, $\eta_p^2 = .19$, and a significant interaction, $F(30, 810) = 1.80$, $MSE = 879.71$, $\eta_p^2 = .06$. Thus, evidence for illusions of competence were found for each pair type, however overestimations only emerged at the highest JOL ratings when the cue was most predictive of the target in the forward direction.

Discussion

Experiment 1 investigated the influence of the directional association of cue-target pairs on JOLs and cued recall. Our results replicated illusion of competence patterns reported by Koriat and Bjork (2005) in which JOLs were inflated for backward, but not forward, associates. Our study eliminated any potential item effects between these two pair types as backward and forward pairs were the same pairs just presented in different orderings and counterbalanced across participants. Of importance, our experiment also found an illusion of competence pattern for symmetrical and unrelated pairs. Symmetrical pairs, in which pairs had similar association in forward and backward directions, were of particular interest in our study given Castel et al. (2007) who showed a similar overestimation pattern using identical cue-target pairs. The pattern

found for our symmetrical pairs suggests that symmetrical associates can similarly produce an illusion of competence even when the pairs are not identical.

We further analyzed the correspondence of JOLs and recall accuracy by plotting measures relative to a calibration line. Our analyses found JOL overestimations tended to occur for associative pairs only when recall was relatively high, but for unrelated pairs, overestimations occurred across recall rates save for the lowest JOL ratings. The calibration plots revealed that illusions of competence were present for all pair types, though there were qualitative differences in the JOL ratings in which these overestimations emerged. These plots are therefore important as they can reveal important differences in the correspondence between JOLs and later recall that are not available if one only compares means collapsed across JOL ratings.

JOL ratings provided in Experiment 1 were made when study was self-paced, which may have affected the relative calibration between JOL ratings and later recall as individuals may have varied the amount of time allocated to encoding and providing a JOL based on their perceived difficulty of each pair. Hertzog, Dixon, Hultsch, and MacDonald (2003) reported that individuals spend more time studying pairs that are perceived to be more difficult to remember. Indeed, an analysis of encoding durations revealed significant differences across pairs, $F(3, 81) = 22.69$, $MSE = 200094$, $\eta_p^2 = .35$, with encoding duration slowest for backward pairs (4749 ms; $SD = 594$ ms) followed by forward pairs (4700 ms; $SD = 415$ ms), symmetrical pairs (4270 ms; $SD = 393$ ms), and unrelated pairs (3925 ms; $SD = 448$ ms). All pairs differed from each other statistically, ts > 3.67, ds > 1.06, with the exception of forward and backward pairs which were equivalent, $t < 1$. Given the recall rates reported above which may be an indicator of pair difficulty, we would have expected that participants would have spent more time studying unrelated pairs relative to the other pair types, though this was not the case. Instead, since more

time was allocated towards study of the backward and forward pairs, it is possible that participants noticed the asymmetrical association and placed additional efforts towards studying the pair. If so, the relative magnitude of the illusion of competence may have been moderated by study duration which could have affected recall rates (and JOL ratings). Thus, to control for differences in study durations, Experiment 2 evaluated whether restricting the amount of time participants studied each word pair and provided a JOL, affected illusion of competence patterns.

Experiment 2: Response Deadline at Study

In Experiment 2, we tested whether calibration patterns between JOLs and later recall found in Experiment 1 would hold when a deadline was used to restrict the amount of time participants studied each pair and provided a JOL. We expected that JOL overestimations would increase relative to Experiment 1, as participants would not be able to increase their encoding duration in response to pairs perceived as being more difficult to improve later recall.

Methods

Participants and Stimuli

Thirty-four University of Southern Mississippi undergraduates participated for partial course credit. All participants were native English speakers and had normal or corrected-to-normal color vision. Data screening followed the same procedure outlined in Experiment 1, and less than 1% of the trial level data was imputed across associative direction groups. Two participants were found to be missing more than 10% of their total JOL trials and were removed from the final dataset, resulting in 32 participants available for the analyses.

Procedure

The same procedure from Experiment 1 was followed with the exception that during the study phase, participants were required to study the word pair and make a JOL response within 5

s. This deadline was based upon mean study durations found in Experiment 1 averaged across pair types (4411 ms) plus approximately 500 ms to allow for a small buffer to ensure that participants would be able to study and provide JOL ratings for all pairs. Thus, this time window was expected to provide participants with adequate time to both study the word pair and provide their JOL, while preventing excessively long study durations. If a JOL rating was not made by the deadline, the computer automatically advanced to the next pair, and the experimenter would politely remind the participant to make their JOL responses within the time period. The word pair and the JOL rating box were presented simultaneously on the computer screen. Participants were provided with instruction regarding the deadline prior to study and completed a practice list to familiarize themselves with the procedure.

Results

Figure 1 (second panel) plots mean JOL ratings and cued-recall rates as a function of word pair type for Experiment 2. Using the same ANOVA as Experiment 1, an effect of measure was detected, $F(1, 31) = 17.99$, $MSE = 772.82$, $\eta_p^2 = .13$, indicating that JOL rates were greater than subsequent recall rates (52.83 vs. 41.91). An effect of pair type was also found, $F(3, 93) = 233.47$, $MSE = 105.03$, $\eta_p^2 = .63$, indicating that JOL/recall rates were greatest for symmetrical pairs (63.24), followed by forward pairs (63.19), backward pairs (45.40), and unrelated pairs (17.64). Differences were significant across all comparisons, ts ≥ 11.21, ds ≥ 1.39, except for symmetrical and forward pairs, $t < 1$. A significant interaction was also found, $F(3, 93) = 56.41$, $MSE = 74.91$, $\eta_p^2 = .14$, which indicated a significant illusion of competence pattern for backward pairs as JOLs were greater than later recall (59.08 vs. 31.72), $t(31) = 9.06$, $SEM = 3.14$, $d = 1.71$, for symmetrical pairs (67.18 vs. 59.30), $t(31) = 2.74$, $SEM = 3.00$, $d = 0.54$, and for unrelated pairs (23.97 vs. 11.33), $t(31) = 4.26$, $SEM = 3.09$, $d = 1.20$, though again, the latter two

pair types were at a lower magnitude. Again, for forward pairs, JOL ratings were equivalent to later recall (61.07 vs. 65.31), $t(31) = 1.39$, $SEM = 3.18$, $p = .18$.

We again constructed calibration plots to evaluate recall rates at 10% JOL increments (Figure 3). Consistent with Experiment 1, overestimations emerged at different JOL ratings for each pair type. Overestimations were found for unrelated and backward pairs at relatively low JOL ratings (> 20% and 40%, respectively), but at a higher JOL rating for symmetrical pairs (> 70%) and only at the highest JOL rating (100%) for forward pairs. These patterns were confirmed by effects of pair type, $F(3, 93) = 95.86$, $MSE = 1365.79$, $\eta_p^2 = .76$, JOL increment, $F(10, 310) = 5.57$, $MSE = 1321.93$, $\eta_p^2 = .15$, and a significant interaction, $F(30, 930) = 2.98$, $MSE = 793.78$, $\eta_p^2 = .09$.

Discussion

The results of Experiment 2 largely followed Experiment 1 in which initial JOL ratings exceeded later recall for backward, symmetrical, and unrelated pairs—a pattern that was particularly robust for backward pairs. For forward pairs, in which the cue word is strongly predictive of the target, JOLs closely approximated later recall rates, indicating that participants were well calibrated for these pairs. Calibration plots also yielded similar patterns to Experiment 1 in which overestimations emerged at early JOL ratings for unrelated pairs, at higher ratings for backward and symmetrical pairs, and only at the highest recall rates for forward pairs. Thus, in contrast to our expectation, the study/rating deadline produced the same illusion of competence pattern as Experiment 1.

Although study deadlines restricted the maximum amount of time for participants to study the pair and provide a JOL rating, they only ensured that participants responded before a deadline, meaning that participants still may have still encoded pairs at different rates. An

analysis of encoding durations indicated that study/rating durations were equivalent across the four pair types, $F < 1$. Thus, whether participants are given self-paced study or are required to study pairs within a 5 s deadline, there are no differences in the correspondence between JOL ratings and later recall, a pattern that was similarly reported by Castel et al. (2007).

Since self-paced versus restricted encoding durations do not appear to affect the illusion of competence, we next evaluated whether the illusion would hold using the delayed JOL task—a method that has been shown to elicit more accurate JOL ratings (Nelson & Dunlosky, 1991). The delayed JOL task requires that participants provide JOL ratings when word pairs are removed from view and are not readily available. Dunlosky and Nelson (1992) proposed that immediate JOLs are less accurate due to noise from short-term memory that is present at encoding but absent at recall. Therefore, removing pairs from participants while JOL ratings are made may encourage participants to process pairs as if they were retrieving them, thereby increasing JOL accuracy. Rhodes and Tauber (2011) confirmed this pattern in a meta-analysis, showing that JOLs made after a delay are consistently more accurate and even provide a small boost to recall performance versus immediate JOLs. Thus, delayed JOLs may be more accurate as the judgment is being elicited in the absence of the studied information, mimicking conditions at recall.

Experiment 3: Delayed Judgments of Learning

In Experiment 3, we tested whether a delayed JOL manipulation would reduce the illusion of competence by producing JOL ratings that more accurately reflect performance at test. Based on Rhodes and Tauber (2011) we expected that delayed JOLs will enhance JOL accuracy, reducing the illusion of competence. Given the illusion of competence was robust for backward pairs, we anticipated that the illusion would be reduced, rather than eliminated. We similarly constructed calibration plots to qualitatively evaluate JOLs as a function of recall accuracy.

Methods

Participants

Thirty-three University of Southern Mississippi undergraduates completed the study for partial course credit. Data screening followed the same procedure used in Experiment 1, and no participants were eliminated. All participants were native English speakers with normal or corrected-to-normal color vision.

Materials and Procedure

All materials and procedure in Experiment 3 were identical to that of Experiment 1 (including self-paced study) with one exception. Specifically, participants viewed a single word pair for each study trial but pressed a key on the keyboard which advanced them to a new screen which removed the word pair and provided a dialog box to enter their JOL response. Thus, participants made JOLs after study when the word pair was no longer available.

Results

Data screening and imputation followed the same procedure used in Experiment 1, which removed two participants. Overall, data were imputed for less than 1% of trials, which were randomly distributed across associative direction conditions. Figure 1 (third panel) reports mean JOLs and percent correct recall as a function of pair direction.

JOL ratings were again found to exceed later recall rates (62.32 vs. 43.88), $F(1, 32) = 29.04$, $MSE = 423.65$, $\eta_p^2 = .28$. Additionally, JOLs/recall rates were also found to differ across pair types, $F(3, 96) = 282.36$, $MSE = 123.96$, $\eta_p^2 = .60$. Overall, JOL ratings/recall rates were greatest for forward pairs (69.54), followed by symmetrical pairs (67.14), backward pairs (52.30), and unrelated pairs (23.43). Post-hoc tests indicated that comparisons across all pair

types differed significantly, ts ≥ 9.85, ds ≥ 1.35, with the exception of forward and symmetrical pairs, which was marginal, $t(32) = 1.92$, $SEM = 1.29$, $p = .06$, $d = 0.18$.

A significant interaction was again found, $F(3, 96) = 40.15$, $MSE = 48.44$, $\eta_p^2 = .13$, and follow up tests indicated a similar pattern of overestimation as Experiments 1 and 2. Overall, the illusion of competence was greatest for backward pairs as JOLs greatly exceeded later recall accuracy, (70.50 vs. 34.09), $t(32) = 9.28$, $SEM = 4.08$, $d = 2.87$, and similar patterns of overestimation were observed for symmetrical pairs (74.29 vs. 60.00), $t(32) = 3.39$, $SEM = 4.38$, $d = 0.92$, and unrelated pairs (32.93 vs. 13.94), $t(32) = 4.80$, $SEM = 4.12$, $d = 1.24$, but again, JOL ratings and recall rates were equivalent on forward pairs (71.58 vs. 67.50), $t(32) = 1.19$, $SEM = 3.55$, $p = .24$.

Calibration plots (Figure 4) showed JOLs following similar overestimation patterns as Experiment 1 in which JOL overestimations emerged at low JOL rates for unrelated pairs (20%) and at higher rates for backward (50%) and symmetrical pairs (80%). Overestimations were again found on forward pairs, but only at the highest JOL ratings (90-100%). These patterns were confirmed by effects of pair type, $F(3, 96) = 63.41$, $MSE = 1243.58$, $\eta_p^2 = .73$, JOL increment, $F(10, 320) = 7.96$, $MSE = 1297.96$, $\eta_p^2 = .20$, and a significant interaction between the two, $F(30, 960) = 2.15$, $MSE = 849.07$, $\eta_p^2 = .06$.

Discussion

The results of Experiment 3 were largely consistent with Experiments 1 and 2. Overestimation occurred most often for backward and unrelated word pairs and these overestimations occurred across nearly all JOL ratings as shown in the calibration plots. For both conditions, correct recall never surpassed 50% at any JOL level, even for JOL ratings of 50% or greater.

By requiring participants to delay their JOL ratings, we reasoned that the calibration between these initial judgments and later recall would improve. This pattern was not in evidence. As we mention in the General Discussion, there are some differences between our delayed JOL procedure and others used in the literature (cf. Rhodes & Tauber, 2011) which may explain why our delayed JOL procedure did facilitate JOL calibration. If anything, overall JOL ratings actually increased when collapsed across pair types relative to the standard procedure in Experiment 1 (61.42 vs. 57.97), though this finding was not significant, $t(59) = 1.26$, $SEM = 2.79$, and there was no improvement in overall recall rates between the two experiments (43.88 vs. 42.59), $t < 1$. Thus, the delayed JOL procedure did not provide any benefits to JOL accuracy relative to standard study instructions.

Cross Experimental Analyses

Prior to discussing our experiments in more detail, we report cross-experimental analyses of JOLs/recall rates and calibration plots. To determine whether correspondence between JOLs and recall accuracy differed across experiments, we conducted a 2 (Measure: JOL vs. Recall) × 4 (Pair Type: Forward vs. Backward vs. Symmetrical vs. Unrelated) × 3 (Experiment 1-3) mixed ANOVA. No significant interactions with Experiment were found, including the three-way interaction (largest $F(2, 90) = 1.26$, $MSE = 567.37$, $p = .29$) indicating similar patterns across experiments despite different methodologies.

Given these similarities, we pooled our experiments to more powerfully evaluate the effects of pair type on JOLs and recall accuracy. Figure 1 (bottom panel) reports pooled mean JOLs and recall rates for each pair type. Starting with mean JOLs/recall, JOLs were found to be greater overall relative to recall, $F(1, 92) = 69.69$, $MSE = 570.51$, $\eta_p^2 = .73$, and JOLs/recall rates were found to differ across the different pair types, $F(3, 276) = 63.41$, $MSE = 113.02$, $\eta_p^2 = .63$.

Importantly, an interaction was found, $F(3, 276) = 122.90$, $MSE = 61.92$, $\eta_p^2 = .13$, which indicated a robust illusion of competence for backward pairs (64.44 vs. 33.18), $t(92) = 15.46$, $SEM = 2.05$, $d = 2.02$, with a smaller illusion for symmetrical (71.41 vs. 58.76), $t(92) = 6.26$, $SEM = 2.05$, $d = 0.85$, and unrelated pairs (26.63 vs. 12.09), $t(92) = 7.24$, $SEM = 2.02$, $d = 1.08$, No evidence for the illusion of competence was found for forward pairs (66.56 vs. 66.42), $t < 1$.

We similarly conducted a cross-experimental analysis on calibration plots and similarly found no significant interactions with experiment (largest $F(6, 162) = 1.40$, $MSE = 768.63$, $p = .22$). A pooled analysis (Figure 5), showed that JOL overestimations emerged at low JOL rates for unrelated pairs (20%), and at higher rates for backward (40%), and symmetrical pairs (70%), but again, only emerged for backward pairs at very high JOL ratings (90-100%). These patterns were confirmed by effects of pair type, $F(3, 81) = 8.70$, $MSE = 667.38$, $\eta_p^2 = .24$, JOL increment, $F(10, 270) = 76.87$, $MSE = 1207.88$, $\eta_p^2 = .74$, and a significant interaction between the two, $F(30, 810) = 10.72$, $MSE = 986.88$, $\eta_p^2 = .28$.

General Discussion

The primary goal of this study was to further examine JOL overestimations on word pairs with different associative directions including symmetrical associates in which forward and backward strength are equivalent. Across three experiments, we found that backward, symmetrical, and unrelated pairs produced an illusion of competence in which JOL ratings exceeded later recall rates. This illusion was particularly robust for backward pairs in which the backward direction made recall of target items particularly difficult. In fact, our cross-experimental data showed that on average, JOLs for backward pairs exceeded recall rates by 34%. For symmetrical associates and unrelated pairs, this illusion was much more modest (9% and 15%, respectively), demonstrating that backward pairs are highly deceptive. For forward

pairs, in which the target was highly predictive from the cue at test, participants were well-calibrated across experiments.

Calibration plots were constructed to provide a more fine-grained examination of the correspondence between JOLs and recall by examining recall rates at each 10% JOL increment relative to a calibration line. These calibration plots indicated that all pair types showed an illusion of competence at some JOL level, however, unrelated and backward pairs which had the lowest recall rates showed an overconfidence for most JOL ratings, whereas forward and symmetrical pairs only showed overconfidence in the highest JOL ratings. This pattern indicates that even when cues are highly predictive of a later target, as in forward pairs, an illusion of competence is likely to emerge for high JOLs. Inclusion of calibration plots is particularly important given "traditional" analyses on mean JOL/recall rates indicated that forward pairs showed no illusion of competence. Thus, the calibration plots indicate that the illusion can be moderated by JOL level.

Experiment 2 further examined JOL accuracy when encoding and JOL rating duration were limited to 5 s versus self-paced encoding. We reasoned that the self-paced encoding in Experiment 1 may have encouraged participants to slow their responses when presented with word pairs that they perceive as difficult to remember. While it was expected that restricting encoding time would likely inflate the illusion of competence given participants would not be able to adjust their encoding durations (thereby reducing correct recall), recall rates were similar between the experiments and the illusion of competence patterns persisted. Together, these experiments are consistent with Castel et al. (2007) who also showed similar JOL/recall patterns on associative pairs when comparing self-paced and timed study durations.

In an attempt to improve JOL accuracy, Experiment 3 utilized a delayed JOL manipulation in which JOLs were provided in the absence of the studied word pair. Contrary to our expectations however, delayed JOLs were ineffective at reducing JOL overestimations and in fact, were actually greater relative to non-delayed JOLs used in Experiments 1 and 2 (see Figure 1 for comparisons). These inflated JOLs may be linked to the amount of time between when the stimuli pair is presented for study and when participants are asked to provide their JOL rating for the pair. In our delayed task, participants were presented with the cue-target pair and were then moved to a new screen where they were asked to provide a JOL rating for the pair that they had just studied. While the JOL ratings in our delayed manipulation were still elicited in the absence of the studied information, the short delay between study and rating may not provide sufficient time for a delayed JOL effect to arise. Thus, the delayed JOL effect may only be effective for increasing JOL accuracy if ratings are solicited after a substantial delay (e.g., Rhodes & Tauber, 2011) Indeed, previous work by Nelson and Dunlosky (1991) has shown this to be the case. By using mixed lists of immediate and delayed JOLs that were structured to have multiple immediate JOL trials spaced between the presentation of a delayed JOL study pair and its corresponding judgment prompt, they showed that delayed JOLs were more accurate relative to those made immediately after study. Thus, delayed manipulations are only effective at increasing JOL accuracy when ample time is provided between study and rating.

Although our delayed JOL manipulation did not enhance JOL accuracy, our experiments importantly build upon existing work on JOLs and associative pairs (e.g., Koriat & Bjork, 2005; Castel et al., 2007) through other means. For instance, our experiments directly compared forward, backward, symmetrical, and unrelated pairs, to more thoroughly catalogue JOL estimations. To this end, we were careful to control for potential item effects when constructing

word pairs that could potentially affect either JOL ratings or recall accuracy. Specifically, associated pairs were all matched in associative strength and forward and backward pairs were created by simply flipping the pair order across counterbalances, making them perfect controls for each other. We were also careful to match all pairs on the basis of frequency, word length, and concreteness. Based on these efforts, we can have greater confidence that the effects reported are due to differences in associative direction and not item differences.

Despite the reliability of the data patterns reported across the experiments, we note two departures from the literature that are worthy of discussion. First, while our experiments showed that participants were generally well calibrated for forward pairs, Koriat and Bjork (2005) and Castel et al. (2007) showed that recall rates for forward pairs exceeded initial JOLs. Second, Castel et al., showed that JOLs were well calibrated overall for unrelated pairs whereas we consistently found an illusion of competence pattern. We ascribe these differences between studies to either (1) differences in lexical/semantic characteristics across pair types that were not controlled for in previous studies, or, (2) that there were considerable differences in the number of pairs that participants were presented with at study, thus affecting recall rates, which we believe is a more likely possibility. For instance, across Koriat and Bjork's experiments, participants studied between 24-72 pairs and in Castel et al. participants studied 48 pairs. However, in our experiments, participants studied a total of 180 items split between two blocks, a larger number of pairs which could have negatively impacted correct recall by increasing interference. Indeed, correct recall rates tended to be 15-25% lower in our experiments relative to these previous studies, though the JOL rates were relatively consistent. This latter possibility is interesting because it suggests that methods that affect recall rates may be important for whether an illusion of competence is found or not. Methods to enhance memory for the target item such

as the use of deep levels-of-processing tasks at study may be more effective at improving the calibration between JOLs and recall by improving recall to match typical JOL overestimations. An encoding task could possibly be paired with a set of instructions designed to encourage participants to temper their JOLs. Indeed, Koriat and Bjork (2006) have shown some success at improving JOL accuracy with such instructions.

Our preceding discussion on methods to improve JOL accuracy therefore leads us to the question: What drives JOL overestimations in the first place? According to the cue-utilization framework set forth by Koriat (1997), metacognitive judgments are based on three domains: The readily observable characteristics of the items to be studied (i.e., intrinsic cues such as the characteristics of the studied pairs, such as item difficulty, associative strength, etc.), manipulations undertaken at the time of encoding (i.e., extrinsic cues such as stimulus duration, study strategy, etc.), and mnemonic cues that inform participants of how well they have learned a given item and to what extent they will be required to remember the item later. Koriat (1997) showed that both intrinsic and extrinsic factors influenced JOL strengths, though only intrinsic factors were shown to influence both JOLs and recall rates equally. As briefly reviewed above, the cue-utilization framework then suggests that JOL overestimation should arise when participants are basing their JOL ratings on extrinsic cues, as these are cues are more likely to disproportionately affect recall rates. However, the present study shows that the direction of association (which by nature is an intrinsic cue) is powerful enough to induce an overconfidence bias. Specifically, the direction of the association may disrupt the mnemonic cues that inform participants of how well they are learning the studied information (i.e., participants may perceive pairs as being more related and thus less difficult to recall) and the conditions in which they will need to retrieve the information. As participants appear to focus primarily on the semantic

relatedness of paired items when making their JOLs, pairs where the retrieval conditions are less certain (such as symmetrical pairs) or unusual (e.g., backward pairs) may result in instances where JOL ratings consistently surpass recall rates.

Alternatively, the robustness of the illusion of competence may be explained by comparing JOLs to the related Judgment of Associative Memory (JAM; see Maki, 2007a, for a review). In a JAM task, individuals are presented with paired associates and are asked to rate the associative strength of the pair (i.e., how many individuals out of 100 would respond to the cue word with the presented target), mimicking the free association process used to create associative overlap norms. JAM ratings are also prone to overestimation, and previous research (Maki, 2007a, 2007b; Valentine & Buchanan 2013) has shown that individuals typically perform poorly on such tasks. Maki (2007a) proposed that this increase in JAM ratings for forward pairs resulted from the presented target activating items related to the cue that tend to be activated less often when only the cue item is shown (Koriat & Bjork, 2006). This may extend to judgments of learning: If individuals have inflated notions of how related paired items are, they may display a tendency to inflate JOLs. However, this explanation seems unlikely, as this study showed that the illusion of competence replicates even after we controlled for the effects of association strength by equating all study lists on FAS and BAS and by having the forward and backward pairs be comprised of the same individual items. Thus, we conclude that the direction of the association is the primary factor driving the illusion of competence.

Conclusion

The present study provides a critical examination of how the associative direction of cue target pairs affects the calibration between JOL ratings and recall. Our data provide further evidence for the illusion of competence first described by Koriat and Bjork (2005) and show that

it extends beyond backward associates and identical item pairs (Castel et al., 2007). Calibration plots allowed us to determine the point at which JOLs became overestimated for each of the pair types. These plots revealed an important finding in that JOL overestimations occurred across pair types, but forward and symmetrical pair types were only overestimated at the highest JOL ratings. Collectively, our experiments provide greater understanding of how associative direction influences metacognitive judgment making and can be informative for developing methods to reduce such metacognitive illusions.

References

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology*, *81* (1), 126–131.

Aust, F., & Barth, M. (2018). papaja: Create APA manuscripts with R Markdown. R Package. Retrieved from https://github.com/crsh/papaja.

Balota, D. A., Yap, M. J., Hutchsison, 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. doi:10.3758/BF03194036

Dunlosky, J., & Matvey, G. (2001). Empirical analysis of the intrinsic–extrinsic distinction of judgments of learning (JOLs): Effects of relatedness and serial position on JOLs. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(5), 1180-1191.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, *20*(4), 374–380.

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), 175–91. doi:10.3758/BF03193146

Hertzog, C., Dixon, R. A., Hultsch, D. F., & MacDonald, S. W. S. (2003). Latent change models of adult cognition: Are changes in processing speed and working memory associated with changes in episodic memory? *Psychology and Aging, 18*(4), 755-769.

Koriat, A. (1981). Semantic facilitation in lexical decision as a function of prime-target association. *Memory & Cognition, 9*(6), 587-598.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349-370.

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. doi:10.1037/0278-7393.31.2.187

Koriat, A., & Bjork, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test, *Memory & Cognition, 34* (5), 959–972.

Maki, W. S. (2007a). Judgments of associative memory. *Cognitive Psychology*, *54* (4), 319–353. doi:10.1016/j.cogpsych.2006.08.002

Maki, W. S. (2007b). Separating bias and sensitivity in judgments of associative memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 33*(1), 231–237.

Miller, T. M., & Geraci, L. (2014). Improving metacognitive accuracy: How failing to retrieve practice items reduces overconfidence. *Consciousness and Cognition 29*, 131-140.

Nelson, D. L., McEvoy, C. L., & Dennis, S. (2000). What is free association and what does it measure? *Memory & Cognition*, *28* (6), 887–899. doi:10.3758/BF03209337

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. doi:10.3758/BF03195588

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. In: *The psychology of learning and motivation*, ed. G. Bower. American Psychologist.

Psychology Software Tools, Inc. [E-Prime 3.0]. (2016). Retrieved from https://www.pstnet.com.

Roediger, H. L., Wixted, J. H., & DeSoto, K. A. (2012). The curious complexity between confidence and accuracy in reports from memory. In Nadel L, Sinnott-Armstrong WP (Eds.), *Memory and Law*. Oxford University Press, pp. 84–108.

Rhodes, G. M., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General. 137*(4), 615 – 625.

Rhodes, G. M., & Tauber, S. K. (2011). The Influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin*, *137*, 131-48. 10.1037/a0021705.

Sauer, J., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence –accuracy relationship for eyewitness identification. *Law and Human Behavior, 34*, 337-347.

Tiede, H. L., & Leboe, J. P. (2009). Metamemory judgments and the benefits of repeated study: Improving recall predictions through the activation of appropriate knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(3), 822-828.

Tulving, E. (1974). Cue-dependent forgetting. *American Scientist, 62*(1), 74-82.

Valentine, K. D., & Buchanan, E. M. (2013). JAM-boree: An application of observation oriented modelling to judgements of associative memory. *Journal of Cognitive Psychology, 25*(4), 400–422.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Mulitvariate imputation by chained equations in R. Journal of Statistical Software, 45(3), 1–67. Retrieved from https://www.jstatsoft.org/v45/i03/

Footnotes

[1] Analyses were also conducted on datasets with no imputation and with the imputation done only for participants missing 5% or less of their total JOL responses. Since similar data were found using each imputation method, we report the results using the 10% cutoff criterion which maximized the number of observations available for analyses. Datasets using no imputation and the 5% cutoff criterion are available via our OSF page (https://osf.io/hvdma/).

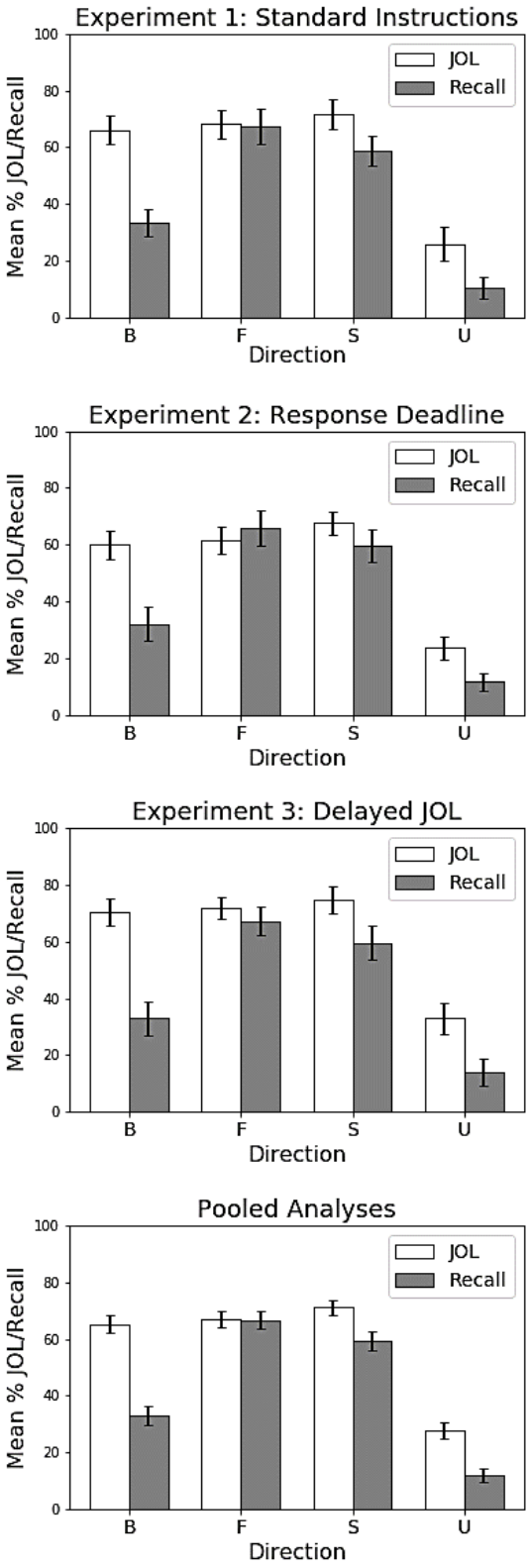

Figure 1. Comparison of mean JOL ratings and recall rates across each of the three experiments. Error bars represent 95% confidence intervals. B = Backward pairs; F = Forward pairs; S = Symmetrical pairs; U = unrelated pairs.

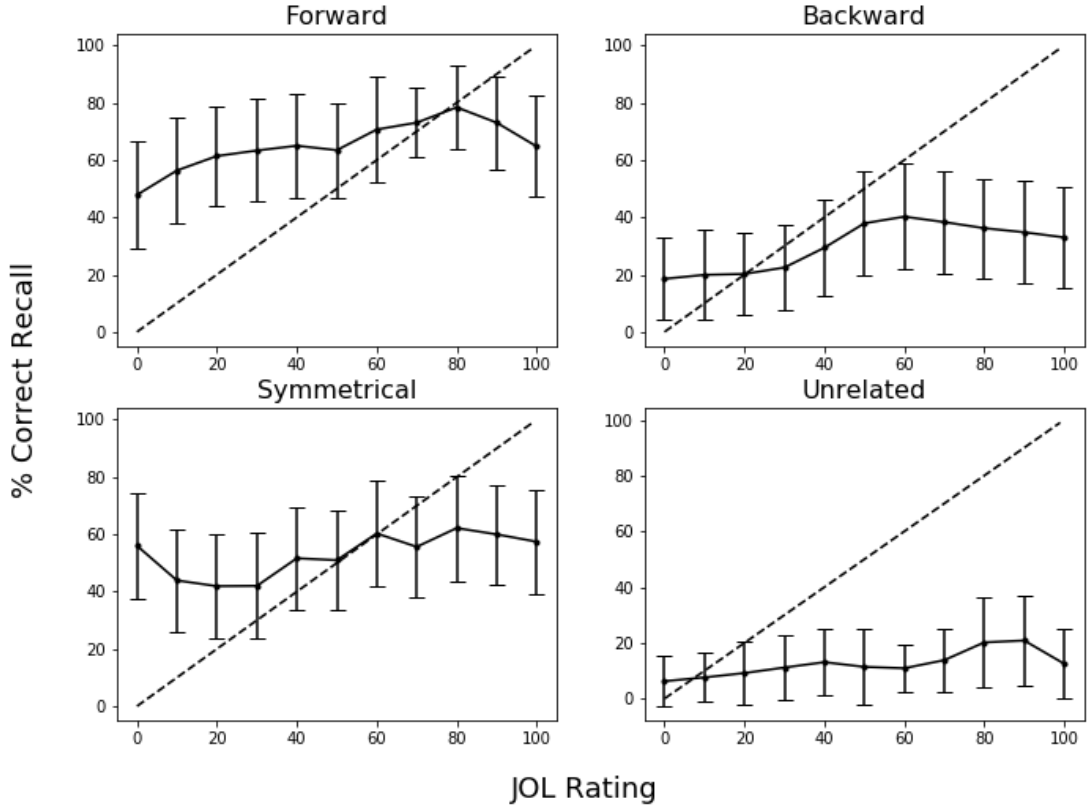

Figure 2. Calibration plots as a function of pair type in Experiment 1. Dashed lines indicate perfect calibration between JOL ratings and proportion of correct cued-recall. Overconfidence is represented by points falling below the calibration line. Data were smoothed over three adjacent JOL ratings. Bars represent 95% confidence interval.

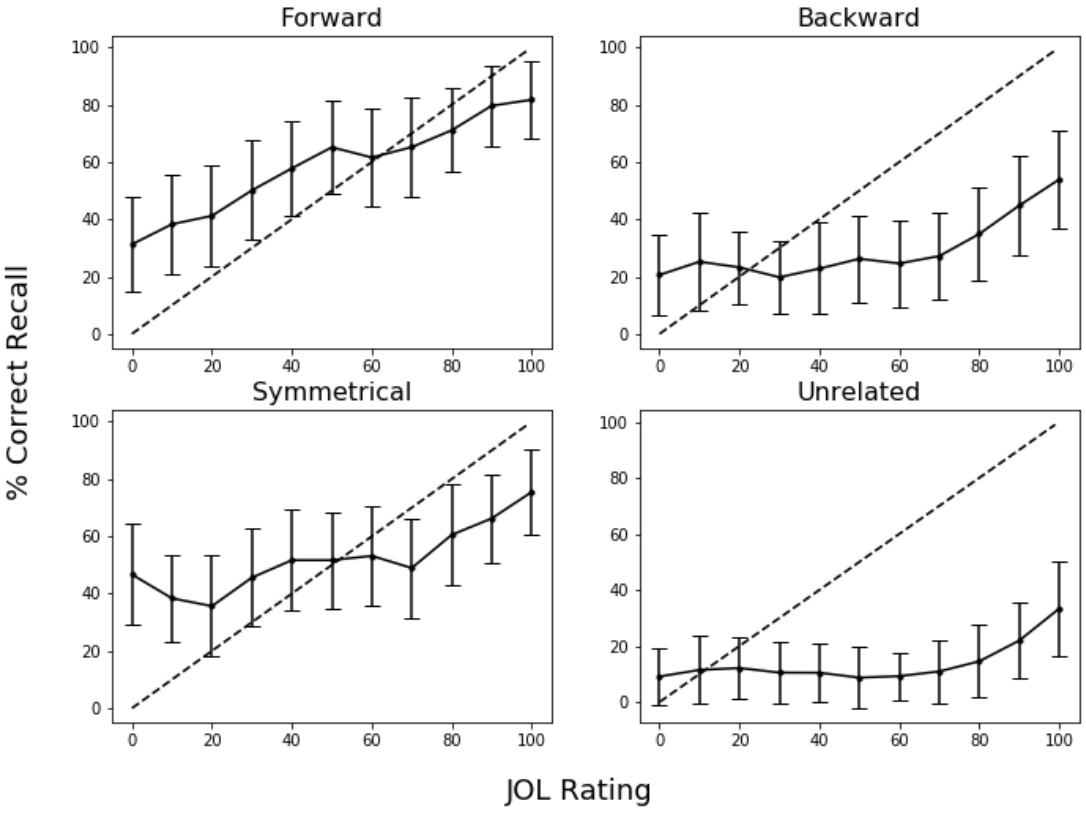

Figure 3. Calibration plots as a function of pair type in Experiment 2. Dashed lines indicate perfect calibration between JOL ratings and proportion of correct cued-recall. Overconfidence is represented by points falling below the calibration line. Data were smoothed over three adjacent JOL ratings. Bars represent 95% confidence interval.

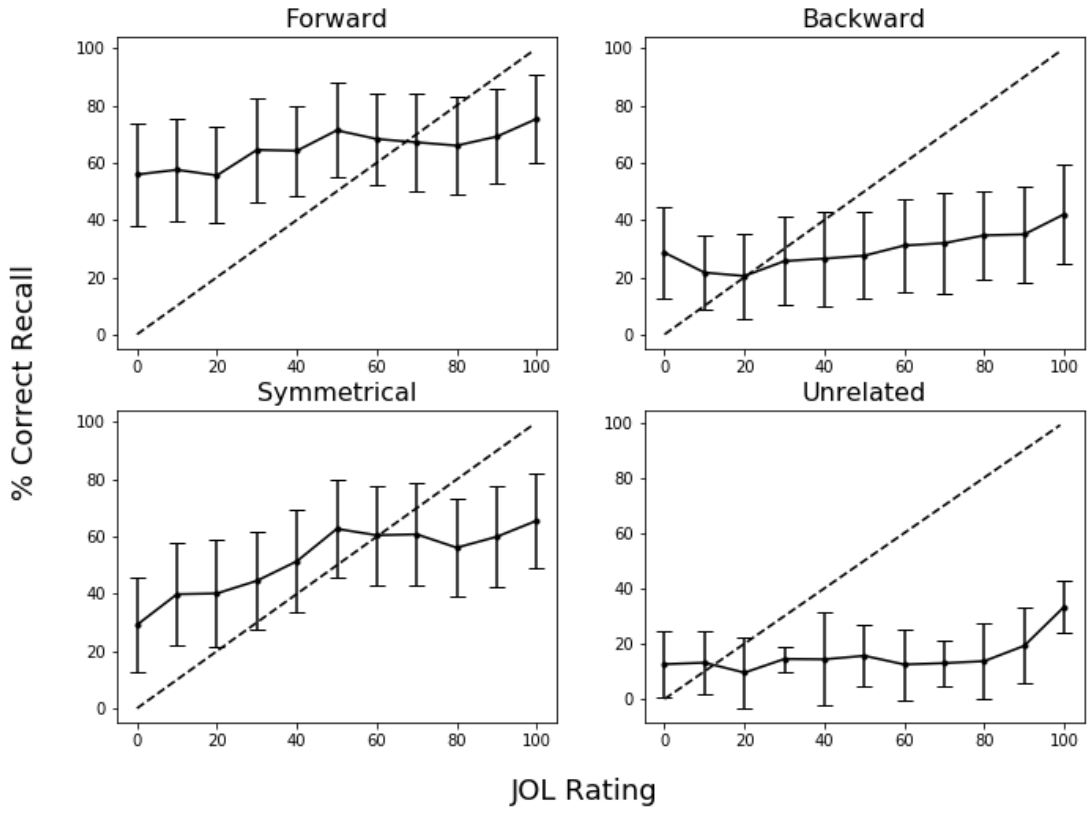

Figure 4. Calibration plots as a function of pair type in Experiment 3. Dashed lines indicate perfect calibration between JOL ratings and proportion of correct cued-recall. Overconfidence is represented by points falling below the calibration line. Data were smoothed over three adjacent JOL ratings. Bars represent 95% confidence interval.

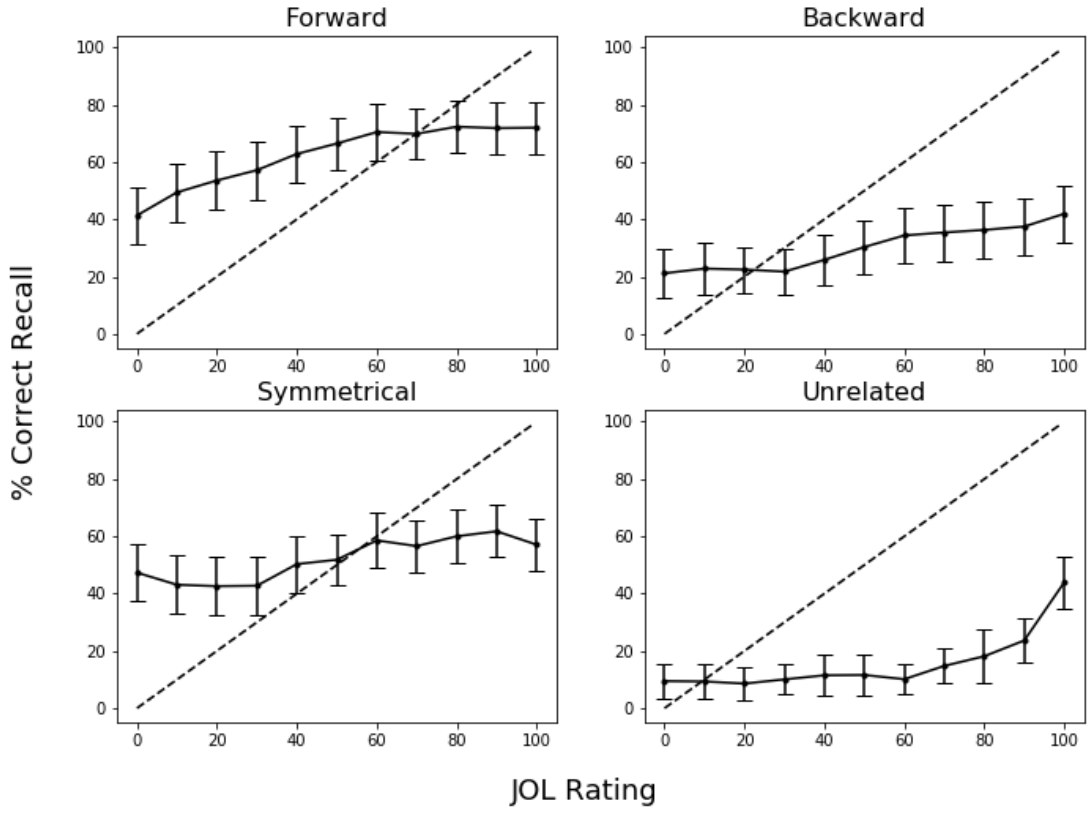

Figure 5. Calibration plots as a function of pair type pooled across Experiments 1-3. Dashed lines indicate perfect calibration between JOL ratings and proportion of correct cued-recall. Overconfidence is represented by points falling below the calibration line. Data were smoothed over three adjacent JOL ratings. Bars represent 95% confidence interval.

Appendix

Table A1

Summary Statistics for associative overlap variables.

Condition	Variable	*M*	*SD*	*Min.*	*Max.*
Forward	FAS	.37	.26	.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

Notes. Values are grouped by JOL condition. FAS and BAS values for unrelated pairs are not included as by definition these pairs have not been normed. Mean FAS and BAS values were computed by taking the average association strength for each pair.

Table A2

Summary statistics for cue and target item properties.

Condition	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/Target	Concreteness	4.70	1.38
		Length	5.21	1.94
		Frequency	3.23	0.67
Unrelated	Cue/Target	Concreteness	4.63	128
		Length	5.21	1.52
		Frequency	2.49	0.85

Notes. Values are grouped by JOL condition. Forward and backward pairs are grouped by position within cue-target pair. Symmetrical and unrelated pairs are averaged across cues and targets, as they did not differ by position within the pairs. 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

Comparison of mean JOL ratings and correct recall percentages across all associative direction groups for each experimental manipulation and pooled analysis.

Experiment	Task	Group	*M*	*95% CI*	F	B	S
Exp. 1	JOL	Forward	68.21	5.03			
		Backward	66.09	4.97	0.16		
		Symmetrical	71.64	5.33	0.24	0.39	
		Unrelated	26.96	5.96	2.84*	2.70*	2.99*
	Recall	Forward	67.41	6.11			
		Backward	33.48	4.80	2.28*		
		Symmetrical	55.84	5.25	0.56	1.87*	
		Unrelated	10.63	3.86	4.12*	1.97*	3.88*
Exp. 2	JOL	Forward	61.53	4.92			
		Backward	59.86	4.90	0.12		
		Symmetrical	67.52	4.02	0.46	0.59	
		Unrelated	23.66	4.08	2.91*	2.77*	3.75*
	Recall	Forward	65.97	6.15			
		Backward	32.02	5.91	1.94*		
		Symmetrical	59.60	5.85	0.37	1.62*	
		Unrelated	11.53	3.20	3.84*	1.49*	3.53*
Exp. 3	JOL	Forward	71.84	3.71			
		Backward	70.46	5.00	0.11		
		Symmetrical	74..63	4.71	0.22	0.29	
		Unrelated	33.10	5.51	2.81*	2.42*	2.76*
	Recall	Forward	67.30	6.00			
		Backward	33.06	4.90	2.14*		
		Symmetrical	59.70	5.80	0.48	1.54*	
		Unrelated	14.03	4.89	3.72*	1.19*	2.89*
Pooled	JOL	Forward	66.44	5.35			
		Backward	66.57	4.96	0.13		
		Symmetrical	71.41	4.69	0.31	0.42	
		Unrelated	26.63	3.99	2.85*	2.63*	3.17*
	Recall	Forward	66.42	6.09			
		Backward	33.17	5.20	2.12*		
		Symmetrical	58.76	5.63	0.47	1.67*	
		Unrelated	12.09	4.87	3.89*	1.55*	3.43*

Note. The three right-most columns indicate Cohen's *d* effect sizes for post-hoc comparisons, * = $p < .05$.