Psychological Research
https://doi.org/10.1007/s00426-020-01342-z

ORIGINAL ARTICLE

The deceptive nature of associative word pairs: the effects of associative direction on judgments of learning

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

Received: 13 December 2019 / Accepted: 10 April 2020

Abstract

The accuracy of judgments of learning (JOLs) in forecasting later recall of cue–target pairs is sensitive to associative direction. JOLs are generally well calibrated for forward associative pairs (e.g., credit-card), but recall accuracy is often overestimated 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 unrelated pairs and symmetrical associates (e.g., salt–pepper)—a novel comparison. The correspondence between initial JOLs and recall accuracy was examined when study was either self-paced with concurrent JOLs (Experiment 1), when study/JOL duration was equated across pair types (Experiment 2), when JOLs were made immediately following study (Experiment 3), and when JOLs were made after a delay (Experiment 4). 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.

Introduction

Metacognitive judgments are important for successful learning. At study, individuals must accurately monitor their own ability to learn new information to modify study strategies and maximize retention (Nelson & Narens, 1990). One method for gauging metacognitive judgments is the judgment of learning (JOL) paradigm, in which individuals estimate their likelihood of accurately retrieving a target word when given a cue word on a later test (e.g., 100% = definitely remember; 0% = definitely not remember). While JOL ratings can be accurate, certain factors have been shown to produce inconsistencies between predicted and actual performance. For instance, JOL accuracy has shown sensitivity 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 and magnitude of cue–target pairs (e.g., root–plant vs. plant–root; 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 following study so as to improve JOL accuracy.

Interest in the correspondence between memory predictions at study and later recall accuracy for word pairs are 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 more easily remembered. More recently, Koriat and Bjork (2005) showed that the associative direction of cue-target pairs can similarly affect the correspondence between JOL memory predictions and later recall. In particular, two directions of associations were suggested to affect the correspondence between JOLs and recall: a priori and a posteriori. A priori associations

R 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).

✉ Nicholas P. Maxwell
nicholas.maxwell@usm.edu

[1] The University of Southern Mississippi, 118 College Dr, Hattiesburg, MS 39406, USA

Published online: 24 April 2020

correspond to forward associations (e.g., door–open) and refer to the likelihood that a cue will elicit a target word. A posteriori associations refer to a perceived association between cue and target that is only apparent when both are presented simultaneously. A posteriori associations include weak associates (e.g., door–stop) and strong associates presented in the backward direction (e.g., knob–door; see too Koriat, 1981). Koriat and Bjork reported that initial JOL ratings were generally predictive of later recall but showed an *illusion of competence* on a posteriori pairs in which JOLs often exceeded later recall rates. Subsequent experiments indicated that the illusion of competence on a posteriori pairs was dependent upon the direction of the association rather than the associative strength as JOLs were well calibrated to recall for weak forward associates—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 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, direction and magnitude are often indexed through free-association norms. Such norms are collected using a free-association task in which participants report the first word that comes to mind in the presence of a cue word. From these norms, the probability of responding to word A with word B (forward-associative strength, FAS) can 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 computed (see Nelson, McEvoy, & Dennis, 2000). Free association norms are useful for assessing the associative strength and direction of cue–target pairs when evaluating their effects on JOLs and recall accuracy. These norms are commonly used across several judgment tasks, including both JOL tasks and the related judgment of associative memory task (JAM; Maki, 2007), which has shown that individuals routinely over-estimate the associative relatedness of paired associates (measured in FAS), especially for weak associates (i.e., when a posteriori relatedness is high but a priori relatedness is low). This has implications for JOL studies, as individuals may perceive paired associates as being more related than their normed strengths imply, and if so, they may assign an inaccurate JOL when asked to make predictions about recall.

Using the Nelson, McEvoy, and Schreiber (2004) free-association norms, Castel et al. (2007) further evaluated the correspondence between JOLs and recall accuracy using strong and weak forward associates or unrelated pairs. Additionally, their study also contained identical pairs to evaluate pair similarity effects given identical cue–target pairs are generally poorly remembered (e.g., Tulving, 1974). The authors reasoned that for identical pairs, participants may rely upon item similarity (vs. cue effectiveness) 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 the forward strong and weak associates) and this pattern was found both when study duration was self-paced and timed. 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 lower than strong associates, contributing to the illusion.

Although prior 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 resulted from a perfect perceptual and semantic match, or because identical pairs are symmetrical associates. 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 the cue effectiveness for recalling the target when providing JOLs. Therefore, strong associates would likely encourage JOL ratings that would exceed later recall when the retrieval target was ambiguous such as backward associates. A similar pattern would likely also emerge for symmetrical associates, given the cue does not directly converge upon an obvious target (as forward associates). Furthermore, the presence of forward and backward associations within symmetrical pairs may make them particularly susceptible to overestimation. Participants may rely on both forward and backward associations when providing JOLs, yet only the forward associations will provide useful retrieval cues at test. Finally, participants may not necessarily interpret symmetrical associates as different than asymmetric associates. For example, the finding that participants provide equivalent JOL ratings for forward and backward pairs in previous studies (e.g., Koriat & Bjork, 2005) suggests that participants do not readily discriminate between different types of associates, despite large differences in recall between pairs. Therefore, evaluating the relationship between JOLs and subsequent recall for symmetrical pairs is important for (1) determining how association affects JOLs outside of when cue–target pairs are perceptually/semantically identical and (2) further assessing the generality of the illusion of competence.

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 associates versus unrelated pairs. To date, no study has investigated the illusion of competence for JOLs on symmetrical associates. Unlike identical pairs used by Castel et al. (2007), symmetrical associates contain equivalent FAS and BAS without word repetitions. Since symmetrical associates are semantically related, they can be compared to forward and backward associates using word association norms. We were, therefore, able to equate associate types on associative strength (FAS and BAS), allowing us to control for these associative variables given cued recall is generally sensitive to associative strength (Nelson, et al., 2004).

Measuring JOL accuracy

Metacognitive studies traditionally distinguish between two types of JOL accuracy: absolute and relative accuracy. Absolute accuracy refers to the overall difference between predicted and actual performance such that the magnitude of an individual's predictions corresponds to their performance at test (Scheck, Meeter, & Nelson, 2004; Connor, Dunlosky, & Hertzog, 1997). For example, absolute accuracy between JOLs and recall would be considered perfect if items given JOL ratings of 50 were subsequently recalled correctly 50% of the time. We refer to this relationship as the calibration between JOLs and recall. Item calibration has been researched extensively within the context of confidence ratings across various research domains, including eyewitness testimony (Tekin and Roediger, 2017; Juslin, Olsson, & Winman, 1996) and facial recognition (Weber & Brewer, 2003).

JOL accuracy has also been assessed in terms of the relative accuracy between predicted and actual performance. For example, if items A and B are studied together, and item A receives a higher JOL than item B, a person's relative accuracy would be perfect if the likelihood of recall at test is greater for item A than item B (i.e., high JOLs are recalled more frequently than low JOLs; Van Overschelde & Nelson, 2006). We refer to this type of accuracy relationship as the resolution between judgments and recall. Though the present study is primarily interested in assessing accuracy in terms of the calibration between JOLs and recall, we report Goodman–Kruskal gamma correlations alongside all calibration plots (Goodman & Kruskal, 1954) as a metric of JOL resolution.

Though calibration and resolution provide unique insights into the relationship between JOLs and recall, for the present study, we primarily focus on JOL calibration to investigate whether the illusion of competence occurs similarly across all levels of JOL ratings (i.e., are participants consistently providing high JOLs to backward pairs even when they are not correctly recalling the pairs at test?). To this end, 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 JOL rating level at which the illusion of competence emerges (i.e., low vs. high JOL ratings) and in doing so, we can more accurately characterize the effects of associative direction on JOLs relative to single JOL scores (e.g., mean calibration scores for absolute accuracy, gamma scores for relative accuracy).

Comparisons of different associative pairs and their respective calibration plots were conducted across four experiments to test the reliability of the illusion of competence. We first evaluated the calibration between JOLs and recall when study and JOL ratings for associates were self-paced and JOLs were made concurrently at the time of study (Experiment 1), when the time provided to study word pairs and provide a JOL rating was held constant across associate types (Experiment 2), when JOLs were elicited immediately following presentation of a cue–target pair (Experiment 3), and when JOL ratings were elicited after a delay as a means of increasing JOL accuracy (Experiment 4; see Dunlosky & Nelson, 1992; Rhodes & Tauber, 2011). To preview, illusions of competence were found for backward, symmetrical, and unrelated pairs, but not for forward associates, 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: concurrent JOLs with self-paced study

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. Our goal was to replicate illusion of competence findings for backward associates and compare this pattern to forward and symmetrical associates and unrelated pairs while also investigating how these four pair types affected the calibration of the JOL/recall relationship. We expected that an illusion of competence would be related to the effectiveness of the cue to elicit the target word. For backward pairs, though the cue and target are ostensibly related, the associative direction between the items makes the cue a poor predictor of the target at test, 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. For backward, unrelated, and symmetrical associates, it is expected that the illusion of competence will result in both poor calibration and poor resolution, with this being most noticeable for backward associates. 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 for partial course credit. Three participants were excluded for failure to report 10% or more of their JOL responses (described below), leaving 28 participants for analysis. All participants were native English speakers with 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 (0.80) to detect a small effect size (Cohen's $d = 0.27$) or larger.

Materials

One hundred and 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 buffers to control for primacy and recency effects. Pairs were equally distributed across two study lists, each consisting of 20 symmetrical, forward, backward, and unrelated pairs and 10 buffers. All stimuli pairs used for each pair type have been made available via our OSF page (https://osf.io/hvdma/).

Participants were presented with both lists which were separated into two study-test blocks, the order of which was counterbalanced across participants. Both study lists were organized such that five buffer pairs 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 from the English Lexicon Project (Balota et al., 2007). Associative strength and semantic/lexical properties of the pair types are reported in the Appendix (Tables 1, 2). 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 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. Test order was randomized anew for each participant.

Procedure

All participants were tested individually via computers running *E-Prime 3* software (Psychology Software Tools, Pittsburgh, PA, USA). 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 pairs, participants were instructed to provide a JOL rating. Specifically, they were told to rate the likelihood that they would be able to remember the target word in the presence of the cue word at test using a 0–100 scale in which 0 indicated that they would be unable to correctly recall the target word, while a response of 100 indicated full certainty that they would recall target word. Participants were encouraged to use the full range of the scale when making their judgments to limit anchoring on 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 provided JOL ratings while the pair was displayed.

Following the first list, participants completed an arithmetic filler task for 2 min followed by a cued-recall test in which participants were presented with the cue 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 min.

Results

Prior to conducting analyses, all data were screened for missing responses and outliers (i.e., JOLs 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 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 skipped recall responses were scored as incorrect, but misspellings of correct items were counted as correct.

A $p < 0.05$ 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. All post hoc comparisons were Bonferroni corrected. Figure 1 (top left) 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 3). Finally, we compute Goodman–Kruskal gamma correlations for all experiments and report them in the figure captions. However, given our interest in JOL calibration, we focus our analyses on absolute accuracy.

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 pair types. A significant effect of measure was found, $F(1, 27) = 21.49$, $MSE = 616.80$, $\eta_p^2 = 0.24$, 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, $F(3, 81) = 266.52$, $MSE = 108.88$ $\eta_p^2 = 0.67$, 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 differed statistically, $ts \geq 13.89$, $ds \geq 1.97$, except for symmetrical and forward pairs, which was marginal, $t(27) = 1.87$, $SEM = 1.44$, $p = 0.07$, $d = 0.28$. Critically, a significant interaction was also found, $F(3, 81) = 29.41$, $MSE = 81.89$, $\eta_p^2 = 0.14$. 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. However, for forward pairs, 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 Fig. 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, $F(3, 81) = 71.70$, $MSE = 1471.60$, $\eta_p^2 = 0.73$, JOL increment, $F(10, 270) = 6.35$, $MSE = 1204.60$, $\eta_p^2 = 0.19$, and a significant interaction, $F(30, 810) = 1.80$, $MSE = 879.71$, $\eta_p^2 = 0.06$. Thus, evidence for illusions of competence was found across pair types however, overestimations only emerged at the highest JOL ratings for forward associates.

[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/).

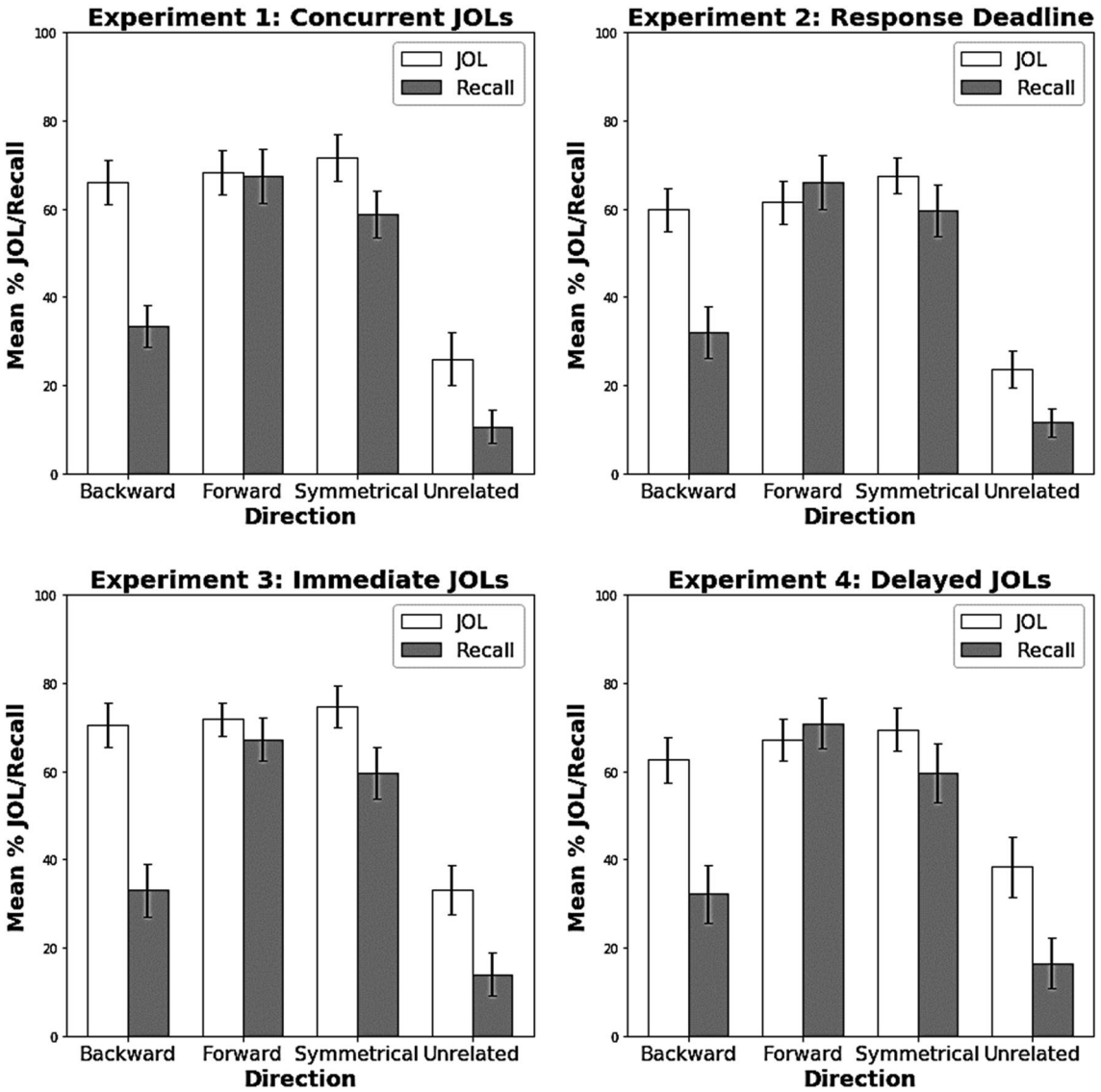

Fig. 1 Comparison of mean JOL ratings and recall rates across each of the four experiments. Error bars represent 95% confidence intervals. *B* backward pairs, *F* forward pairs, *S* symmetrical pairs, *U* unrelated pairs

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 potential item effects between these pair types, as backward and forward pairs were the same pairs in different orderings and counterbalanced across participants. Of importance, our experiment also found an illusion of competence 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 words.

We further analyzed the correspondence of JOLs and recall accuracy by plotting measures relative to a calibration line. Our analyses found that JOL overestimations tended to occur for associative pairs only when recall was relatively high, but for unrelated pairs, overestimations occurred across recall rates except 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.

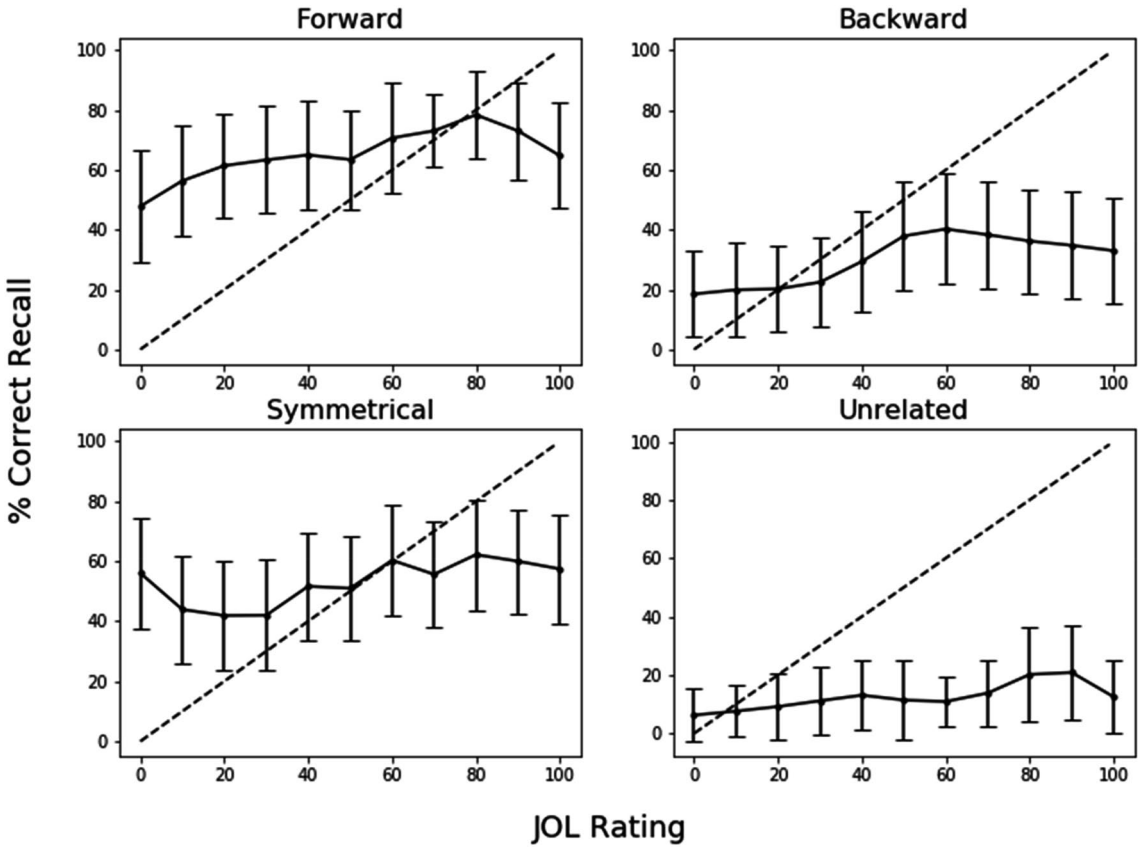

Fig. 2 Calibration plots as a function of pair type in Experiment 1. Dashed lines indicate perfect calibration between JOL ratings and proportion of correct cued recall. Overconfidence is represented by points falling below the calibration line. Data were smoothed over three adjacent JOL ratings. Bars represent 95% confidence interval. Forward pairs: $g = 0.21$, backward pairs: $g = 0.30$, symmetrical pairs: $g = 0.25$, unrelated pairs: $g = -0.34$

These plots are, therefore, important as they can reveal qualitative 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 recall as individuals may have strategically varied the amount of time allocated to encoding 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 as more difficult to remember. Analysis of encoding durations revealed significant differences across pairs, $F(3, 81) = 22.69$, MSE = 200,094, $\eta_p^2 = 0.35$, 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, except forward and backward pairs which were equivalent, $t < 1$. Given the recall rates reported above which may indicate pair difficulty, we would have expected that participants would have spent more time studying unrelated pairs, though this was not in evidence. Instead, more time was allocated towards study of the backward and forward pairs, suggesting that participants noticed the asymmetrical association and placed additional efforts towards studying the pair. If so, the magnitude of the illusion of competence may have been moderated by study duration which could have affected recall rates and/or JOL ratings. To control for this possibility, Experiment 2 equated study duration for all pair types by providing a deadline for participants to study each pair and provide a JOL rating.

Experiment 2: concurrent JOLs with a study deadline

In Experiment 2, we tested whether calibration between JOLs and recall found in Experiment 1 would hold when a study deadline was used to restrict the amount of time participants studied each pair and provided a JOL. We expected JOL overestimations would increase relative to Experiment

1, as participants would not be able to strategically increase 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 were native English speakers and had normal or corrected-to-normal vision. Data screening followed the same procedure in Experiment 1, and less than 1% of the trial level data were imputed across associative direction groups. Two participants were found to be missing greater than 10% of their total JOL trials and were removed, resulting in 32 participants available for the analyses.

Procedure

The same procedure from Experiment 1 was followed, except that during study, 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 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 remind the participant to provide JOL responses within the time period. The 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 (top right) plots mean JOL ratings and cued-recall rates as a function of pair type for Experiment 2. Using the same ANOVA as Experiment 1, an effect of measure was found, $F(1, 31) = 17.99$, MSE = 772.82, $\eta_p^2 = 0.13$, indicating that overall JOL rates exceeded subsequent recall rates (52.83 vs. 41.91). An effect of pair type was also found, $F(3, 93) = 233.47$, MSE = 105.03, $\eta_p^2 = 0.63$, 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, $ts \geq 11.21$, $ds \geq 1.39$, except for symmetrical and forward pairs, $t < 1$. A significant interaction was also found, $F(3, 93) = 56.41$, MSE = 74.91, $\eta_p^2 = 0.14$, which indicated a significant illusion of competence pattern for backward pairs as JOLs exceeded 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. For forward pairs, JOL ratings were equivalent to later recall (61.07 vs. 65.31), $t(31) = 1.39$, SEM = 3.18, $p = 0.18$.

We again constructed calibration plots to evaluate recall rates at 10% JOL increments (Fig. 3). Consistent with Experiment 1, overestimations emerged at different JOL ratings for each pair type. Overestimations were found for unrelated and backward pairs at low JOL ratings (> 20% and 40%, respectively), but at a higher JOL ratings 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, $\eta_p^2 = 0.76$, JOL increment, $F(10, 310) = 5.57$, MSE = 1321.93, $\eta_p^2 = 0.15$, and a significant interaction, $F(30, 930) = 2.98$, MSE = 793.78, $\eta_p^2 = 0.09$.

Discussion

The results of Experiment 2 largely followed Experiment 1: JOL ratings exceeded recall for backward, symmetrical, and unrelated pairs which was particularly robust for backward pairs. For forward associates, JOLs closely approximated later recall rates, indicating that participants were well calibrated. 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 prediction, 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 (cf. Castel et al., 2007).

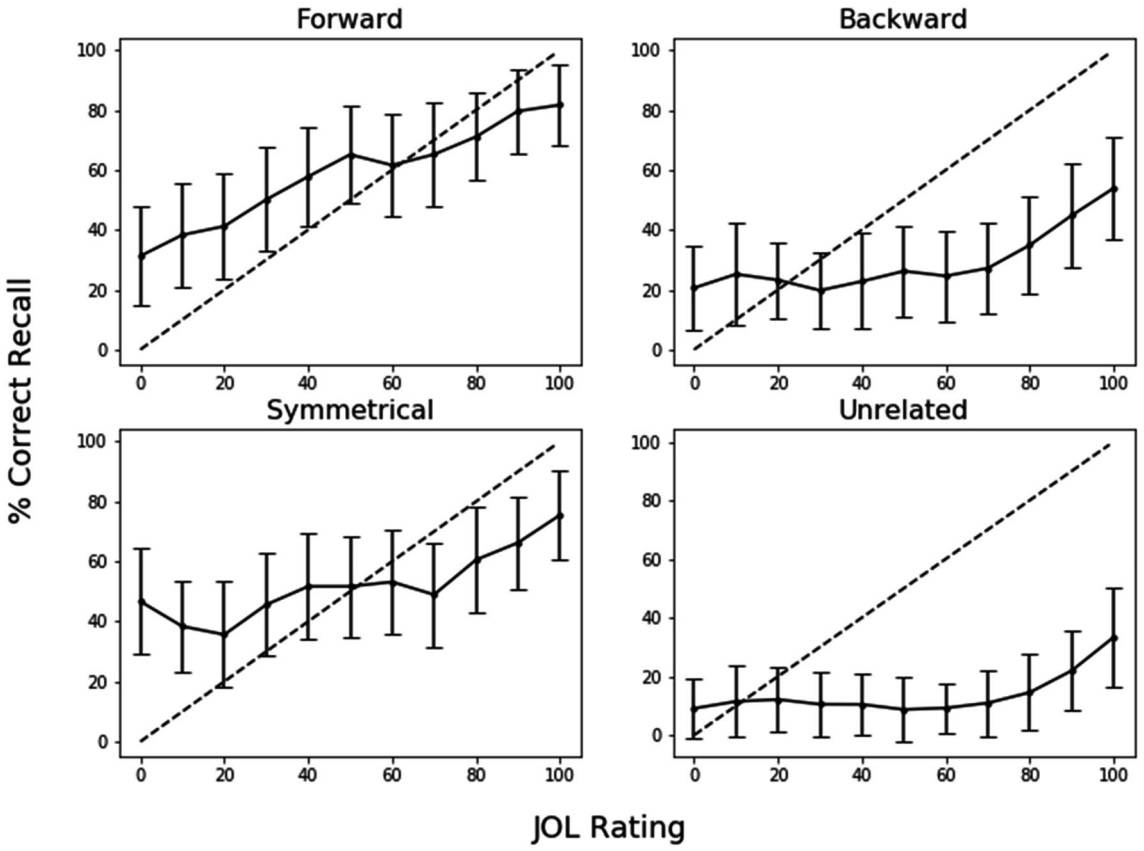

Fig. 3 Calibration plots as a function of pair type in Experiment 2. Dashed lines indicate perfect calibration between JOL ratings and proportion of correct cued recall. Overconfidence is represented by points falling below the calibration line. Data were smoothed over three adjacent JOL ratings. Bars represent 95% confidence interval. Forward pairs: $g = 0.22$, backward pairs: $g = 0.23$, symmetrical pairs: $g = 0.22$, unrelated pairs: $g = -0.32$

Experiment 3: immediate JOLs

Since self-paced versus restricted encoding durations do not appear to affect the illusion of competence, we next evaluated whether the illusion would hold when using an immediate JOL task, in which JOLs were elicited immediately following the removal of the study pair (i.e., pairs were not available for reference when providing a JOL). Since the pair is no longer accessible, participants may be more likely to modulate their JOL ratings, leading to improved calibration. Furthermore, whereas our first two experiments used a concurrent JOL task, Experiment 3 allowed us to directly replicate the JOL task employed by Koriat & Bjork (2005) while extending it to include the symmetrical pairs incorporated in Experiments 1 and 2. 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. Participants were native English speakers with normal or corrected-to-normal 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

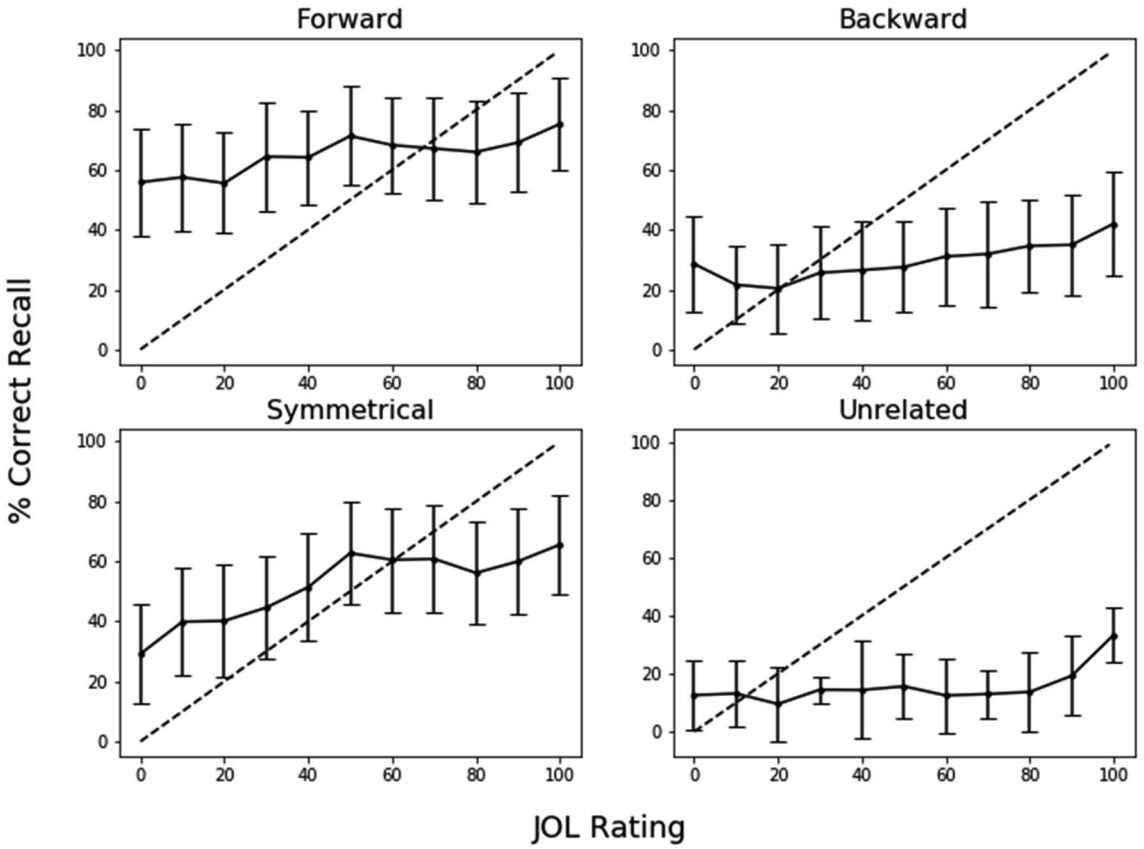

Fig. 4 Calibration plots as a function of pair type in Experiment 3. Dashed lines indicate perfect calibration between JOL ratings and proportion of correct cued recall. Overconfidence is represented by points falling below the calibration line. Data were smoothed over three adjacent JOL ratings. Bars represent 95% confidence interval. Forward pairs: $g = 0.23$, backward pairs: $g = 0.26$, symmetrical pairs: $g = 0.24$, unrelated pairs: $g = -0.09$

keyboard which advanced them to a new screen in which the word pair was removed and replaced with a dialog box to enter their JOL response. Thus, participants made JOLs after study when the pair was no longer available to reference.

Results

Data screening/imputation followed the same procedure as Experiment 1. Data were imputed for less than 1% of trials, which were randomly distributed across associative direction conditions. Figure 1 (bottom left) reports mean JOLs and percent correct recall as a function of pair direction.

Overall JOL ratings were again found to exceed later recall rates (62.32 vs. 43.88), $F(1, 32) = 29.04$, $MSE = 423.65$, $\eta_p^2 = 0.28$. Additionally, JOLs/recall rates were also found to differ across pair types, $F(3, 96) = 282.36$, $MSE = 123.96$, $\eta_p^2 = 0.60$. 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, $ts \geq 9.85$, $ds \geq 1.35$, with the exception of forward and symmetrical pairs, which was marginal, $t(32) = 1.92$, $SEM = 1.29$, $p = 0.06$, $d = 0.18$.

A significant interaction was again found, $F(3, 96) = 40.15$, $MSE = 48.44$, $\eta_p^2 = 0.13$, 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 (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), $t(32) = 1.19$, $SEM = 3.55$, $p = 0.24$.

Calibration plots (Fig. 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, $\eta_p^2 = 0.73$, JOL increment, $F(10, 320) = 7.96$, MSE = 1297.96, $\eta_p^2 = 0.20$, and a significant interaction between the two, $F(30, 960) = 2.15$, MSE = 849.07, $\eta_p^2 = 0.06$.

Discussion

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

By requiring participants to postpone their JOL ratings until completing the study task (vs. concurrently), we reasoned that the calibration between these initial judgments and later recall would improve, as participants would be less prone to overestimation if they were making judgments in the absence the pair. This pattern was not in evidence, as JOLs were similar to those elicited in Experiment 1 (61.42 vs. 57.97) $t(59) = 1.26$, SEM = 2.79, and there was no difference in overall recall rates between experiments (43.88 vs. 42.59), $t < 1$. Thus, the immediate JOL procedure did not provide any benefits to JOL accuracy relative to concurrent JOL study instructions.

Experiment 4: delayed JOLs

Given the results of the previous three experiments, we next assessed whether implementing a delayed JOL task would reduce the illusion of competence. 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. 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. Based on this, we expected that delayed JOLs would enhance the accuracy of JOLs thereby reducing the illusion of competence. Given that the illusion of competence was robust for backward pairs, we anticipated that the illusion would be reduced, rather than eliminated. However, this reduction may not necessarily be reflected in the calibration plots, as research by Van Overschelde and Nelson (2006) has shown that delayed JOLs decrease calibration relative to resolution. Thus, Experiment 4 sought to test whether delayed JOLs would decrease the illusion of competence by either increasing mean recall or decreasing the magnitude of JOLs (or potentially doing both) and whether or not any potential changes would be detected when assessing the calibration between JOLs and recall.

Methods

Participants

Thirty-nine undergraduates were recruited from the University of Southern Mississippi undergraduate research pool and completed the study for partial course credit. Data screening was consistent with the procedure used in Experiment 1, and three participants were eliminated, leading to a total of 36 participants included in the analyses.

Materials and procedure

Materials in Experiment 4 were identical to those in the previous experiments. The procedure closely followed that of Experiment 3 with the following exception. After participants viewed a single cue–target pair, they pressed a key which advanced them to a new screen in which the cue–target pair was removed and participants were asked to solve an arithmetic problem modeled after the OSPAN task (Turner & Engle, 1989). After completing an OSPAN problem, participants were then presented with the cue–item only (e.g., credit–?) and were asked to type their JOL rating into a dialogue box. Thus, all JOL ratings were elicited after a delay but without an intact pair.

Results

Figure 1 (bottom right) displays mean JOLs and percent correct recall for each of the four pair types. Consistent with the previous experiments, the same general patterns of results were found. JOL ratings again exceeded later recall rates (59.79 vs. 44.81), $F(1, 35) = 19.12$, MSE = 800.37, $\eta_p^2 = 0.15$. JOLs/recall rates also differed across pair types, $F(3, 105) = 266.07$, MSE = 97.03, $\eta_p^2 = 0.46$, with JOLs/recall greatest for forward pairs (69.16), followed by symmetrical pairs (64.47), backward pairs (47.31), and unrelated pairs (26.98). Post hoc tests indicated that comparisons across all pair types differed significantly, $ts \geq 4.27$, $ds \geq 0.12$.

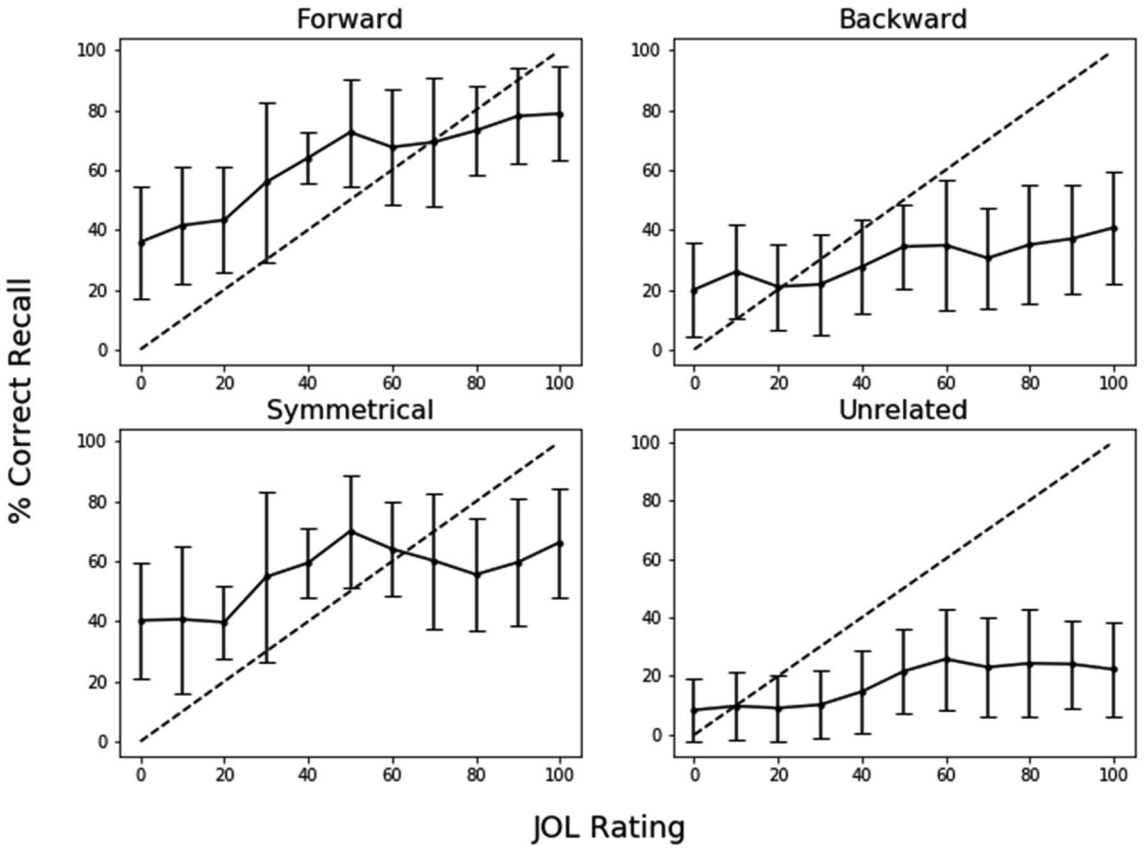

Fig. 5 Calibration plots as a function of pair type in Experiment 4. Dashed lines indicate perfect calibration between JOL ratings and proportion of correct cued recall. Overconfidence is represented by points falling below the calibration line. Data were smoothed over three adjacent JOL ratings. Bars represent 95% confidence interval. Forward pairs: $g = 0.25$, backward pairs: $g = .26$, symmetrical pairs: $g = .25$, unrelated pairs: $g = .03$

A significant interaction was again detected, $F(3, 105) = 77.93$, MSE = 51.04, $\eta_p^2 = 0.12$, and follow-up tests indicated similar patterns of overestimation reported above. Again, the illusion of competence was greatest for backward pairs as JOLs were significantly greater than recall, (62.60 vs. 32.17), $t(35) = 8.61$, SEM = 3.66, $d = 1.69$. This pattern was again found for symmetrical pairs (69.53 vs. 59.73), $t(35) = 2.84$, SEM = 3.57, $d = 0.55$, and unrelated pairs (38.33 vs. 16.48), $t(35) = 4.99$, SEM = 4.54, $d = 1.13$, though JOL ratings and recall rates were equivalent for forward pairs (67.15 vs. 70.91), $t(35) = 1.22$.

Finally, calibration plots (Fig. 5) showed that JOLs followed a similar overestimation pattern to that observed in Experiment 1, with JOL overestimations emerging at low JOL rates for unrelated pairs (20%) and at higher rates for backward (50%) and symmetrical pairs (80%). We again found overestimations of forward pairs, but only for the highest JOL ratings (90–100%). These patterns were once again confirmed by significant effects of pair type, $F(3, 81) = 60.36$, MSE = 1779.92, $\eta_p^2 = 0.53$, JOL increment, $F(10, 270) = 10.92$, MSE = 1338.91, $\eta_p^2 = 0.29$, the interaction between the two, $F(30, 810) = 2.46$, MSE = 919.50, $\eta_p^2 = 0.08$.

Discussion

Findings from Experiment 4 were largely consistent with the previous experiments. Overestimation occurred most frequently across backward and unrelated word pairs, and again, these overestimations occurred for almost all JOL ratings that these two pair types received. Indeed, correct recall of these

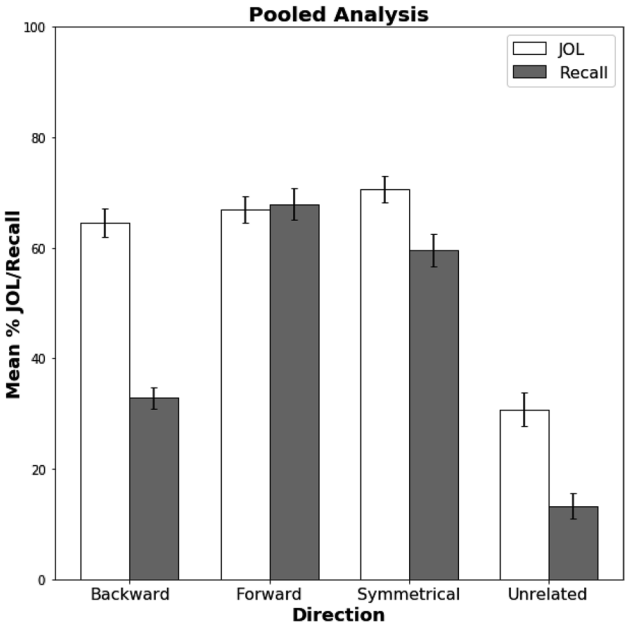

Fig. 6 Comparison of mean JOL ratings and recall rates pooled across each experiment. Error bars represent 95% confidence intervals. *B* backward pairs, *F* forward pairs, *S* symmetrical pairs, *U* unrelated pairs

pairs never surpassed 60% at any JOL level, even when receiving JOL ratings of 60 or higher.

Having participants provide their JOL ratings after completing the math task (vs. concurrently or immediately) was expected to reduce the illusion of competence, as delayed JOLs have been repeatedly shown throughout the literature to improve the relationship between predicted and actual recall performance (Rhodes, 2016). Our findings did not replicate the delayed JOL effect across any of the pair types. We discuss some methodological differences in the General discussion below which may account for this discrepancy, but first turn to a set of cross-experimental analyses to confirm the consistency of our data patterns across experiments.

Cross-experimental analyses

To examine the consistency between JOLs and recall across experiments, we combined the datasets of each of the previous four experiments. While differences across samples may be of concern when conducting cross-experimental analyses, we note that the participants tested in each experiment were recruited from the same participant pool and tested within the same academic year. Thus, participant differences across experiments would likely be minimal.

We first conducted a 2 (measure: JOL vs. recall) × 4 (pair type: forward vs. backward vs. symmetrical vs. unrelated) × 4 (Experiment 1–4) mixed ANOVA. The effect of experiment was marginally significant, $F(3, 125) = 2.18$, MSE = 758.28, $p = 0.09$, $\eta_p^2 = 0.02$, which reflected greater JOL/recall for Experiment 2 compared to Experiment 3 (53.10 vs. 47.38), $t(61) = 2.65$, SEM = 2.21, $d = 0.13$, with all other comparisons unreliable, all ts < 1.86, ps > 0.07, but importantly, the three-way interaction was non-significant, $F(9, 375) = 1.48$, MSE = 63.17, $p = 0.15$. We therefore pooled our data across experiments to examine the combined differences between JOL and recall accuracy as a function of pair type (see Fig. 6).

The pooled analysis indicated that JOLs were greater overall relative to recall (58.21 vs 43.38), $F(1, 128) = 86.08$ MSE = 659.34, $\eta_p^2 = .18$, and JOLs/recall rates differed across pair types, $F(3, 384) = 986.08$, MSE = 114.68, $\eta_p^2 = 0.57$. Importantly, an interaction was found, $F(3, 384) = 185.43$, MSE = 63.88, $\eta_p^2 = 0.12$, which indicated a robust illusion of competence for backward pairs (64.50 vs. 32.84), $t(128) = 17.68$, SEM = 1.81, $d = 1.96$, with a smaller illusion for symmetrical (70.62 vs. 59.50), $t(128) = 6.03$, SEM = 1.86, $d = 0.71$, and unrelated pairs (30.70 vs. 13.28), $t(128) = 4.99$, SEM = 4.54, $d = 1.11$. Of note, even with the additional statistical power from the pooled analysis, no illusion of competence emerged for forward pairs (67.00 vs. 67.89), $t < 1$.

We similarly conducted a cross-experimental analysis on the calibration plots and found no main effect or interactions with experiment (largest $F = 1.07$). A pooled analysis (see Fig. 7) showed that JOL overestimations emerged at low JOL rates for unrelated pairs (20%), at higher rates for backward (40%) and symmetrical pairs (70%), but again, only emerged for forward pairs at the highest JOL ratings (90–100%). These patterns were confirmed by effects of pair type, $F(3, 294) = 150.07$, MSE = 1431.87, $\eta_p^2 = 0.61$, JOL increment, $F(10, 980) = 33.26$, MSE = 1300.05, $\eta_p^2 = 0.25$, and a significant interaction, $F(30, 2940) = 7.51$, MSE = 88.20, $\eta_p^2 = 0.02$. Thus, despite efforts to manipulate the illusion of competence by varying the context in which JOLs were provided, these methods had little effect on JOL calibration, both when average JOL ratings were compared to mean recall accuracy and on calibration plots.

General discussion

The primary goal of our 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

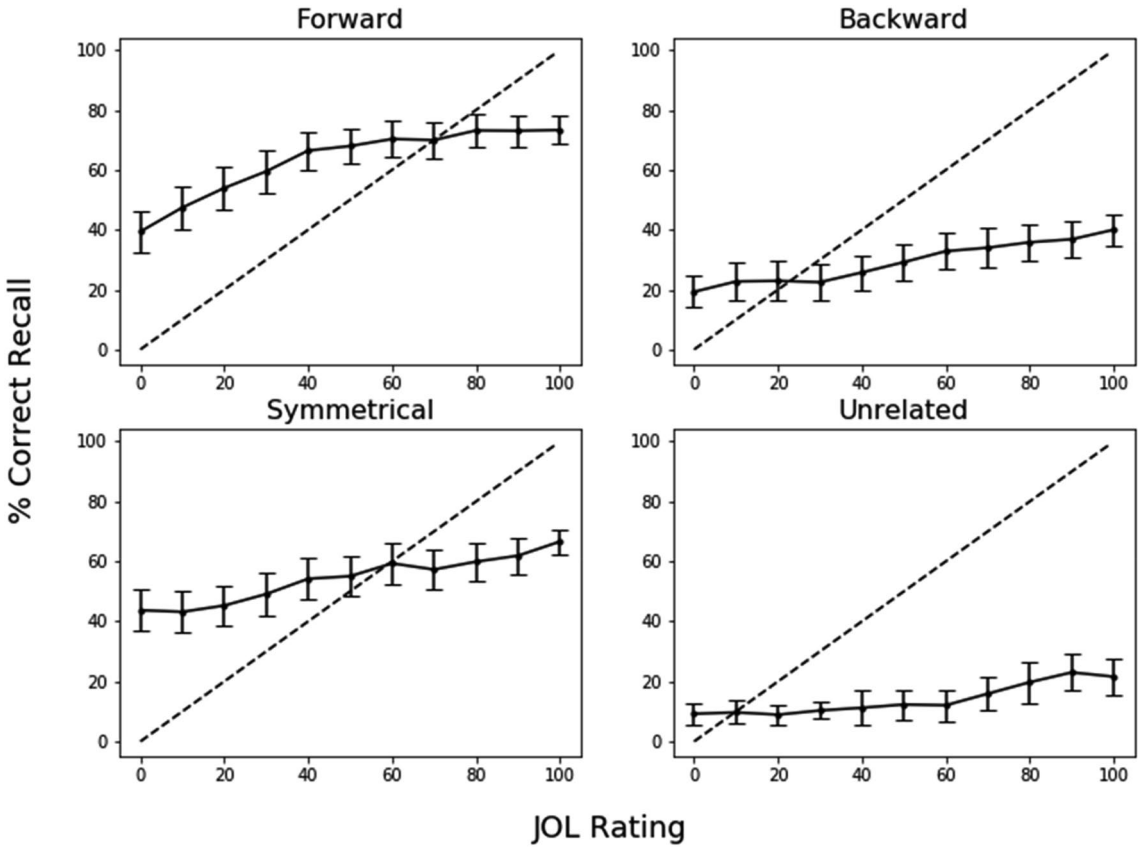

Fig. 7 Calibration plots as a function of pair type pooled across Experiments 1–3. Dashed lines indicate perfect calibration between JOL ratings and proportion of correct cued recall. Overconfidence is represented by points falling below the calibration line. Data were smoothed over three adjacent JOL ratings. Bars represent 95% confidence interval. Forward pairs: $g = 0.23$, backward pairs: $g = 0.26$, symmetrical pairs: $g = 0.24$, unrelated pairs: $g = -0.16$

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. Our cross-experimental data showed that on average, JOLs for backward pairs exceeded recall rates by 32%. For symmetrical associates and unrelated pairs, this illusion was much more modest (11% and 18%, 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 absolute accuracy 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 for 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 can still be detected. The 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 provide additional information regarding the correspondence between JOLs and recall accuracy that is unavailable if mean rates are examined in isolation.

Experiment 2 further examined JOL accuracy when encoding and JOL ratings were restricted to 5 s. We reasoned that self-paced encoding used in Experiment 1 may have encouraged participants to slow their encoding when

presented with word pairs perceived as difficult to remember. While we expected that restricting encoding time would 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 pattern persisted. Together, these experiments are consistent with Castel et al. (2007) who also showed similar JOL/recall patterns when comparing self-paced and timed study durations.

In Experiment 3, an immediate JOL manipulation was used in which study pairs were removed when providing JOLs which occurred immediately following rather than concurrently with study. This manipulation was ineffective at reducing JOL overestimations and replicated prior research (e.g., Koriat & Bjork, 2005).

Finally, Experiment 4 used a delayed JOL manipulation in which participants completed an OSPAN math problem between studying each cue-target pair and typing their JOL. Though we expected the delayed task to reduce the illusion of competence increases overall JOL accuracy and calibration, this was not the case as the same illusion of competence remained. One explanation for this pattern is that our delayed manipulation deviated from the one traditionally used in the literature. Nelson and Dunlosky (1991) used mixed (vs. pure) lists of immediate and delayed JOLs in which the delay consisted of a series of immediate JOL trials interleaved between the study and rating phase of the delayed trial. Under these conditions, delayed (vs. immediate) JOLs were found to be more accurate. Furthermore, participants in our manipulation only completed one problem between study and rating. Thus, the length of our delay may not have been sufficient to increase accuracy (e.g., Rhodes & Tauber, 2011).

For delayed JOLs to be successful at improving accuracy, the task completed between study and judgment must effectively remove the studied pair from working memory such that participants rely on long-term memory when making their JOLs (Dunlosky & Nelson, 1992). Though we reasoned that our OSPAN manipulation would be sufficient to remove the studied pairs from working memory, the illusion of competence persisted.

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) in novel ways. For instance, our experiments directly compared forward, backward, symmetrical, and unrelated pairs, to more thoroughly catalogue JOL estimations. To this end, we controlled for potential item effects when constructing word pairs that could potentially affect either JOL ratings and/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 based on frequency, length, and concreteness. From these efforts, we 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 such that the difference between JOLs and recall was non-significant, Koriat and Bjork (2005) and Castel et al. (2007) showed that recall rates for forward pairs consistently exceeded JOLs. Second, Castel et al., showed that JOLs were well calibrated overall for unrelated pairs, whereas we consistently found an illusion of competence pattern for this pair type. We ascribe these differences between studies to either differences in lexical/semantic characteristics across pair types that were left uncontrolled, or that there were considerable differences in the number of pairs that participants were presented with at study, 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 pairs split between two blocks. The greater number of pairs could have negatively impacted correct recall through increased 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, as JOLs may not be affected in the same way. 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. Such encoding tasks could further be paired with a set of instructions designed to encourage participants to temper their JOLs. Indeed, Koriat and Bjork (2006) showed 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 (Koriat, 1997), metacognitive judgments are based on three domains: Readily

observable characteristics of the study items (i.e., intrinsic characteristics of the pairs, such as item difficulty, associative strength, etc.), manipulations at 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 showed that intrinsic and extrinsic factors influenced JOL strengths, though only intrinsic factors were shown to influence JOLs and recall rates equally. The cue-utilization framework then suggests that JOL overestimation should arise when participants are basing 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 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 emphasize the semantic relatedness of pairs when providing 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 JAM ratings (Maki, 2007). 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 overlap norms. JAM ratings are also prone to overestimation, and previous research (Maki, 2007; Valentine & Buchanan 2013) has shown that individuals typically perform poorly on such tasks. Maki (2007) proposed that this increase in JAM ratings for forward associates 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 JOLs: if individuals have inflated notions of associative strength, they may be more likely 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.

Finally, though not focal to the current study, our data may lend some clarity to the ongoing debate between the roles of perceptual fluency versus belief processes when generating JOLs (i.e., Blake & Castel, 2018; Mueller, Dunlosky, & Tauber, 2015). Perceptual fluency is generally conceptualized as the ease of processing word pairs, whereas beliefs correspond to expectancies regarding the memorability of the word pair. Fluency is likely related to speed of processing a cue–target word pair, and, therefore, we noticed a comparison to the semantic priming literature which has extensively examined processing speed of pairs including symmetrical and asymmetrical associates. In a comprehensive review, Hutchison (2003) indicated that semantic priming, computed as the response latency difference for targets when they are preceded by a related versus unrelated prime, was equivalent between BAS-matched forward and backward associates when averaged across several studies. This pattern suggests that processing fluency is, therefore, equivalent across forward and backward pairs. When examining JOL rates across pair types in our pooled analysis, we report a similar patten across pair types. Specifically, JOLs for backward pairs were within 6% of JOL rates for forward and symmetrical pairs. These JOL similarities are consistent with semantic priming patterns, which due to the implicit nature of semantic priming, is consistent with a fluency-based account versus an expectancy/belief-based account. Of course, our study was not designed to test these accounts as we did not attempt to assess fluency or beliefs for word pairs, but we note that priming patterns may be related to JOL rates which may be informative regarding the mechanisms behind JOLs.

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 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 qualitative 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.

Open practices statement

The data for all experiments have been made available at https://osf.io/hvdma/ and none of the experiments were preregistered.

Compliance with ethical standards

Ethical approval The studies reported were approved by the University of Southern Mississippi Institutional Review Board (Protocol #IRB-19-429) and found to be in accordance with the 1964 Helsinki Declaration ethical principles. Informed consent was obtained from all individuals who participated in this study. The authors report no completing interests.

Appendix

See Tables 1, 2, and 3.

Table 1 Summary Statistics for associative overlap variables

	Variable	*M*	SD	Min	Max
Forward	FAS	0.37	0.21	0.05	0.81
	BAS	0.00	0.00	0.00	0.00
Backward	FAS	0.00	0.00	0.00	0.00
	BAS	0.37	0.21	0.05	0.81
Symmetrical	FAS	0.19	0.13	0.01	0.46
	BAS	0.19	0.13	0.02	0.52

Values are grouped by pair type. 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 2 Summary statistics for cue and target item properties

	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

Values are grouped by pair type. 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 3 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*
Exp. 4	JOL	Forward	67.15	4.81			
		Backward	62.60	5.17	0.30*		
		Symmetrical	69.53	4.89	0.16	0.45*	
		Unrelated	38.33	6.89	1.59*	1.30*	1.71*
	Recall	Forward	70.91	5.71			
		Backward	32.17	6.51	2.06*		
		Symmetrical	59.73	6.66	0.59*	1.37*	
		Unrelated	16.48	5.72	3.11*	0.84*	2.28*
Pooled	JOL	Forward	67.00	2.39			
		Backward	64.50	2.60	0.17*		
		Symmetrical	70.62	2.42	0.29*	0.42*	
		Unrelated	30.70	3.04	2.29*	2.10*	2.56*
	Recall	Forward	67.89	2.85			
		Backward	32.84	2.97	2.08*		
		Symmetrical	59.50	2.97	0.50*	1.55*	
		Unrelated	13.28	2.33	3.62*	1.26*	2.99*

Mean JOL and recall rates for each associative direction condition across each experiment. The three rightmost columns indicate Cohen's *d* effect sizes for post hoc comparisons

*$p < 0.05$

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. https://github.com/crsh/papaja. Accessed 19 Dec 2019.

Balota, D. A., Yap, M. J., Hutchsison, K. A., Cortese, M. J., Kessler, B., Loftis, B., et al. (2007). The English lexicon project. *Behavior Research Methods, 39*(3), 445–459.

Blake, A. B., & Castel, A. D. (2018). On belief and fluency in the construction of judgments of learning: Assessing and altering the direct effects of belief. *Acta Psychologica, 186*, 27–38.

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.

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology & Aging, 12*(1), 50–71.

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), 175–191. https://doi.org/10.3758/BF03193146.

Goodman, L., & Kruskal, W. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*, 732–764.

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.

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review, 10*(4), 785–813.

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1304–1316.

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

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

Mueller, M. L., Dunlosky, J., & Tauber, S. K. (2015). Why is knowledge updating after task experience incomplete? Contributions of encoding experience, scaling artifact, and inferential deficit. *Memory & Cognition, 43*(2), 180–192.

Nelson, D. L., McEvoy, C. L., & Dennis, S. (2000). What is free association and what does it measure? *Memory & Cognition, 28*(6), 887–899. https://doi.org/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. https://doi.org/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 G. Bower (Ed.), *The psychology of learning and motivation*. Washington, DC: American Psychologist.

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

Rhodes, G. M. (2016). Judgments of learning: Methods, data, and theory. In J. Dunlosky & S. K. Tauber (Eds.), *The oxford handbook of metamemory* (pp. 90–117). Oxford: Oxford University Press.

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*, 131–148. https://doi.org/10.1037/a0021705.

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

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.

Scheck, P., Meeter, M., & Nelson, T. O. (2004). Anchoring effects in the absolute accuracy of immediate versus delayed judgments of learning. *Journal of Memory and Language, 51*, 71–79.

Tekin, E., & Roediger, H. L. (2017). The range of confidence scales does not affect the relationship between confidence accuracy in recognition memory. *Cognitive Research: Principles and Implications, 2*(1), 49–62.

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.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language, 28*, 127–154.

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.

Van Overschelde, J. P., & Nelson, T. O. (2006). Delayed judgments of learning cause both a decrease in absolute accuracy (calibration) and an increase in relative accuracy (resolution). *Memory & Cognition, 34*(7), 1527–1538.

Weber, N., & Brewer, N. (2003). The effect of judgment type and confidence scale on confidence-accuracy calibration in face recognition. *Journal of Applied Psychology, 88*(3), 490–499.