Summer 2023

Judgments of Learning and Retrospective Confidence Judgments: A Qualitative Exploration of Difference in Processes

David Hengerer
Claremont Graduate University

Judgments of Learning and Retrospective Confidence Judgments:

A Qualitative Exploration of Difference in Processes

By

David Hengerer

Claremont Graduate University

2023

Approval of the Dissertation Committee

This dissertation has been duly read, reviewed, and critiqued by the Committee listed below, which hereby approves the manuscript of David Hengerer as fulfilling the scope and quality requirements for meriting the degree of Doctor of Philosophy in Psychology.

Dr. Stacey Wood, Chair

Scripps College

Molly Mason Jones Professor of Psychology

Dr. Andrew R. A. Conway

New Mexico State University

Professor of Psychology and Psychology Department Head

Dr. Lori E. James

University of Colorado, Colorado Springs

Professor of Cognitive Psychology

Dr. Yaniv Hanoch

University of Southampton

Professor of Decision Sciences

Judgments of Learning and Retrospective Confidence Judgments:

A Qualitative Exploration of Difference in Processes

David Hengerer

Claremont Graduate University: 2023

Many studies of metamemorial confidence have found differences in calibration and resolution between two similar confidence judgments – judgments of learning (JOLs) and retrospective confidence judgments (RCJs). These findings have led to competing theories of the processes involved in JOLs and RCJs, and whether they make use of the same processes or different processes. This study critically tested two such explanations for JOLs and RCJs – the dual process descriptive model of confidence and the target accessibility model of confidence. Participants provided written justifications of their metamemorial confidence judgments for JOLs and RCJs for unrelated word-pairs. Justifications were analyzed using three different but complementary text data analyses – Latent semantic analysis, n-gram word frequency analysis, and support vector machine analysis – to determine whether both JOLs and RCJs utilize the dual process descriptive method, or if RCJs instead only utilize target accessibility. Results indicated that both JOL and RCJ justifications are characterized by a cue-familiarity check at lower levels of confidence and increasing amounts of partial target information as confidence increases. These findings support the dual process descriptive model of confidence, a model that states that confidence judgments are comprised of a cue-familiarity check followed by a retrieval attempt and associated partial target information. Additionally, results indicated that RCJ justifications made greater use of cue-based information than they did target-based information. This finding challenges theories that RCJs only utilize target accessibility as the source of metamemorial confidence and suggests other processes are involved.

Dedication

I would like to dedicate this work to my partner Megan for her unending support throughout this entire process, and in loving memory of our dog Marsha for her unlimited unconditional positive regard. I would also like to thank my mother and brother for their support and for believing in me. I would like to thank my grandfather for inspiring me to pursue a higher degree and for teaching me to love learning and discovering new knowledge.

Acknowledgements

I would like to thank my committee for their direction and help throughout my dissertation process – this was not a small ask of them and they all selflessly contributed their time and expertise to ensure I was not just stumbling through on my own. Finally, I would like to thank my advisor, Dr. Stacey Wood, for her unflagging support and guidance throughout my degree program, and for being willing to take a chance on me when no one else would. I would not be here, at the completion of the very arduous process of my graduate program without all her help, and for that I am eternally grateful.

Table Of Contents

Judgments of Learning and Retrospective Confidence Judgments:

A Qualitative Exploration of Difference in Processes

Metacognition is one's thinking about their own thinking (Arbuckle & Cuddy, 1969; Cutting, 1975; Flavell, 1979; Hart, 1967; Luttrell et al., 2013). It is often conceptualized as being comprised of two components, monitoring and control (Koriat, 1997; Koriat & Goldsmith, 1996; Nelson & Narens, 1990). These components are used to keep track of the current state of information in mind (monitoring) and update or change behaviors (control) based on the state of that information (Ackerman & Goldsmith, 2008; Dunlosky & Bjork, 2008; Goldsmith, et al., 2002; Weber & Brewer, 2008; Yaniv & Foster, 1995). Metamemorial confidence is a self-evaluation of memory performance – it can be thought of as the output of the metacognitive monitoring process (Ackerman & Goldsmith, 2008; Koriat & Goldsmith, 1996; Metcalfe & Finn, 2008; Nelson & Narens, 1990). It is thought to be used in decision and memory processes to determine what course of action to take (Azevedo & Hadwin, 2005; Bol & Hacker, 2012; Miller et al., 1960). Perhaps the most apt description of the utility of metacognition can be found in Nelson & Narens (1990, p. 128), "A system that monitors itself (even imperfectly) may use its own introspections as input to alter the system's behavior". In contrast, "poor self-monitoring capacity necessarily entails poor selection and execution of relevant control processes: if you do not know what you do not know, you cannot rectify your ignorance" (Benjamin et al., 1998, p.65).

Approaches to Metamemorial Confidence

There are two main approaches to explaining metamemorial confidence (MC) – direct access and inference. Direct access arises from Hart (1965, 1967) and Arbuckle and Cuddy (1969). According to Schwartz (1994) and Koriat (1997), the direct access approach involves the

individual having "direct access" to the memorial representation upon which the MC judgment is being made, as well as the ability to assess the strength of said representation. Importantly, this does not involve retrieval of information from memory, but instead privileged access to non-retrieved or to be retrieved information. It can be thought of as similar to Cowan (2008, 2017)'s conception of working memory, in that the information being monitored "directly" is held in a state of heightened availability, but not actually retrieved or in the focus of attention. Although direct access was chronologically the first approach to define the source of information used in MC judgments, it has fallen out of favor in recent years. This is because the direct access approach has two major implications that do not stand up to empirical scrutiny. First, any factor that affects recall or recognition performance should equally affect the MC judgment. This implication is rejected by studies such as that by Kornell et al. (2011), who found a differential effect of font size and study time on recall performance and MC judgments, such that increased font size led to greater magnitude of MC judgments but did not affect recall performance, whereas greater study time increased recall performance but did not affect magnitude of MC judgments. The second implication made by the direct access approach is that MC judgments should be at or near perfect accuracy for assessing retrieval, and that there should be no systematic factors that affect accuracy. This implication is rejected by looking at essentially any feeling-of-knowing (FOK) study (e.g., Costermans et al., 1992) – FOKs are slightly more accurate than chance at predicting future recognition success for items that were not recalled successfully – or looking at the delayed JOL effect (Dunlosky & Nelson, 1992) – immediate JOLs are significantly and systematically less accurate than are JOLs made after a brief delay.

The inference approach, on the other hand, states that all confidence is inferential and is based on a variety of cues – information available at the time of the decision/memory task. The

inferential approach explicitly precludes the possibility of privileged access to the specific memory trace upon which the primary task is based or the memory mechanisms that would assess trace strength prior to retrieval. There are several different inferential theories of MC judgments, but perhaps the most influential is Koriat's (1993, 1997) *cue-utilization* theory. Like all inferential theories of MC, cue-utilization theory posits that MC judgments do not involve direct/privileged access to the memory trace involved in the primary decision, and that MC judgments are based on cues. It goes further to define and give examples of three different types of cues: Intrinsic cues, extrinsic cues, and mnemonic cues. Intrinsic cues are factors specific to the items being studied or the decision being made. This can include the perceived ease or difficulty of committing items to memory and the associative relatedness between words in specific word pairs. Extrinsic cues are factors that pertain to the conditions in which the attempted learning/decision occurs in (the environment). Examples include the amount of time the to-be-learned material is presented, whether presentations are repeated, and the spacing between said repetitions (massed vs. spaced practice). Additionally, factors such as the level of processing involved in learning the item or making the decision are extrinsic cues, as are beliefs the individual may have about the effects of said processing. Finally, mnemonic cues, according to Koriat (1997), are internal phenomenological states that accompany the processing of information, which he refers to as "sheer, subjective experiences". Mnemonic cues include the accessibility of relevant information (Dunlosky & Nelson, 1992; Koriat et al., 1980; Nelson & Dunlosky, 1991), the ease with which information comes to mind (Kelley & Lindsay, 1993; Koriat, 1993), cue familiarity (Koriat, 1997; Metcalfe et al., 1993; Schwartz & Metcalfe, 1992;), and even the outcome of previous retrieval attempts (Finn & Metcalfe, 2007, 2008), among others. According to Koriat (1997), mnemonic cues are sensitive to intrinsic and extrinsic cues.

This is to say, the "sheer subjective experience" of the retrieval attempt changes depending on the presence or absence of the various intrinsic and extrinsic cues.

There are many different cues that have been found to affect metamemorial confidence. As already mentioned about mnemonic cues, information accessible as part of a retrieval attempt, such as partial target information, is used in MC judgments (Koriat, 1993; Koriat, et al., 1980; Son & Metcalfe, 2005). The fluency of encoding (Hertzog et al., 2003), retrieving (Kelley & Lindsay, 1993) or processing information (Koriat & Ma'ayan, 2005; Undorf & Erfelder, 2015) can also serve as a source of information for MC judgments, as can the motoric fluency associated with producing information (Susser, et al., 2017). However, it should be noted that there is some debate over whether it is fluency itself, or people's beliefs about fluency that actually affect their MC judgments (Mueller et al., 2013; Susser et al. 2017). Another prominent source of information for MC judgments is one that is somewhat related to partial target information - the memory for past test heuristic (Finn & Metcalfe, 2007; Finn & Metcalfe, 2008; Hertzog et al., 2013). This heuristic states that people remember their item specific performance from the previous test of the same information and use that performance as a basis for their MC judgments (Tauber & Rhodes, 2012). Interestingly, people can and often do use this heuristic when making MC judgments about other people's future memory performance (Serra & Ariel, 2014), although it tends to be slightly less accurate than when using it to make predictions about one's own future memory performance - perhaps due to the absence of the idiosyncratic mnemonic cues available from personally attempting retrieval (Maki, 2008; Serra & Ariel, 2014; Tullis & Fraundorf, 2017). Still other cues used for MC judgments include cue familiarity (Metcalfe, 1993; Schwartz & Metcalfe, 1992), which can be likened to the effects of prior exposure to information and has been found to be used frequently at the lowest levels of

confidence (specifically the absence of cue familiarity; Jersakova et al., 2017; Metcalfe & Finn, 2008b; Son & Metcalfe, 2005).

Metamemory Judgments

There are many different types of metamemory judgments, however, the three most studied are judgments of learning (JOL), retrospective confidence judgments (RCJ), and feelings of knowing (FOK). Judgments of learning are predictions of the likelihood of eventual memory performance (Dunlosky & Nelson, 1992). A judgment of learning is quite simply an answer to the question: How confident are you that in X units of time, you will be able to recall a target piece of information? JOLs often are solicited during paired associate word learning tasks, and are therefore in the form of: How confident are you that in X units of time, you will be able to recall the target word when presented the cue word? Importantly, JOLs are all predictive, they are asking the individual to assess their ability to retrieve information in the future. Judgments of learning and retrospective confidence judgments are similar in content, albeit different in performance. Similarly to a JOL, a retrospective confidence judgment (RCJ) is also a subjective assessment of confidence in one's own knowledge (Koriat, 2012). RCJs ask almost the same question about confidence in retrieving information as do JOLs, but the chronological direction of said task is the opposite. Instead of predicting how confident an individual is that they will be able to retrieve information in the future, a retrospective confidence judgment instead asks the individual how confident they are that they have correctly retrieved information at some point in the past. Feelings-of-knowing are slightly different, they are judgments made exclusively on information that was not successfully retrieved during a retrieval attempt and are more often used in retrieval tasks involving semantic information instead of episodic information. The individual is asked to estimate their likelihood of successfully recognizing said unretrieved information

(Leonesio & Nelson, 1990). Notably, FOKs tend to be significantly less accurate at predicting future recognition than JOLs are at future recall or than RCJs are at assessing the correctness of prior recall (Leonesio & Nelson, 1990).

Studies comparing RCJs and JOLs

Several studies have compared JOLs and RCJs, often hoping to determine if one is superior in terms of accurately assessing memory performance. Nguyen et al. (2017) conducted a head-to-head comparison of JOLs and RCJs in facial recognition tasks. They had participants engage in a standard cross-race effect eyewitness memory paradigm, and randomly assigned participants to engage in either immediate JOLs or JOLs made after a brief (~30 second) filled delay after studying a series of faces. Participants then engaged in an old/new recognition test, after which they made RCJs (the authors refer to them as postdictive metamemory judgments, which is not strictly speaking accurate given the procedure they used). They found that delayed JOLs were more accurate at predicting recognition performance than were immediate JOLs, replicating the delayed JOL effect (Nelson & Dunlosky, 1991), and ran a mixed effects linear model comparing the accuracy of delayed JOLs with RCJs. Although delayed JOLs were a statistically significant predictor in the best fitting model, RCJs were a much stronger predictor of memory accuracy. The authors followed up this analysis with a correlational analysis looking at the relationship between JOLs and RCJs and found the relationship to be weak and to not differ whether immediate or delayed JOLs were used. The authors took these findings as indicative of JOLs and RCJs involving different underlying cognitive processes but were not able to specify what the differences in processes were.

Dougherty et al. (2005) attempted to determine whether RCJs and JOLs assessed memory differently, looked at the relative accuracy of each metamemory judgment, and explored whether

making either confidence judgment improved later recall performance. In their first experiment, the authors had participants study word pairs and provided instructions that they would be later asked to retrieve the second word of each prompted with the first. Word pairs were presented for either three or twelve seconds each, varying between word pairs. After studying between 277- and 330-word pairs, each participant completed a self-paced forced response recall test. Immediately after taking this test, participants each made both RCJs and JOLs for each tested word pair – importantly, for both RCJs and JOLs the participant was presented with only the cue of the to be judged word pair and asked to make the metamemorial judgment accordingly. After making metamemorial judgments, participants were immediately presented with a second self-paced forced response recall test. After completing this second recall test, participants again made RCJs on all word pairs they were tested on. To assess whether RCJs and JOLs were based on the same information, the authors examined the correlation between both RCJs and JOLs with the initial recall test performance, the latency of said initial recall test, and assessed whether both RCJs and JOLs were equally affected by the study time manipulation.

The authors hypothesized that both MC judgments were based at least partially on retrievability (i.e., partial target information retrieval). Consistent with this hypothesis, the authors found that RCJs and JOLs both correlated strongly with initial recall test performance, but that the correlation was significantly higher for RCJs than it was for JOLs. There was no effect of study time on the strength of the observed correlation between either MC judgment and initial recall performance. The authors further examined the correlations between MC judgment magnitude and latency on the initial retrieval test. When ignoring whether the MC judgment was correct, Dougherty et al. (2005) found that both RCJ and JOL magnitude correlated significantly with retrieval latency, but that the correlation was significantly larger for RCJs, such that

although longer retrieval latencies were associated with lower magnitude MC judgment for both JOL and RCJ, this was even more the case for RCJs. The authors suggest that this finding is strong support for the hypothesis that RCJs are based on retrievability. Interestingly, when looking at correct and incorrect MC judgments separately, only correct JOLs did not correlate significantly with retrieval latency, indicating that speed of retrieval was only related to JOL magnitude for incorrect JOLs. Additionally, the authors plotted the distribution of confidence judgments for both RCJs and JOLs conditionalized on correctness of MC judgment. The figures revealed that both RCJs and JOLs appear to handle incorrectly recalled items using the same processes and information but handled correctly recalled items differently – whereas RCJs for correct items tended towards the high extreme of the confidence scale, JOLs for correct items made more use of the middle of the confidence scale. The authors took this finding as suggesting that specifically for correct items, participants were either using different information or different processes to make JOLs as opposed to RCJs. A final analysis comparing the initial recall results and the MC judgments concerned the absolute magnitude of MC judgments. Although both RCJs and JOLs were higher for items studied for 12 seconds as opposed to 3 seconds, RCJs were higher than were JOLs for both the 12 and 3 second conditions. The authors took this finding as further support that there is a procedural difference between RCJs and JOLs.

Dougherty et al. (2005) interpreted the totality of their findings in their first experiment as indicative of RCJs and JOLs being based on different information. They referenced three main findings to support this conclusion. First, the differential effect of the study time manipulation on the correlation between MC judgments and final recall performance is evidence that people utilized the study time cue differently in forming said judgments. Second, the correlation between RCJs and prejudgment recall latency but absence of correlation between JOLs and

prejudgment recall latency indicates that people used retrieval fluency more in forming their RCJs than they did for their JOLs. Finally, the distributions of judgments for RCJs and JOLs differed considerably among correctly recalled items.

However, the authors were concerned that their findings might have been an artifact of their within-subjects design, and that participants were deliberately making their MC judgments differently as a result of said design. To test for this possibility, Dougherty et al. (2005) conducted a second experiment, replicating their procedure but with a between-subjects design. They found that although both JOLs and RCJs were strongly correlated with prejudgment recall, RCJs were once again more strongly correlated with prejudgment recall. However, in the between-subjects design, both RCJs *and* JOLs correlated with latency of the prejudgment recall, and there was not a significant difference between the correlations for type of MC judgment and prejudgment recall latency. Despite this, the correlation between latency for incorrectly recalled items and JOLs was significantly higher than was the correlation between latency for correctly recalled items and JOLs, whereas RCJs were equally well correlated with latency for both correct and incorrect items. The authors took this finding as support for the notion that latency of retrieval (i.e., retrieval fluency) is more important for RCJs than it is for JOLs. This finding also suggests that for correctly recalled items participants made use of cues in addition to retrieval fluency, but for incorrectly recalled items, retrieval fluency was a major source of information. This finding is particularly interesting when taken in conjunction with Son and Metcalfe's (2005) dual-process theory and findings that items given the lowest and highest confidence ratings were judged the fastest, but that the lowest confidence ratings were typically associated with an absence of cue familiarity (see also Metcalfe & Finn, 2008b). Further results from Dougherty et al.'s (2005) second experiment once again revealed that participants tended to use the mid-point

of the scale more when making JOLs as opposed to RCJs – a finding the authors qualify as possibly indicative of different processes involved in the MC judgments, but also possible different use of response scale between the two MC judgments. The authors followed up on this possibility by examining the part-scale correlations between MC judgment magnitude and final recall accuracy. Interestingly, their results indicated that RCJs were more able to successfully predict future recall, but that JOLs were more able to discriminate between items that received a MC judgment magnitude of zero as opposed to non-zero magnitude. This finding is in line with the dual-process descriptive model for JOLs (Jersakova et al., 2017; Metcalfe & Finn, 2008b; Son & Metcalfe, 2005). Finally, participants' response latency when making JOLs was significantly longer than was their response latency for making RCJs, once again suggesting different processes being involved in the two judgments. Taking the findings of both experiments together lends support to JOLs utilizing a dual-process method involving cue familiarity and retrievability and to RCJs relying more on retrievability alone.

Other studies have also explored the differences between JOLs and RCJs, generally by examining the correlation between RCJs and recall as opposed to JOLs and recall and tend to find that RCJs are a more accurate assessment of performance and basis for restudy decisions (Hines et al., 2009; Perfect & Hollins, 1996; Robey et al., 2017; Ryal et al., 2016; Siedlecka et al., 2016; Wattier & Collins, 2011).

Studies examining justifications for Metamemorial Confidence

Koriat et al. (1980) conducted perhaps the first study to examine justifications for confidence judgments. They conducted a two-experiment study looking at peoples' retrospective confidence in their answers to general knowledge and specifically whether people's confidence ratings were biased by their justifications of said answers. The authors were concerned about the

high rate of overconfidence observed in a review by Lichtenstein et al. (1982) and suspected that people selectively focus on evidence in support of selected responses and ignore or disregard evidence against their selected response. This suspicion would later be supported by a finding from a study of perceptual discrimination confidence by Maniscalco et al. (2016) – people appear to track evidence for and against alternatives in separate accumulators, and only the accumulator containing evidence for an alternative is considered during a confidence judgment. To test their supposition, Koriat et al. (1980) had participants review a series of two-alternative forced choice general knowledge questions, select one of the alternatives, and then had participants in the experimental condition list reasons for *and* against their choice (Experiment 1) or for *or* against their choice - or both for and against their choice (Experiment 2), as well as rating the perceived strength of each reason. All participants then provided a retrospective confidence judgment assessing the probability of their choice being correct. In their first experiment, the authors found that participants in the control condition (those who did not list reasons for or against their selected choice) demonstrated the expected overconfidence outcome; the participants in the experimental condition (those listing reasons for and against their choice) also demonstrated overconfidence, but to a statistically significantly lesser degree. In their second experiment, Koriat et al. (1980) found that improvements to confidence calibration were strongest in the group of participants who only listed reasons contradicting their selected response – a finding the authors took as supporting their hypothesis that overconfidence stems at least partly from a tendency to focus on congruent and disregard incongruent evidence for a selected alternative.

Relevant to the current study, the authors analyzed the reasons for and against any given selection. They found that on average, participants provided more reasons for (ex. “I know for

sure…") than against (ex. "I vaguely remember…") their selected alternative and that the strongest correlation with confidence ratings was the sum of the strength of reasons for a specific alternative. They took this finding as suggesting that confidence is determined based on the strength of evidence retrieved for an alternative; a finding that Koriat later expanded upon and developed into his target accessibility view of metamemorial confidence (Koriat, 1993). Taken together, this as well as some of Koriat's later work suggest that RCJs are based on partial retrieved target information such that higher confidence ratings are associated with more partial information having been retrieved. It should be noted, however, that the justifications in Koriat et al. (1980) are for the primary decision task, and not technically for the confidence judgment itself. This discrepancy could potentially prove to be problematic when attempting to apply the target accessibility model of metamemorial confidence to actual confidence justifications and leaves open the possibility that, as per Hanczakowski et al. (2013) and Selmeczy and Dobbins (2014), RCJs are in fact made in a manner consistent with a dual-process model like that of JOLs (Jersakova et al., 2017; Metcalfe & Finn, 2008b; Son & Metcalfe, 2005).

Gardiner et al. (1998) examined peoples' reported phenomenological experiences of remembering, knowing, and guessing following a yes/no recognition task. The authors collected transcripts of participants' descriptions of their experiences and subjected those transcripts to a semi-formal content analysis. This content analysis was primarily concerned with the nature of responses in each possible response category, and how the content of responses differed between categories. Remember responses generally involved intra-list and extra-list associations, as well as imagery and sometimes superficial word characteristics. The authors took these responses as reflecting the use of effortful strategies and associations as mnemonic aids. Remember responses also included indications of words triggering involuntary/automatic recollection of personal

memory from participants' everyday lives. Remember responses were also formally coded for content. Two raters independently studied the transcripts and assigned the responses into five different categories: intra-list association, extra-list association, item-specific image[s], item's physical features, and self-reference[s]. Raters agreed on 81% of the 90 responses – the authors took this as an acceptable level of concordance. Taking these results together, the authors suggested a distinction between remembering due to voluntary use of study strategies and remembering as a result of involuntary associations with personal experiences. The authors go on to posit that for both types of remembering, the remembered words triggered an engagement of consciousness beyond just being aware of the presence of the word.

Know responses, on the other hand, generally lacked specific contextual or evidence of perceptual experience. Responses instead indicated familiarity, "just knowing", and thinking that the word had occurred in the list. They were also characterized by reported absences of the type of information reported for remember responses. The authors also note that both remember and know responses were accurate at distinguishing between studied and unstudied words, with both occurring significantly more for studied than unstudied words. The final type of response, guessing, differed from both remember and know responses in that guess responses did not accurately discriminate between studied and unstudied words. Guess responses were characterized by inferences and strategies that were unrelated to memory for the studied word. Additionally, guess responses tended to organically indicate ratings of uncertainty, including phrases such as "not sure", "not confident", and "I think". These expressions of uncertainty were generally not present in the know responses. On the other hand, ratings of certainty occurred with some frequency in the know responses but did not occur at all in the guess responses. These results indicate different levels of certainty, and therefore confidence, associated with the

different subjective experiences of recognition. The authors also note that both guess and know responses included indications of familiarity.

Taken together, the results suggest a clear distinction between phenomenology for remembering, knowing, and guessing, as well as provide support for a dual-process mechanism of recognition memory. In terms of the current study, these results provide some support for a dual-process mechanism for metamemorial confidence, as assessing whether a particular memory decision is characterized as remembering, knowing, or just guessing is itself a self-assessment of performance and can therefore be loosely regarded as a sort of confidence judgment in and of itself. However, remember/know judgments are not necessarily solely based on confidence (McCabe et al., 2011; Rajaram et al., 2002), therefore this study alone is not sufficient evidence of a dual-process mechanism for metamemorial confidence.

Williams et al. (2013) examined whether the content and nature of people's subjective reports was sufficient for observers to recover confidence and experiential state information from other people's memory tests. The authors used the 270 justifications reported in Gardiner et al. (1998) as stimuli, and had participants assign a confidence rating to each justification (Experiment 1) or categorize the justifications as describing subjective experiences of remembering, knowing, being familiar with, or guessing the original words (Experiments 2 & 3), after the authors had manipulated the confidence level associated with each justification as determined in Experiment 1.

For their first experiment, the authors posited that if participants' confidence ratings were able to differentiate justifications for different subjective experiences it would provide support for there being a true relationship between confidence and Remember/Know responses such that use of subjective experience generates confidence. Participants each reviewed the cue word and

justification for 27 cue/justification pairings from Gardiner et al. (1998). They then were presented with the question 'How confident do you think this participant was that they had accurately recognized this word?" and responded via a 0-100 (in increments of 5) scale. Remember items had the highest confidence rating, followed by Know items, and then finally Guess items. Analysis of variance revealed a significant main effect for type of subjective experience and post hoc t-tests indicate that confidence ratings differed significantly between each subjective experience. The authors took these results as indicating a reliable relationship between judgments of confidence and subjective memorial experience, even for other people's subjective memorial experiences. Additionally, the authors split the Know justifications into justifications indicating knowing and justifications indicating familiarity. They then repeated the earlier analysis of variance but using the now four subjective experience categories as a factor in place of the previous three level factor. They once again found a significant main effect, such that remember justifications were classified as higher confidence than were know justifications, which were in turn classified as higher confidence than were familiar justifications, which were once again in turn classified as higher confidence than were guess justifications. The authors took this finding as evidence that people can distinguish between the experiences of knowing and of familiarity when using others' memorial reports. Together, the results provide support to the supposition that subjective awareness leads to confidence.

For their second and third experiments, Williams et al. (2013) selected justifications that were associated with low, medium, and high confidence in their first experiment, and then tasked participants with determining whether those justifications belonged to the remember, know, familiar, or guess categories of subjective experience. In their experiment 2, they found that participants were able to correctly classify remember and familiar justifications, but that know

and guess justifications were affected by the manipulated confidence rating, such that items with higher confidence ratings were categorized as know more frequently than familiar or guess, and items with lower confidence were more frequently categorized as guess as opposed to familiar, regardless of their actual categorization. The authors suggested that these findings could be because the categories of remember and familiar directly describe the processes involved in the recognition decision, whereas know and guess do not directly describe the processes involved in recognition. Overall, the results indicate that people's judgments of subjective experience were influenced by confidence ratings, albeit more so for know and guess subjective experiences. In their third experiment, Williams et al. (2013) directly manipulated which confidence ratings were associated with specific justifications, such that specific justifications were presented with different confidence ratings. They once again hypothesized that if confidence influences interpretations of subjective experience, different confidence levels associated with specific justifications could lead to participants interpreting the justifications as belonging to a different category of subjective experience than the one it had been initially classified as. They found that, unlike in experiment 2, confidence ratings did not influence to which category participants classified each justification as belonging. Instead, participants generally correctly classified each justification regardless of the assigned confidence rating. Taken together, the authors posit that their results support the proposition from both Tulving (1985) and Gardiner et al. (1998) that confidence is derived from subjective experience. They justify this conclusion as drawn from their finding that when both a justification statement and a confidence rating are provided, participants categorized the justification using the text of the justification itself rather than the confidence value, and that this was because participants considered their understanding of memory experience to determine that subjective experience was more important than confidence.

This finding and the authors' conclusion match well with Koriat (1993, 1997)'s conception of confidence as being driven by the sheer, subjective experience of retrieval, and not vice versa. The authors also posit that their findings are inconsistent with single-process models, but acknowledge that given the results of their second experiment, there is some room for interpretation. It is possible that these findings imply that much like the findings of Maki (2008), when the sheer, subjective experience of retrieval is available as a cue for confidence, it will be preferred over other possible sources of information.

Selmeczy and Dobbins (2014) examined how people spontaneously defined and justified remembering and knowing experiences in a recognition memory paradigm. They had participants engage in an encoding task – judging how many syllables a word contained for their Experiment 1 or just viewing a word for their Experiment 2, for 100 words. They then tested participants on their recognition memory by having participants judge whether words were old or new. Participants also provided a confidence rating for each old/new recognition decision, and on a small number of items participants also were asked to justify their confidence rating by describing in detail why they chose the specific confidence rating they provided. The authors then conducted three different text data analyses to examine participants' understanding and experience of remembering and knowing as they were associated with different confidence levels. The text data analyses were n-gram analysis, rater scorings of justifications, and Support Vector Machine. They expected that high confidence reports would contain recollection-linked content not found in the medium confidence justifications, as per Tulving (1985)'s characterization of autonoetic and noetic consciousness.

Their first n-gram analysis looked at frequency of words and short phrases that participants used in their justifications for items that were correctly recognized as old (hits),

examining whether frequency of specific n-grams differed reliably from chance at the different levels of confidence. They found that n-grams fell into two broad categories: those indicating intensity of memory and those indicating the conscious experience of remembering. Additionally, some high confidence justifications also included temporal information, whereas medium confidence justifications did not. Ultimately their n-gram analysis indicated that participants expressed awareness that the word they were providing a confidence judgement for was part of a past personal experience. This finding supports Tulving (1985)'s conception of autonoetic consciousness as well as the authors' hypothesis that people can recognize differences between the phenomenology of remembering as opposed to knowing. Interestingly, the n-gram analysis also revealed that participants used the absence of remembering as justification for medium confidence ratings and for judging new items. Further, although participants did not use the word "know" more frequently for their medium confidence justifications, they did use the word "familiar" more frequently for this level of confidence than for high confidence justifications. Also appearing more frequently for medium as opposed to high confidence justifications were modifiers that indicated reduced certainty, such as the words "but" and "if", as well as negations and the first-person present tense of the verb to be, "am", which the authors speculated as reflecting that familiarity-based recognition exists in the perceptual present. Interestingly, a similar pattern of word usage was found in the medium confidence justifications for correct rejections – a finding the authors suggest indicates that medium confidence ratings for both old and new items exist on the same dimension.

The second n-gram analysis looked at frequency of words and short phrases that participants provided in their justifications for items that were correctly recognized as new (correct rejections). High confidence correct rejection justifications had a high rate of the word

"remember", in this case used to indicate a clear absence of remembering an item as evidence that it was new. This finding suggests that people use subjective memorability heuristics during recognition judgments, as initially proposed by Brownet al. (1977). On the other hand, medium confidence correct rejections indicated higher levels of uncertainty and often contained negations. Taken together, the authors suggest that their n-gram analyses indicate that high confidence is associated with remembering or its absence, and that medium confidence is more associated with feelings of familiarity.

Selmeczy and Dobbins (2014) also had raters analyze participants' confidence justifications. Raters categorized responses into 4 categories that arose from an initial review of said responses: personal experiences outside of the experiment; imagery, feelings, and thoughts; notable absences of memory; and strategies to memorize words. For each category of response, the authors conducted a repeated measures analysis of variance, with item type (hits vs. correct rejections) and confidence levels (high vs. medium) as the factors. The ANOVA for category 1, personal experiences, revealed a main effect for the confidence factor but not for item type and no interaction. The authors state that this finding indicates that personal experiences were mentioned more often for high rather than medium confidence justifications, for both hits and correct rejections. The ANOVA for category 2, imagery, thoughts, and feelings, revealed a main effect of confidence, and marginal effects for item type and the interaction between factors. The authors interpreted these results as suggesting that high confidence justifications included more instances of imagery, thoughts, and feelings than did medium confidence reports, and that this finding was stronger for hits than it was for correct rejections. The ANOVA for category 3, notable absences of memory, revealed a significant main effect for both confidence level and item type, as well as a significant interaction. The authors took these results to mean that notable

absences of memory occurred more during correct rejections than during hits, but also occurred more during high confidence correct rejections than during medium confidence correct rejections. Finally, the ANOVA for category 4, strategies to memorize words, revealed a main effect for confidence level, a marginal effect for item type, and a significant interaction between the factors. Category 4 words appeared significantly more often in high confidence hits as opposed to medium confidence hits and did not differ in frequency between levels of confidence for correct rejections. Taken together, the results for ratings of justifications indicate more instances of remembrance (or a notable absence of remembrance) for high confidence as opposed to medium confidence justifications. This finding is both in line with the results of the n-gram analyses Selmeczy and Dobbins (2014) conducted and with Tulving's (1985) conceptualization of autonoetic consciousness.

Finally, Selmeczy and Dobbins (2014) also conducted a Support vector machine (SVM) analysis to parse high and medium confidence hit justifications. Support vector machine is a machine learning algorithm used to classify objects into categories based on the maximal distance between cases of those categories. It does this by drawing hyperplanes called support vectors and then checking the distance between the support vector and the objects to each side of them and repeating this process until the distance from the support vector to the objects in each direction are at the maximum possible value (Selmeczy & Dobbins, 2014; Hamel, 2009). The SVM analysis was conducted to test whether recognition judgments were based on a unidimensional strength signal or instead a dual process model. If recognition judgments were based on a unidimensional strength signal, then the SVM would have difficulty distinguishing between levels of confidence, as the only difference in terms of words used for each confidence level would be the intensity of said words (ex. somewhat versus very confident; might recall

versus will recall). If, instead, recognition confidence is based on dual process signal detection, such as that proposed by Yonelinas (1994), the SVM should be able to effectively classify responses referring to a familiarity process as opposed to those involving a "conscious recollection of prior contextual information". In terms of SVM, sensitivity refers to the ability to distinguish between categories; specificity refers to whether the distinction used by the SVM is unique among different possible categories. Selmeczy and Dobbins (2014) made four predictions for their SVM results: high sensitivity for high as opposed to medium confidence hits, high specificity for high as opposed to medium confidence correct rejections, low sensitivity to medium as opposed to low confidence hits, and no specificity for medium as opposed to low confidence hits. Their results supported all their predictions. The SVM results indicated the existence of features of the confidence justifications that clearly differentiated high as opposed to medium confidence hits, and that those features were not present in high versus medium confidence correct rejections. These findings indicate a categorically different experience taking place in high confidence hits as opposed to medium confidence hits; this finding supports the Dual Process Signal Detection model for recognition judgments (Yonelinas, 1994). The SVM was also only moderately able to differentiate medium as opposed to low confidence hits and was even less able to differentiate medium versus low correct rejections. These last two findings suggest that low and medium confidence judgments exist on a single dimension of familiarity and are best distinguished by the intensity of familiarity.

In terms of the current study, Selmeczy and Dobbins (2014) results provide support for a dual-process model of MC judgments, even for retrospective confidence judgments (RCJs), at least for recognition paradigms. This finding is in line with findings from Jersakova et al. (2017) that JOLs use a dual-process mechanism involving familiarity and recollection at differing levels

of confidence, but also directly contrast findings from Koriat et al. (1980) that indicate that RCJs only involve a unidimensional target accessibility process and do not involve cue familiarity. This study also pioneered the method used in Jersakova et al. (2017) that is also used in the current study.

Jersakova et al. (2017) conducted a three-experiment study comparing justification content of scale and binary delayed JOLs. In each experiment, the authors presented participants with cue-target word pairs, then predict their ability to later successfully recognize (Experiments 1 and 2) or recall (Experiment 3) the target words when presented with the cue word of each word pair. Participants made these predictions on a 6-point numeric confidence scale (Experiment 1), a binary yes/no JOL in conjunction with a 3-point confidence rating (sure, maybe, guess) (Experiment 2), or randomly assigned participants to make the numeric or binary JOL (Experiment 3). Following some of the JOLs in each experiment, participants were asked to provide a written justification for their JOL. The authors analyzed these justifications using three complementary text data analysis techniques: the n-gram frequency analysis and SVM technique used in Selmeczy and Dobbins (2014) and a latent semantic analysis (LSA, Landauer & Dumais, 1997). The latent semantic analysis was used to determine whether specific justifications were more likely to refer to cue or target terms of each word pair. Results indicating references to cue terms at any JOL level serve as evidence of a dual process model of JOL confidence (Selmeczy & Dobbins, 2014; Son & Metcalfe, 2005; Yonelinas, 1994), whereas results indicating only references to the target term would support the target accessibility model of JOL confidence (Koriat, 1997; Koriat, et al., 1980; Nelson & Dunlosky, 1991). The second text data analysis the authors conducted was an n-gram frequency analysis. Similarly to how it was used by Selmeczy and Dobbins (2014), the authors compared frequency of unique phrases (in this case only

bigrams and trigrams) in justifications for JOL categories against justifications for other JOL categories. The authors specifically looked for phrases that indicated familiarity and remembering. Finally, the authors used Support Vector Machine (SVM) to find the point of maximum difference between justifications for different levels of JOL confidence. If a difference in processes used in making JOLs at different confidence levels exists, then SVM should have relatively high classification accuracy. If, on the other hand, the only difference between JOLs of different confidence levels is magnitude of the same process, SVM classification accuracy would be low. In line with Metcalfe and Finn's (2008b) prediction that 0% JOLs should reference the cue term, whereas other JOLs should increasingly reference the target term, the authors predicted that SVM classification for the numeric JOLs should be high for 0% JOLs as opposed to all other JOLs. The authors did not have a prediction on SVM classification accuracy for binary JOLs with confidence ratings.

LSA results for Experiment 1 revealed that 0% and 20% JOL confidence justifications were more likely to refer to the cue term than to the target term of the word pair. The 100% JOL confidence justifications were more likely to refer to the target term than to the cue term, and JOLs in the 40%-80% range shifted gradually from cue term referencing to target term referencing. LSA results for Experiment 2 indicated that justifications of "Guess" confidence responses for both Yes and No JOLs were more likely to refer to the cue term than to the target term of the word pair. In Experiment 3, justifications for 0%, 20%, and 40% JOLs all referred to the cue term more than the target term, justifications above 40% did not differ in their reference to cue term as opposed to target term. Justifications for binary JOLs at the "No-sure" and "No-maybe" level referred to the cue term more than to the target term, references did not differ for other levels of binary JOL. Taken together, the authors posited that the LSA results indicated that

both cue and target information was used in making JOLs of different levels, a finding that they took as supporting the dual-process model of JOL confidence (Metcalfe & Finn, 2008b; Son & Metcalfe, 2005; Yonelinas, 1994).

N-gram word frequency analysis for Experiment 1 revealed that the 0% JOL confidence level was best characterized by an absence of remembering and a lack of cue familiarity. The 20% JOL confidence level justifications demonstrated a vague sense of familiarity and a lack of explicit recollection. The 40% JOL justifications expressed some level of cue familiarity, and the 40% and 60% justifications both included explicit references to possible target accessibility. The 80% JOL justifications included references to some levels of certainty, and the 100% JOL justifications indicated explicit remembering of the target of the word pairs. Experiment 2 n-gram word frequency results indicated references to high levels of certainty in justifications for "Yes-sure" and "No-sure" responses, with those justifications referring to remembering and absence of remembering, respectively. Justifications referring to familiarity (or a lack thereof) were predominantly in the various "No" responses. The authors took these results as suggesting that the primary distinction between "Yes" and "No" JOLs was whether the cue term was familiar to participants. Experiment 3's n-gram word frequency analysis results were generally consistent with the results from Experiments 1 and 2, with 0% and 20% JOL justifications referring to not remembering and remembering seeing the cue term, respectively, for the numeric JOL ratings. Similarly, the "yes-sure" and "no-sure" justifications in the binary JOL condition also referenced not remembering the cue term and remembering the target term, respectively. These results further support the dual-process model of JOL confidence.

Finally, the SVM analysis for Experiment 1 revealed high classification accuracy for 0% as opposed to 20% JOL justifications, whereas all other JOL level comparison had lower

classification accuracy. This finding supports Metcalfe and Finn (2008b)'s descriptive model of JOL confidence, such that the difference in processes used occurs at the lowest points of the JOL scale. SVM analysis for Experiment 2, on the other hand, had highest classification between "yes-maybe" and "yes-sure" justifications, even higher than the classification accuracy between "yes" and "no" responses. However, classification for all responses categories was accurate at above chance levels, but that "yes" justifications were generally more clearly differentiated than were "no" justifications. The authors believed that these results indicated participants were using the different response scales in different manners, and that they did not map directly on to one another. Experiment 3 SVM results were generally consistent with Experiment 1 and 2's results, the highest classification accuracy for numeric JOLs was 0% to 20%, and no other classifications were at above chance levels for numeric JOLs. Binary JOL classification results were also similar to the results of Experiment 2, once again with the highest classification accuracy at the "yes-sure" and "yes-maybe" comparison point. Additionally, like Experiment 2's results, "yes" response justifications were more clearly differentiated than were "no" response justifications. The authors took these results as indicating that 1) participants used the numeric JOL scale similarly for recognition and recall JOLs, and 2) the two different JOL scales (numeric and binary) were not used in the same manner, even within a given mode of remembering.

In terms of the current study, Jersakova et al. (2017) provides strong evidence for the dual-process view of JOL confidence in both recognition and recall - in conjunction with Selmeczy and Dobbins (2014), these studies provide a particular methodology that is extremely well suited for comparing processes used in different types of confidence judgments. With that in mind, the current study is based on the methodology used in these two studies to compare the processes involved in JOLs as opposed to RCJs.

The Current Study

Taking all of the above mentioned studies together, there is ample evidence that JOLs operate following a dual-process model (Dougherty et al., 2005; Jersakova et al., 2017; Metcalfe & Finn, 2008b; Son & Metcalfe, 2005) and that JOLs and RCJs appear to require different processes, at least to some extent (Dougherty et al., 2005; Maki, 2008; Nguyen et al, 2017; Robey et al., 2017; Watier & Collin, 2011). There is also mixed evidence on whether RCJs also use a dual-process method (Gardiner et al., 1998; Hanczakowski et al., 2013; Selmeczy & Dobbins, 2014; Williams et al., 2013) or instead rely solely on target retrievability (Dougherty et al., 2005; Koriat et al., 1980). With this discrepancy in mind, the purpose of this study is to test the dual-process model of JOLs and to assess whether RCJs are also based on the same dual-process model or instead based solely on target retrievability/accessibility. It does so by utilizing a method very similar to that of Selmeczy and Dobbins (2014) and Jersakova et al. (2017) – looking at participants' justifications for their confidence judgments for both JOLs and RCJs via three different and complementary text-data analyses. These three text-data analyses (described further in the results sections) are LSA, N-gram frequency analysis, and SVM. Each of these methods of analysis are described in detail further in the results section but are also introduced briefly here, along with the predicted pattern of results for each type of analysis. It should be noted that the null hypothesis for these analyses, when comparing *across* the two types of confidence judgments, is that JOLs and RCJs both operate on the same dual-process model; the alternative hypothesis is that JOLs and RCJs invoke different processes, with JOLs operating on the dual-process model and RCJs only involving target-accessibility.

The LSA provides information about the semantic similarity of justifications to the cue and target terms of the word pairs upon which the justifications are based. The LSA analysis

returns a cosine value for each justification in regard to both the cue term and the target term the justification is based on. Higher cosine values indicate more semantic similarity. Greater semantic similarity between the cue term and the justification is evidence of cue familiarity being used as a process for making the MC judgment; greater semantic similarity between the target term and the justification is evidence of target accessibility being used as a process for making the MC judgment. Cue-justification and target-justifications cosines that both are significantly different from zero but not significantly different from each other, in conjunction with the n-gram analysis for these confidence judgments indicating the presence of n-grams referring to familiarity *and* n-grams referring to recollection, could suggest that both processes are used simultaneously in making a confidence judgment, or that one or more different processes entirely are used. For JOLs, it is predicted that the cosine value for the similarity between cue terms and justifications will be greater than the cosine value for similarity between target terms and justifications at the lower end of the JOL scale, particularly 0% JOLs and 20% JOLs. For 40% JOLs through 80% JOLs, it is predicted that both cue-justification cosine and target justification cosine will be significantly greater than zero but that there will not be a significant difference between these cosine values. For 100% JOLs, it is predicted that the cosine value will be greater for target terms and justifications as opposed to cue terms and justifications. All these predictions are in line with the findings of Jersakova et al. (2017).

There are two possible predicted patterns of results for the LSA for RCJ justifications – one pattern would be indicative of RCJs operating via the same dual process model as JOLs, the other would be indicative of RCJs operating solely on target accessibility. If RCJs operate via the same dual process model as do JOLs, the LSA results for RCJs should be reminiscent of the LSA results for JOLs. Specifically, at the 0% and 20% RCJ levels, there should be more semantic

similarity between cue terms and justifications than between target terms and justifications. At the 40% to 80% RCJ levels, there should be relatively equal amounts of semantic similarity between cue terms and justifications as there is between target terms and justifications, but cosines for both cue and target similarity should be different from zero. Finally, at the 100% RCJ level, there should be more semantic similarity between target terms and justifications than there is between cue terms and justifications. Alternatively, RCJs may operate solely by target accessibility. If this is the case, the LSA results for RCJs should be such that all levels of RCJ (0%-100%) demonstrate greater semantic similarity between target terms and justifications than between cue terms and justifications.

The n-gram frequency analysis looks at how rates of occurrence of words and short phrases differ between different levels and types of confidence judgments. The n-gram analysis informs as to the specific content of justifications for the different levels of confidence judgments. If, for example, the 0% JOL justifications have a higher frequency of the phrase "not familiar" than do the 20% JOL justifications, this is revealed by the n-gram analysis. The n-gram analysis also allows for capturing specific changes in justification content across levels of confidence, such as increases in recollection-specific terminology as confidence increases (Jersakova et al., 2017) or differential use of intensity modifiers to indicate varying levels of certainty. Of particular interest for this analysis is whether specific terms indicating cue-familiarity (ex. "vaguely remember seeing", "I remember seeing") , or a lack thereof (ex. "not remember seeing", "do not recall"), occur at the 0% and 20% confidence levels and if the frequency of said occurrence differs between the two types of confidence judgments. Also of interest would be the frequency of occurrence of words or phrases indicating recollection at all levels of confidence judgment, but particularly for 0% and 20% RCJs. A higher frequency of

words or phrases indicating recollection, or a lack thereof, at the 0% and 20% RCJs as opposed to the 0% and 20% JOLs could be evidence that only target accessibility is used during RCJs.

The final text analysis is SVM, a supervised machine learning technique used for classification of data. The SVM will be fed vectors of data which are comprised of the different words and phrases identified in the n-gram analysis. The SVM is trained on a set of pre-identified data and then should be able to classify additional data based on whether it fits better into one of two categories. This technique provides a unique opportunity to differentiate justifications both within a given type of confidence judgment at each level of confidence, and across both types of confidence judgment for a given level of confidence. For the within confidence type SVMs, the two possible outcome categories are two adjacent confidence levels, for example 0% and 20%. For the between confidence type SVMs, the two possible outcome categories are the two different MC conditions, JOL and RCJ. As was the case with both Selmeczy & Dobbins (2014) and Jersakova et al. (2017), the training dataset is a random sample of 50% of the justifications for each type of confidence judgment at each level of confidence. Each justification in the training dataset can be tagged for whether it was in response to a JOL or an RCJ. Separate SVM analyses are conducted looking at classification accuracy at each level of confidence within a given type of confidence judgment, as well as looking at classification accuracy for each type of confidence judgment within a given level of confidence.

For the first set of the SVM analyses, looking within a given type of confidence judgment, the pattern of results should help illuminate the process involved in making that type of confidence judgment. If RCJs are solely based on target accessibility (i.e. quantity/strength of evidence, as per Koriat et al., 1980), it is expected that SVM classification accuracy will be low at all confidence levels for RCJs. If RCJs are in fact based on a dual-process model then SVM

classification should be high, specifically at low (0%-20%) levels of confidence. It is expected that SVM classification accuracy for JOLs should be reasonably high, specifically in distinguishing very low JOLs (i.e., 0%-20%) from all other JOLs, similar to the findings of Jersakova et al. (2017).

The second set of the aforementioned SVM analyses instead compares the processes involved in the two different types of confidence judgments. If RCJs and JOLs involve different processes then SVM classification accuracy should be high, particularly for the 0% and 20% confidence levels. If instead the two confidence judgments use the same processes than SVM classification accuracy should be low at all levels of confidence.

In summary, the purpose of this study is to answer two questions: 1) do JOLs use the dual-process model proposed by Son & Metcalfe (2005) and further expanded upon by Metcalfe & Finn (2008b); and 2) do RCJs operate on the same dual-process model as do JOLs (Gardiner et al., 1998; Hanczakowski et al., 2013; Selmeczy & Dobbins, 2014; Williams et al., 2013), or instead operate solely on target-retrievability/accessibility, as per Dougherty et al (2005) and Koriat et al. (1980)?

Method

Participants

Power analysis to determine the mandatory minimum number of participants for an analysis with a power of 0.8 was conducted using the R package "WebPower" (Zhang & Yuan, 2018). The power analysis indicated that a total of 165 pairs of observations were needed to adequately power a paired sample t-test comparing the LSA cosines for cue-terms and target-terms at any level of confidence, assuming an effect size equal to the smallest significant effect in Jersakova et al. (2017), Cohen's d = 0.22. An additional power analysis indicated 325

observations per group for an independent samples t-test comparing the LSA cosines for JOL and RCJ cue/target terms at specific levels of confidence. The study was designed such that each participant gave approximately six justifications at each level of confidence (three justifications per confidence level per block), and there are two conditions, so a minimum of 110 participants total, or 55 participants in each condition, providing a total of 660 justifications per confidence level (330 per condition) are needed for a power of 0.8. To allow for a small number of possible invalid or unusable responses, a total of 160 participants were included in the study. Keep in mind that the level of analysis for this study is justifications, not participants. 160 participants, each providing three justifications per confidence level per block for two blocks should result in approximately 480 justifications per confidence level per condition, well in excess of the minimum of 325 justifications needed for a power of 0.8 for the between subjects comparisons and the 165 pairs of justifications needed for a power of 0.8 for the within subjects comparisons. Participants were sourced from the Prolific.co participant pool and were remunerated $8.00 for their time, as the study took slightly less than one hour to complete and Prolific.co has a required minimum hourly rate of $8.00 per participant.

A total of 226 participants expressed interest in and began the study. Of these 226 participants, 160 completed the study, an additional 10 did not provide consent, 51 completed less than half of the questions and were thereby excluded from analyses, and 5 started the study but ceased participating before completing the first phase of the study, and indicated either technical issues that prevented them from doing so (n=4), or stopped due to extreme boredom (n=1).

Participants were young adults (Mean Age = 22.33, SD = 1.99, Range = 18-25) who resided in the United States of America and were fluent in English. Of the 160 participants who

completed the study, 62 (38.75%) identified as female, 80 (50.00%) identified as male, 4 (2.50%) identified as non-binary, 1 (0.63%) identified as a transman, 1 (0.63%) identified as a transwoman, 1 (0.63%) identified as agender, and an additional 10 (6.25%) participants did not report their gender. There was some variety in participant highest educational attainment, with the largest plurality of participants having completed a bachelor's degree (45, 28.13%), high school (43, 26.88%), or some college (40, 25.00%), and a smaller number of participants reporting having an associate's degree (9, 5.63%), a master's degree (5, 3.125%), less than high school (2, 1.25%), a GED (2, 1.25%), Trade School (1, 0.63%), or did not report their education level (13, 8.13%).

Materials

Stimuli used are taken from the same source as those used in Jersakova et al. (2017). They are from a list of 555 common singular English nouns (5-6 letters in length) from the English Lexicon Project (minimum log Hyperspace Analogue to Language frequency 8.026; Balota et al, 2007). These 555 words were then divided into two lists, a list of words to be used as cues and a list of words to be used as targets. Each participant was exposed to 60 cue-target word pairs, 30-word pairs per block for two blocks. Word pairs were randomly selected for each participant by randomly choosing a word from the cue list and the target list and then presenting them together (Selmeczy & Dobbins, 2014; Jersakova et al, 2017); word pairs were not controlled for associative strength – MC judgments are substantially higher for related than for unrelated word pairs (Hertzog, et al., 2002; Mueller et al., 2013). This study requires justifications for all levels of confidence - by allowing associative strength (i.e. relatedness) to vary across word pairs there was a greater chance that participants would respond with a greater variety of confidence levels.

All confidence judgments, both JOLs and RCJs, were made on a 6-point scale ranging from 0% to 100% in increments of 20%. JOLs were prompted with the first word of the cue-target pair and the question, "How confident are you that in about ten minutes from now you will be able to recall the second word of this item when prompted with the first?". The JOL scale anchor point for 0% is "definitely won't recall" and for 100% is "definitely will recall". For RCJs, participants were prompted with the first word of the cue-target word pair and the question, "How confident are you that the reply you gave for this item is correct?". The RCJ scale anchor point for 0% is "definitely not correct" and for 100% is "definitely correct" (Dougherty et al., 2005). All confidence justifications were prompted by the statement 'Please describe in as much detail as possible why you chose this confidence level" (Jersakova et al., 2017; Selmeczy & Dobbins, 2014).

Demographic questions asking for the participant's age, gender, ethnicity, and highest completed level of education were included. They occurred after the participant completed both phases of the procedure, immediately prior to debriefing.

Procedure

The study was designed using PsychoPy (Peirce, Gray, Simpson, & Macaskill, 2019) and Pavlovia (www.pavolvia.org) and administered online over Prolific (www.prolific.co), an online data collection platform. The MC tasks are constrained by the number of word-pairs a participant can reasonably be expected to commit to memory and provide justifications for in a single ~45 minute to 1 hour sitting (the timeframe here is determined by that used in Jersakova et al., 2017, Selmeczy & Dobbins, 2014). Participants were randomly assigned to one of two study conditions – a condition where they made a judgment of learning (JOL) at initial judgment (N = 78, 48.75%), and a condition where they performed a recall test and then made a retrospective

confidence judgment (RCJ) at initial judgment (N = 82, 51.25%). Participants then completed the study in two blocks, each block consisting of three phases, with no delay between blocks or phases. Phase 1 of each block consists of studying 30 cue-target word pairs, each presented for 6000 ms (Jersakova et al, 2017; Selmeczy & Dobbins, 2014) with a fixation cross for 500 ms between each trial. Following the study phase, participants were presented with the cue of each pair and asked to provide either 1) a JOL predicting performance for the target on the subsequent recall memory test (for those participants in the JOL condition, N=78); or 2) the target associated with that specific cue (i.e. they completed the recall memory test), followed by an RCJ assessing their performance on the retrieval task (for those participants in the RCJ condition, N=82). Participants then completed a final recall memory test for all items, where upon presentation of each cue they recalled the cue-matched target from the word pair, and then provided an additional RCJ assessing their performance on this final recall test. The order of items in each phase was randomized, and the items used in each block were different.

Confidence for both JOLs and RCJs was measured on a 6-point scale (0-1-2-3-4-5) whereby each scale point corresponds to a confidence rating between 0% and 100% inclusive (0-20-40-60-80-100%). Average confidence ratings were calculated on a scale of zero through five, but can be converted into their percentage equivalents by dividing by five.

On a subset of judgment trials, immediately after giving a MC judgment, participants were asked to justify their previously rendered MC judgment using a written keyboard-entered (i.e. typed) response. Participants provided a maximum of 9 justifications per block (18 total), with no more than 3 justifications per MC judgment response per participant (Jersakova et al, 2017). No justifications were asked for the first five trials of either block. Justifications were not

asked for response options for which 3 justifications had been solicited, within blocks. Please refer to Figure 1 in Appendix A for a flow chart of the procedure for the study.

Results

Data Preparation

Text Data Preparation

Spelling errors were corrected for all words in each justification. Contractions were completed (e.g., don't was replaced with do not) (Selmeczy & Dobbins, 2014; Jersakova et al, 2017). In keeping with Selmeczy and Dobbins (2014) procedure, words were not reduced to their stems, as it is possible that different tenses of words appear at different frequencies across the different levels of confidence, and these different tenses may differentially reflect familiarity as opposed to conscious recollection. Likewise, stop-words were not removed for the same reason, except for in the SVM analyses, where their removal was necessitated by limitations of computer processing power. Additionally, all punctuation was stripped from the justifications and all text was converted into lowercase format to avoid any case mismatch issues that could arise from the text data analysis in either the LSA algorithm or in R.

General Data Preparation

A total of 226 .csv files containing participant data were imported from the study site hosted on pavlovia.org. Of these 226 data files, 66 files were either completely blank or missing more than 50% of all data and were excluded from further handling and analysis. The remaining 160 .csv files, each containing one participant's data, were aggregated into a single .csv file containing each participant's ID code, the metamemorial judgment condition to which they were assigned, and then a row for each trial the participant engaged in, with columns each of the following: the cue word, the target word, the cue word again, the participant's response as to the

target word (if the participant was in the RCJ condition), the participant's confidence rating in response to the metamemorial judgment prompt at initial judgment, the participant's written justification for their confidence rating at initial judgment, the cue a final time, the participant's response as to the target word for the final test, the participant's confidence rating in response to the metamemorial judgment prompt at final test, the block to which the trial belonged, and finally the participant's reported age, gender, and highest level of education completed. Each participant completed 30 trials per block, and a total of two blocks for the study, so each participant had a total of 60 rows of data in the .csv file used for analysis, for a total of 9600 total observations. All this data was transferred manually from the individual .csv files to the single larger .csv file used for analysis.

Data for target responses at both initial judgment and final test were scored based on whether the response participants provided matched with the target presented with the cue to which they were responding. Scoring was done using a binary scheme, with correct responses scored as "1"s and incorrect responses scored as "0"s.

Analyses

Two separate sets of analyses were conducted to assess group differences in confidence and performance – first using all 60 responses from each of the 160 participants, for a total sample of 9600 observations, and a second set of analyses were run on a reduced sample that only included the 3036 observations that contained justifications for the confidence judgments. The reduced sample is the one used in the later LSA, n-gram word frequency, and SVM analyses. All participants provided at least one justification, and possibly up to thirty justifications.

Confidence and Test Performance - Overall Sample

Confidence

For the overall sample, participants in the JOL condition reported an average confidence at initial judgment that was on the lower middle end of the confidence scale (M = 1.79, SD = 1.92), close to the 40% confident mark. For these same participants, average confidence at final test was somewhat lower (M=1.62, SD =2.01), in between the 20% and 40% confident mark. Confidence was not normally distributed, so change over time in confidence was assessed with a Wilcoxon signed-rank test. The decrease over time in confidence from initial judgment to final test for participants in the JOL condition is significant, V =1298660, $p < .001$, n=4680.

Participants in the RCJ condition, on the other hand, had an average confidence at initial judgment just slightly above the 40% confidence mark (M=2.07, SD = 2.11). This confidence rating was also around the 40% confidence mark at final test (M=2.14, SD = 2.12). Once again, confidence was not normally distributed, so a Wilcoxon signed-rank test was utilized to assess change over time. For the RCJ condition, confidence did increase significantly over time, V = 656269, $p < .001$, n=4920.

Confidence ratings did differ between the JOL and RCJ conditions. This difference was evaluated by Mann-Whitney-Wilcox tests as, once again, confidence was not normally distributed. At the initial confidence judgment, confidence in the RCJ condition was significantly higher than it was in the JOL condition, W = 10812400, $p <.001$, n_{RCJ} = 4920, n_{JOL} = 4680. Likewise, at the final test, confidence was higher in the RCJ condition than it was in the JOL condition, W = 9912582, $p <.001$, n_{RCJ}=4920, n_{JOL}=4680.

Test Performance

Participant performance at final test was also examined. Due to how the final test score was determined - as a series of individual binary scores for each trial - differences in score by

condition are assessed by chi-squared test. Participants in the JOL condition had a lower proportion correct (27.4%) on the final test than did participants in the RCJ condition (33.6%). This difference in proportion correct was significant, $X^2(1) = 44.0692$, $p < .001$. Together, these findings indicate that participants in the RCJ condition both performed better and were more confident in their performance than were participants in the JOL condition.

Metamemorial Resolution

Relative accuracy of metamemorial confidence judgments, i.e., resolution, was assessed by Goodman-Kruskal Gamma correlation (Koriat, 1997; Koriat et al., 2002; Nelson, 1984). Gamma correlations were specifically used to look at relative accuracy of initial MC judgments as related to final test performance. Please note that both the JOL and RCJ conditions make RCJs at their final test MC judgment. Participants in the JOL condition demonstrated very high relative accuracy of their MC judgments, $\gamma = .87$, SE =.01, 95% CI = [.86,.89]. Participants in the RCJ condition also demonstrated very high relative accuracy of their MC judgments, $\gamma = .90$, SE = .01, 95% CI = [.88,.91]. Gamma correlations for both conditions were significant at the $p < .001$ level. These findings indicate that participants in both MC conditions were generally correct in their predictions of whether they would be able to (or had) successfully recall(ed) the correct target when presented with the cue for that target.

Reduced Sample

The 160 participants provided a total of 3089 justifications for their confidence ratings. Of these 3089 observations, 53 were either empty strings (i.e. they did not contain any text), or they were in response to cue/target word pairs where either the cue or target word were not included in the corpus of words the LSA utilized for its analysis – thereby precluding their use in the LSA analyses, and as such were excluded from the reduced sample. The resulting 3036

observations and justifications comprise the reduced sample used in the LSA, n-gram frequency, and SVM analyses, and as such are subject to a separate set of analyses identical to the ones run above on the full 9600 observation sample.

Confidence

As with the overall sample, confidence in the reduced sample was also not normally distributed. As such, comparisons of confidence between MC judgment conditions were conducted using Mann-Whitney-Wilcox tests; comparisons of confidence within a MC judgment condition over time were conducted using Wilcoxon signed-rank tests. Confidence at initial judgment for the JOL condition (M = 2.24, SD=1.83) and the RCJ condition (M=2.34, SD = 1.93) did not differ significantly, $W = 1120727$, $p = .1129$, $n_{JOL}=1444$, $n_{RCJ}=1592$. Confidence at final test, on the other hand, was significantly higher in the RCJ condition (M=2.46, SD=2.04) than it was in the JOL condition (M= 1.96, SD = 2.06), $W = 988330$, $p <.001$, $n_{JOL}=1444$, $n_{RCJ}=1592$. The decrease in confidence over time for the JOL condition was significant, $V = 216099$, $p <.001$, n=1444. The increase in confidence over time for the RCJ condition also was significant, $V = 133174$, $p < .001$, n = 1592. These findings exhibit the same pattern as do those of the full sample, so we can be confident that there are no systematic differences in confidence values between the full and reduced sample.

Performance

As with the full sample, participant performance at final test was examined. Due to how the final test score was determined, as a series of individual binary scores for each trial, differences in score by condition are assessed by chi-squared test. Participants in the JOL condition had a lower proportion correct (32.1%) on the final test than did participants in the RCJ condition (37.5%). The difference in proportion correct was significant, $X^2(1) = 9.359$, p

<.01. These results once again indicate that participants in the RCJ condition performed better at final test than did participants in the JOL condition. We would expect, and indeed hope for this to be the case, as participants in the RCJ condition were tested on the same cue-target pairs twice – once prior to the initial confidence judgment and once at the final test, as opposed to just making a confidence judgment and then taking the test.

Metamemorial Resolution

Participants in the JOL condition of the reduced sample demonstrated a high relative accuracy of their MC judgments, $\gamma = .78$, SE = .01, 95% CI = [.75,.81], $p < .001$. Participants in the RCJ condition demonstrated a slightly higher relative accuracy of their MC judgments, $\gamma = .83$, SE = .01, 95% CI = [.81,.85], $p < .001$. These findings once again indicate that participants in both MC judgment conditions were able to accurately distinguish successful and unsuccessful memory performance - between recalled and unrecalled words for the RCJ condition and between words that will be recalled and words that will not be recalled for the JOL condition.

Latent Semantic Analysis

Three types of text data analyses were used, paralleling the design of Jersakova et al, (2017). The first text data analysis was Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998), which allows for determining whether justifications are referring to the cue or target word of the word pair stimuli. LSA evaluates semantic relationships between single terms and bodies of text. It does so via singular value decomposition – LSA creates a mathematical matrix that represents a body of text, and then maps the semantic relationships between single words and sets of words. The mapping used in LSA relies on the frequency of co-occurrence as well as a weighting function of the importance of a term for a specific body of text (Jersakova et al., 2017; Landauer, Felt, & Laham, 1998). LSA is trained on

a set of texts to create a semantic space that can be used for the mapping and weighting functions. There is an online LSA tool (http://wordvec.colorado.edu/) that is freely available and offers a semantic space with 300 factors (Dennis, 2006) – this LSA tool was used by Jersakova et al. (2017) in their study. As in Jersakova et al. (2017) this study made use of this tool to classify the semantic similarity between MC justifications and cue/target terms for each justification. The tool returns a cosine value for each classification in the range of -1 to 1, with a value of 1 indicating perfect similarity and a value of 0 indicating the absence of similarity. As per Jersakova et al. (2017) and Wandmacher, Ovchinnikova, and Alexandrova (2008), negative cosine values were set to 0, because a value more dissimilar than "not similar at all" is not interpretable.

The cosine values for cue-justification pairs and target-justification pairs were then compared via Wilcoxon signed-rank tests at each level of confidence judgment for both JOL and RCJ confidence judgments (Jersakova et al., 2017). Additionally, the cosine values for cue-justification and target justification at each level of confidence for JOLs were compared via Mann-Whitney Wilcox ranked sum tests with the same term type cosine values for RCJs at the same confidence level (i.e. the 0% JOL cue-justification cosine was compared with the 0% RCJ cue-justification cosine). This comparison indicates whether semantic similarity for cues/targets differs between judgment types at any given level of confidence. Significant differences in cue-justification cosines at the 0% and 20% confidence levels are indicative of different processes being used by the different types of confidence judgments. Additionally, all group comparisons of similarity cosines were assessed via Bayes factor in addition to NHST to determine the weight of evidence for cue as opposed to target similarity at each level of confidence for each type of confidence judgment. Bayes factors were run using the 'BayesFactor' package in R (Morey &

Rouder, 2021). Each written justification refers to a specific memory, so the semantic similarity is likely to be low, unless the justification specifically mentions either the cue or the target (Jersakova et al., 2017). Increases in LSA cosine score might also occur even in the absence of a justification directly mentioning a cue or target if the justification instead includes partial semantic information about the cue or target (Jersakova et al, 2017; Laham, 1997).

LSA specific data preparation and procedure

A subset of the overall data, containing only the cue, target, and justification provided for the confidence judgment in response to the cue, was created for submission to the LSA tool. The LSA tool requires that all data be submitted in long format lists, with each piece of text separated by a hard return, and can only handle ~200 pieces of text at a time. A total of 3089 justifications exist in the dataset, each accompanied by their own cue and target, for a total of 9267 separate entries that had to be grouped into sets of ~200 and run through the LSA tool. Each set of ~200 entries, once submitted to the LSA, would return an ~200x200 matrix containing the cosine similarity between all possible pairings of entries. A total of 54 of such matrices were generated, and the specific cosine similarity values for the cue-justification pairing, the target-justification pairing, and the cue-target pairing were extracted and entered into the dataset used for analysis. It should be noted that there were a total of four words that were present as cues in the cue-target pairings that did not exist in the greater LSA corpus, and as such no cosine similarity values could be calculated on those pairings. As such, these four words: wombat, intro, venue, and reggae, and all trials containing them, were removed from analyses.

LSA Results

The main goals of the LSA analyses are to answer two questions: 1) How does cue cosine similarity compare to target cosine similarity at each level of confidence within each MC

judgment? and 2) How do cue and target cosine similarity within a given level of confidence differ between MC conditions? Answering each of these questions will help paint the picture of exactly how confidence utilizes cue and target information as confidence develops. It will also allow for a direct comparison of cue and target information utilization between JOLs and RCJs. To compare cue and target cosine similarity at each level of confidence within each MC judgment condition, the non-parametric equivalent of paired samples t-tests, Wilcoxon signed-rank tests, were used. Table 1 has the descriptive and inferential statistics for each of these comparisons. For the JOL MC condition, the 0%, 20%, and 40% confidence levels all had higher cue cosine similarity than target cosine similarity, whereas the 60%, 80%, and 100% confidence levels did not have significantly different cue and target cosine similarity. For the RCJ MC condition, the 0%, 20%, and 40% confidence levels once again had higher cue cosine similarity than target cosine similarity, but unlike in the JOL condition, the 100% confidence level also had higher cue cosine similarity than target cosine similarity. The 60% and 80% confidence levels for RCJs did not have significantly different cue and target cosine similarity.

These analyses were also conducted using the Bayesian equivalent of the paired t-test. Interpretations of the Bayes factors are derived from work by Dienes (2014), who states that Bayes factors greater than 3 are evidence in favor of the alternative hypothesis, Bayes factors less than 1/3 are evidence in favor of the null hypothesis, and Bayes factors between 1/3 and 3 are insensitive to the null or alternative hypothesis. The pattern of results was generally the same for the RCJ cue-target justification comparisons, with the tests supporting the alternative hypothesis that the difference in cosine similarity was not zero for confidence at 0%, 20%, 40%, and 100%, but supporting the null hypothesis that the difference in cosine similarity was zero for confidence at 60% and 80%. Results for the JOL cue-target justification comparisons, on the

other hand, were slightly different from the frequentist inferential results, and more in line with the findings of Jersakova et al. (2017) and the theory posited by Metcalfe & Finn (2008b). There was strong evidence for the alternative hypothesis that the difference in cosine similarity was not zero at the 0% JOL confidence level, weak evidence in favor of the alternative hypothesis, or perhaps insensitive evidence at the 20% JOL confidence level, weak evidence in favor of the null hypothesis at the 40% JOL confidence level, and then strong evidence in favor of the null hypothesis at the 60%, 80%, and 100% JOL confidence level. Please refer to Table 1 for the specific Bayes Factors for each comparison.

Table 1

Cue-Justification and Target-Justification LSA scores by MC condition and Confidence Level

		Cue LSA score		**Target LSA score**		**Test Statistics**			
MC Condition	**Confidence Level**	***Mean***	***SD***	***Mean***	***SD***	***V***	***p***	***n***	***Bayes Factor***
JOL	0%	.21	.13	.17	.09	37824	.00000588*	355	39621.41
	20%	.20	.12	.17	.11	19762	.0052*	265	3.57
	40%	.21	.12	.18	.15	12528	.00029*	203	.66
	60%	.20	.12	.19	.15	8527.5	.070	180	.14
	80%	.21	.16	.21	.16	8404.5	.440	184	.09
	100%	.22	.15	.23	.17	16226	.638	257	.08
RCJ	0%	.21	.13	.16	.09	53724	.00000001*	413	7451565
	20%	.21	.12	.17	.10	24862	.000015*	280	1005.37
	40%	.22	.13	.18	.10	8507.5	.0008*	166	35.44
	60%	.22	.13	.20	.13	7900	.157	173	.31
	80%	.24	.15	.21	.12	12337	.050	209	.65
	100%	.25	.15	.21	.11	38058	.000001*	351	19916.06

*Note: *indicates a significant difference at the p <.05 level.*

The second question of interest for the LSA analyses is whether cue and target cosine similarity for a given level of confidence differ between MC conditions. As both cue and target cosine similarity are non-normally distributed, this question was answered through a series of Wilcoxon rank-sum tests. For cue-justification cosine similarity, JOLs and RCJs did not differ significantly at confidence levels of 0%, 20%,40%, or 60%. JOLs and RCJs did differ at

confidences levels of 80% and 100%, however, with RCJs having higher cue-justification cosine similarity than did JOLs. Target-justification cosine similarity did not differ between JOLs and RCJs at any level of confidence.

As with the previous LSA analyses, these analyses were repeated using Bayesian t-tests, please see resultant Bayes factors in Table 2. The null hypothesis for all tests was no difference in cosine similarity between JOL and RCJ conditions, and the alternative was that there was a difference. For cue-justification comparisons, there was evidence for the null hypothesis for confidence levels of 0%, 20%, and 40%, and evidence insensitive to either the null or alternative hypothesis for confidence levels 60%, 80%, and 100%. For target-justification comparisons, there was evidence in support of the null hypothesis for 0%, 20%, 40%, 60%, and 80% confidence levels. Evidence was insensitive to null and alternative hypotheses for the 100% confidence level. These results are similar in pattern to those of the frequentist analyses, but generally a bit more conservative. Overall, they suggest that there is little to no meaningfully different utilization of either cue or target information at any level of confidence between JOLs and RCJs – except for potentially more cue information utilization at higher levels of confidence for RCJs.

Table 2

Cue-Justification and Target-Justification LSA scores by MC condition and Confidence Level

Cue/ Target Cosine Similarity	Confidence Level	JOL Cosine Similarity Score		RCJ Cosine Similarity Score		Test Statistics				
		Mean	*SD*	*Mean*	*SD*	*W*	*P*	*nJOL*	*nRCJ*	*Bayes Factor*
Cue	0%	.21	.13	.21	.13	73389	.979	355	413	.084
	20%	.20	.12	.21	.12	34821	.215	265	280	.169
	40%	.21	.12	.22	.13	16068	.444	203	166	.213
	60%	.20	.12	.22	.13	14140	.136	180	173	.425
	80%	.21	.16	.24	.15	16829	.033*	184	209	.457
	100%	.22	.15	.25	.15	39890	.015*	257	351	1.370
Target	0%	.17	.09	.16	.09	76558	.289	355	413	.138

20%	.17	.11	.17	.10	36866	.899	265	280	.096
40%	.18	.15	.18	.10	15638	.235	203	166	.116
60%	.19	.15	.20	.13	13728	.055	180	173	.203
80%	.21	.18	.21	.12	17612	.150	184	209	.115
100%	.23	.17	.21	.11	47005	.374	257	351	.810

*Note: *indicates a significant difference at the p <.05 level. The descriptive statistics and Ns are identical to those in Table X, they are just arranged in a manner that allows for more easy comparison between MC conditions for a given justification cosine similarity and confidence level.*

N-gram Frequency Analysis

The second text data analysis is an n-gram frequency analysis, which allows for isolating unique phrases that occur with differing frequencies at different levels of confidence as compared to the other levels of confidence. N-gram analysis compares bodies of text by counting the frequency of occurrence of each n-gram across all bodies of text. As per Jersakova et al. (2017), MC justifications only contribute up to one time for any given n-gram count so as to account for some participants writing more than others. In other words, if a specific word or n-gram is present more than once in the same justification, it will only count towards that word/n-gram one time. This also allows for computing the total number of justifications that contain a given n-gram in each MC judgment category. Words and phrases identified as important in Jersakova et al. (2017) and Selmeczy and Dobbins (2014) were given particular attention and included in the analyses – other n-grams that have an occurrence of at least 10 times across all justifications (Jersakova et al. 2017) were also included. N-gram analysis typically involves a binomial test and computes a *p*-value for the proportion of occurrence of an n-gram under a particular response category, assuming a binomial distribution with a *p*-parameter of 0.5 (Jersakova et al., 2017; Selmeczy & Dobbins, 2014). The specific methodology for the n-gram analysis in this study for comparisons within a specific type of confidence judgment, on the other hand, follows the method used in Jersakova et al. (2017), as their n-gram analysis involved comparing multiple categories of MC judgments. This n-gram analysis involves contrasting each

MC judgment category against all other MC judgment categories. The *p*-parameter, instead of being set to 0.5, which would indicate an equal probability of occurrence for each of two categories, was set to the 1/6 for any given category based on the presumption that each n-gram is equally likely to occur at all levels of confidence. This allowed for calculating, for a specific MC judgment category, whether the proportion of occurrence of any n-gram was significantly different from it having equal probability in all MC judgment categories. Additional analyses examining the frequency of words or phrases indicating cue-familiarity, or a lack thereof, for 0% and 20% confidence judgments in JOLs as opposed to RCJs were conducted, but with the *p*-parameter for the binomial test set to 0.5, thereby testing whether terms indicative of cue-familiarity are equally or differentially likely to occur at low confidence for both types of confidence judgments. A similar analysis was conducted on words or phrases indicating recollective experiences at the 0% and 20% confidence level for JOLs and RCJs, this time comparing the relative amount of target accessibility present in these confidence judgments. If the process involved in JOLs and RCJs differs, the relative frequency of terms indicating recollective experience present in the justifications for these levels of confidence should also differ.

N-gram frequency analysis specific data preparation and procedure

The "tokenizers" package in R (Mullen, Benoit, Keyes, Selivanov, and Arnold, 2018) was used to convert all participant justifications into three separate sets of tokens – unigrams (i.e. words), bigrams, and trigrams. This procedure was completed six times – once for the overall dataset, and then once each for just the JOL and just the RCJ metamemorial judgment conditions, and then both with and without removing stop-words. It should be noted that what words are considered stop-words is entirely context dependent, so using a pre-determined list of commonly used stop-words would not work in this particular case – there is a large difference in

the phrase "can *not* remember" and "can remember" – not is often considered a stop-word, but in the case of this study it completely changes the meaning of the n-gram. As such, a custom list of stop-words was generated (Appendix D), by taking the 31 most frequently occurring words that did not contain informational value relevant to the study. The n-gram frequency analyses themselves were conducted on the set of words/n-grams that contained these stop-words, as per Selmeczy & Dobbins (2014). However, for the SVM analyses, the set of words/n-grams with stop-words removed were used, as was necessitated by computing power constraints.

N-gram Analysis Results

Similar to the LSA analyses, there were two specific questions that the n-gram analyses were designed to answer. First, within a given MC condition, are there more words/phrases that are indicative of cue familiarity, or an absence thereof, at lower levels of confidence, and more words/phrases indicative of at least attempts of target retrieval or partial target information as confidence increases? This pattern of results would suggest that both cue familiarity and target accessibility are at play in the formation of confidence. Second, are there differences in frequency of words/phrases across MC judgment conditions? This finding could be indicative of either different processes or different utilization of the same processes used to form confidence.

N-gram analyses addressing the first question are presented in Appendix B, Table 3, for JOLs, and Table 4, for RCJs. Please note, all n-gram analyses were conducted without removing stop-words, as described above. Tables only include n-grams that occur at least 10 times across all justifications. As was the case in both Selmeczy & Dobbins (2014) and Jersakova et al. (2017), unigrams were not included in the results due both to their overwhelmingly large numbers and difficulty interpreting them due to lack of context. Additionally, the tables only display n-grams that have a proportion significantly higher than would be expected if the n-gram

appeared equally at all levels of confidence, as differentiating each level based on what *is* characteristic of that level of confidence, as compared to the other levels of confidence, is necessary for answering the research questions of interest. A full table of all n-grams can be found in the supplementary materials at https://osf.io/2h6ae/.

N-gram Results - JOLs

Although the LSA analyses indicate that both the 0% and 20% JOLs refer more to the cue than the target, semantically, the n-gram analysis indicates a difference in *how* that referencing occurred. As per Table 3, the 0% JOL confidence level can perhaps best be characterized by the inability to remember ("cannot recall", "cannot remember"). As per Jersakova et al. (2017), this can be interpreted as expressing lack of cue familiarity as the participant cannot even remember having seen the cue ("not remember it", "not remember seeing", "not remember this"). The 20% JOL confidence level also has some evidence of inability to remember ("cannot recall", "cannot remember"), but also has evidence to the contrary ("I remember seeing", "recall seeing"), indicating some level of cue familiarity. Additionally, language present at this level of JOL confidence expressed uncertainty ("not sure", "am not sure"). So, whereas the 0% confidence level and 20% confidence level both contained more cue- than target- information, the actual information content of each confidence level differed. The 40% JOL confidence level was predominantly characterized by varying levels of uncertainty and hedging ("do not really", "I believe", "I might", "not really", "not sure"), and is indicative of both the presence of cue familiarity and partial target access. The 60% JOL confidence level can be characterized by uncertainty as well, but more towards ability to correctly retrieve the target ("could remember", "I may", "think I know", "think I remember") and in some cases even relative certainty that the participant will successfully retrieve the target ("I will remember"). These sentiments suggest at

least some amount of target access. The 80% JOL confidence level is characterized by even more certainty in the ability to correctly retrieve the target ("can recall", "can remember", "I remembered"), but still has some uncertainty present ("I think", "think I know"), albeit less uncertainty than was present at the 40% and 60% JOL confidence levels. The 100% JOL confidence level is best characterized by certainty in the participant's ability to correctly retrieve the target when presented with the cue ("I know", "I recall", "I remember it", "it was", "remember this pair"), suggesting access to the target. These findings replicate those of Jersakova et al. (2017) and match the pattern specified by Metcalfe and Finn (2008b), who suggested that the 0% JOL is best characterized by lack of cue familiarity, and that JOLs above that level are best characterized by both the presence of cue familiarity and increasing levels of target access as confidence increases.

N-gram Results - RCJs

As per Table 4, the overall pattern of results for the RCJ n-gram analysis is similar to the JOL results, with a few notable differences. The 0% RCJ confidence level is once again best characterized by an absence of cue familiarity ("cannot remember", "do not recall", "have no idea", "I did not", "I forgot", "not remember seeing", "not remember this"). The 20% RCJ confidence level likewise contained somewhat mixed evidence for the participant's ability to recognize the cue ("I cannot", "I do not", as opposed to "I remember seeing"), as well as uncertainty about their ability to successfully retrieve the target ("am not confident", "am not sure", "I guessed", "sort of"). The 40% RCJ confidence level is characterized by uncertainty ("am not sure", "am unsure", "I believe", "could be"), as is the 60% RCJ confidence level ("I think", "think I remember"). The 80% RCJ confidence level, as was the case with the 80% JOL confidence level, can be characterized by relative certainty in the participant's ability to correctly

retrieve the target when presented with the cue ("am pretty sure", "remember thinking"). Finally, the 100% RCJ confidence level, as was the case with 100% JOL confidence level, can be characterized by certainty that the participant correctly retrieved the target ("am sure", "because I remember", "I know this", "I remembered this"). However, unlike the 100% JOL confidence level, the 100% RCJ confidence level also contains evidence that participants had created associations between the cue and target and utilized those associations to help them correctly retrieve the target when presented with the cue ("a connection", "remember this pairing", "reminded me of"). This finding suggests that participants made use of associative information in their confidence judgments, even when not instructed to do so (Hertzog et al., 2014; Jersakova et al., 2017). Together with the JOL specific analysis, this analysis provides an answer to the first of the two questions the n-gram analysis was designed to answer. It appears that both JOLs and RCJs have words/phrases that indicate cue familiarity (or lack thereof) at lower levels of confidence, and language indicating increasing levels of target access as confidence increases.

N-gram Results – Comparing JOLs & RCJs

The second question of the n-gram analysis was whether there were differences in the frequency of occurrence of specific n-grams across MC conditions. To answer this, it is necessary to examine three separate sets of data. The first is to examine n-grams that occurred in both the JOL and RCJ conditions but had a proportion of occurrence that was significantly higher in one condition or the other (Appendix B, Table 5). The second is to look at n-grams that only occurred in one MC condition or the other, but not both (Appendix B, Table 6). Finally, it would be potentially helpful to look at the relative frequency of inclusion of the specific cue or target terms in any given justification and see if those frequencies differ across MC conditions for any given level of confidence.

As per Table 5 in Appendix B, there were n-grams at each level of confidence that were significantly more likely to appear in either JOL or RCJ justifications. For the 0% confidence level, "not remember this" was significantly more likely to appear in RCJ justifications, whereas "remember it" and "to remember" were both more likely to appear in JOLs. It should be noted, however, that the bigram "remember it" that appeared in JOLs was likely part of the larger trigram "*not* remember it", which dramatically changes the meaning of the n-gram and does not differentially indicate remembering of the target between JOL and RCJ conditions at the 0% confidence level. Instead, this indicates that the word "it" was used predominantly in the JOL condition at 0% confidence, whereas in RCJs the equivalent phrase was more commonly stated as "not remember this" or "not remember that".

There were six n-grams that occurred at differential rates between the JOL and RCJ conditions for the 20% confidence level. Of those six, two occurred more frequently in the RCJ condition ("am not confident", "the first word"), the remaining four occurred more frequently in the JOL condition ("do not recall", "if I", "not recall", "to remember"). The n-grams unique to RCJs and JOLs for this level of confidence do not provide much information useful in differentiating the two MC judgment types, outside of "am not confident" expressing relative uncertainty for the RCJ condition, whereas "do not recall" in the JOL condition is more of an expression of certainty of absence/inability.

Forty percent confidence judgments only had representation from JOLs, there were no instances of n-grams that occurred more frequently in RCJs than JOLs for this level of confidence. The n-grams that occurred more frequently in JOLs for this level of confidence are primarily characterized by uncertainty ("I can", "if I", "might be") with some additional allusions to the target that did not specifically mention the target word itself ("second word",

"the second word"). There were also no instances of n-grams occurring more frequently in RCJs than in JOLs for 60% confidence judgments. The n-grams that occurred more frequently for JOLs than for RCJs at this level of confidence were slightly less uncertain, but generally in line with those of the 40% confidence level ("I can", "I have", "think I").

Eighty percent confidence justifications had n-grams that occurred more in both JOLs and RCJs. N-grams that occurred more in RCJs included "am pretty sure" and "remember the word", both indicating relative certainty that the target was successfully retrieved. N-grams that occurred more in JOLs for this confidence level still indicated some uncertainty in many cases ("I can", "I think I") but in other cases also demonstrated relative certainty that the target could be successfully retrieved ("I know", "remember it"). Finally, the 100% confidence level predominantly had n-grams that occurred more in RCJs than JOLs ("I remember this", "I thought of", "remember this one", "remember this pair") – these n-grams expressed certainty in successful retrieval of the word and allude to associations between the cue and target. The only n-gram that occurred more in JOLs than in RCJs for the 100% confidence level was "the other", a bigram that was part of the trigram "the other word", referring indirectly to the target of the cue-target pairing.

There were several n-grams that only occurred in either JOLs or RCJs, but not both. At the 0% confidence level, the n-grams "could not remember" and "did not remember" occurred in the RCJ condition, whereas the n-grams "not remember it", "will not", and "too many" occurred only in the JOL condition. Note that some of the difference in n-grams here is the tense – RCJ-exclusive n-grams are in past tense, whereas JOL-exclusive n-grams are in present or future tense. This is to be expected given the temporal direction of each metamemorial judgment. The 20%, 40%, and 60% confidence level had additional examples of this tense difference – JOL-

exclusive n-grams included future or future imperfect momentary tenses "I might", "I will", and "will remember" as opposed to past tense RCJ-exclusive n-gram "I remembered". The 80% confidence level only had RCJ-exclusive n-grams, "I remember thinking" and no JOL exclusive n-grams at all. Finally, the 100% confidence level was predominantly comprised of RCJ-exclusive n-grams, such as "easy to", "I imagined", "I remembered this", and "this pairing". The only JOL-exclusive n-gram for this level of confidence was "word is". A full list of these n-grams, as well as the confidence level they occurred at and their total frequency, can be found in Appendix B, Table 6.

The final analysis for n-gram word frequency analyses is not technically an n-gram analysis but is instead a frequency analysis of the presence of either the cue or target term in the justification. Overall, inclusion of the cue in justifications occurred significantly more frequently in the RCJ than the JOL condition, $X^2(1) = 45.54$, $p < .001$. This difference existed at each level of confidence, please see Table 7 below. There was not an overall difference in inclusion of the target in justifications between MC conditions, $X^2(1) = 2.78$, $p = .10$, nor were there differences at any level of confidence.

Table 7

Frequency of Inclusion of Cue in Justification by Confidence Level and MC Condition

		JOL		RCJ		Test Statistics	
Cue/Target	**Confidence Level**	***Count***	***%***	***Count***	***%***	X^2	***p***
Cue	0%	33	9.3%	65	15.7%	6.55	.011*
	20%	27	10.2%	48	17.1%	4.978	.028*
	40%	22	10.8%	31	18.7%	3.95	.047*
	60%	25	13.9%	57	32.9%	16.92	<.001*
	80%	46	25.0%	78	37.3%	6.32	.012*
	100%	73	28.4%	130	37.0%	4.59	.032*
Target	0%	2	0.6%	0	0.0%	.67	.41
	20%	5	1.9%	6	2.1%	0.00	1.00
	40%	10	4.9%	8	4.8%	0.00	1.00
	60%	26	14.4%	34	19.7%	1.35	.246

80%	52	28.3%	68	32.5%	.65	.419
100%	99	38.5%	133	37.9%	.01	.942

*Note: *indicates a significant difference at the p <.05 level. Results for the 0% and 20% confidence Target Inclusion chi-squared tests were re-ran as Fisher's Exact tests, as there were not sufficient observations in each cell to support a chi-squared test. Results for the Fisher's Exact Test indicated the same as the chi-squared tests, no significant difference in frequency of inclusion of the target at either level of confidence.*

Taken together, the findings of these three analyses tell us that generally, there is not a tremendous difference in specific words/phrases used in JOL as opposed to RCJ justifications. The first analysis results only indicate differences in amount of uncertainty at the 20% and 80% confidence levels, and that the 100% confidence level included references to associations between the cue and target terms in the RCJ but not JOL conditions. Otherwise, the same pattern of absence of cue familiarity, followed by cue familiarity and retrieval failure, and then increasing amounts of target access occurs in both MC judgment conditions. The second analysis really only found a difference in tense used in the different justifications, and this can be explained by JOLs asking for a prediction of future performance as opposed to RCJs asking for a rating of past performance. The biggest difference to come out of these analyses was the difference in inclusion of the cue term in the justification itself, at all levels of confidence, between the MC judgment conditions. The cue term itself was significantly more likely to be included in the justifications for RCJs than the justification for JOLs, at all levels of confidence. This finding is in line with the LSA findings and suggests that RCJs may make use of cue information beyond just cue familiarity more so than do JOLs.

Support Vector Machine

Finally, the third text data analysis is Support Vector Machine (SVM). SVM allows for finding the point at which two or more categories maximally differ from one another. In this case, it should allow for distinguishing between different types of evidence used in justifications

at different levels of confidence. For text classification, each unique word in a document is a feature and different coding schemes can be used to quantify the feature values of words in the document. A binomial scheme was used to indicate whether a word/feature is present or absent in the justifications for each participant's confidence justifications. As such, each justification was represented by a vector comprised of the binary values indicating the presence or absence of all possible n-grams. These vectors were treated as the input to the SVM, and the output was the MC judgment category to which the justification belongs. The SVM algorithm was then trained and tested on a term matrix where each row represented a provided confidence justification, and each column a particular word feature (Selmeczy & Dobbins, 2014). The SVM algorithm attempted to locate a boundary that separates two sets of data by the largest magnitude possible, in this case the two sets of data were pairs of different MC judgment categories for the within confidence judgment type analyses, and the same confidence level judgments from each type of confidence judgment for the between confidence judgment type analyses. To compare pairs of MC judgment response categories, the justification responses for each pair of categories were labeled and combined (Jersakova et al., 2017). SVM was implemented using R statistical software (v 4.1.2, R Core Team 2021) and the packages RTextTools (Jurka, Collingwood, Boydstun, Grossman, & van Atteveldt, 2012) and caret (Kuhn, 2022). The SVM was trained on a testing dataset of 50% of the justifications and then validated on the remaining 50% of justifications. The classifier was trained with a linear kernel and a cost value of .10 (Jersakova et al., 2017; Lateef, 2020; Selmeczy & Dobbins, 2014). Performance was evaluated by the classifier's ability to correctly distinguish the response categories from each other and is measured as overall accuracy.

As per both Jersakova et al (2017)'s findings and Metcalfe and Finn (2008b)'s findings, JOLs at the 0% confidence level should refer to the cue and a lack of familiarity with it, whereas all other levels of confidence should instead refer to the target and provide increasing levels of partial target information. This split should be detected by the hyperplane drawn by SVM at the point the justifications provided for levels of confidence are maximally differentiated.

On the other hand if, as according to Dougherty et al. (2005) and Koriat et al. (1980), RCJ justifications refer exclusively to the target of the cue-target pairing then the SVM should have little success at finding a point of differentiation across the levels of confidence. However, if, as findings from Hanczakowski et al. (2013) and Selmeczy and Dobbins (2014) indicate, RCJs do in fact make use of cue familiarity, this should result in SVM successfully finding a point of differentiation between levels of RCJ confidence based on the source of information referred to by the justification.

Data Handling

Support Vector Machine specific data preparation

To run the support vector machine, it was necessary to generate binary matrices containing columns for each word/n-gram that could appear in any given justification and fill those matrices with binary values (1 or 0) indicating presence or absence of each word/n-gram. This was accomplished by concatenating lists of all words/n-grams from the JOL and RCJ conditions separately, and then an additional list for the combined JOL and RCJ conditions. These lists were then used as indices in a series of for loops that iterated through each participant's justification and detected the presence of the specific word/n-gram used as the index of the loop and then flagged that specific row-column combination as "1" if the word/n-gram was present or "0" if it was absent. This allowed for each justification to be re-interpreted as a vector of 1's and 0's, where each component of the vector represented a specific word or n-

gram, as per Jersakova et al (2017). All materials used in this analysis are available at https://osf.io/2h6ae/.

Support Vector Machine (SVM) Analyses

Table 8 and 9 below have the overall classification accuracy for comparisons within judgment types, specifically for adjacent confidence levels, whereas Table 10 contains the comparisons across judgment types. Please note, there are similar tables in Appendix C which contain sample SVM results for all confidence level comparisons. Those tables contain SVM results from a single iteration of SVM. Tables 8, 9, and 10, on the other hand, provide classification accuracies that are the average accuracy over 5 iterations of each SVM, and are thereby more stable estimates of classification accuracy.

Table 8

Bivariate SVM classification accuracy results by for JOLs

Confidence Level	20%	40%	60%	80%	100%
0%	85.37%				
20%		84.29%			
40%			86.91%		
60%				87.58%	
80%					88.00%

Note: The results here are the percentage of test cases classified correctly by the SVM algorithm and indicate the extent to which two levels of JOL confidence can be differentiated based on the content of their justifications.

Table 9

Bivariate SVM classification accuracy results by for RCJs

Confidence Level	20%	40%	60%	80%	100%
0%	88.03%				
20%		84.04%			
40%			88.76%		
60%				92.11%	
80%					88.89%

Note: The results here are the percentage of test cases classified correctly by the SVM algorithm and indicate the extent to which two levels of JOL confidence can be differentiated based on the content of their justifications.

Table 10

Bivariate SVM classification accuracy results by MC Judgment Type

Confidence	RCJ 0%	RCJ 20%	RCJ 40%	RCJ 60%	RCJ 80%	RCJ 100%
JOL 0%	71.96%					
JOL 20%		84.63%				
JOL 40%			85.43%			
JOL 60%				88.75%		
JOL 80%					87.14%	
JOL 100%						78.61%

Note: The results here are the percentage of test cases classified correctly by the SVM algorithm and indicate the extent to which the same confidence level for the two MC judgment types can be differentiated based on the content of their justifications.

The results for the JOL SVM are interesting, but not entirely in line with what would be expected based on the descriptive model of Metcalfe & Finn (2008b). Their descriptive model of confidence predicts that the greatest difference in use of cue familiarity and target access should occur at the boundary between the lowest level of JOL confidence and the next level up (0% and 20%), indicating a shift from solely using cue familiarity at the lowest level to a mix of cue familiarity and target accessibility at the next level up. This would be reflected in the highest classification accuracy of an SVM occurring at this boundary, and the SVMs at higher boundaries (20-40%, 40-60%, 60-80%, 80-100%) having notable lower classification accuracies. This is also in keeping with the findings of Jersakova et al. (2017), with the highest classification accuracy occurring at the lowest adjacent confidence levels, and classification accuracy decreasing as confidence increased. Further, Jersakova et al. (2017)'s results were such that the classification accuracy of the higher confidence levels was not significantly greater than chance. They took this as indicating that all the higher confidence justifications in their study could be best characterized as using different degrees of the same process, target accessibility, and not distinct processes.

Contrary to the predictions of Metcalfe & Finn (2008) and the findings of Jersakova et al. (2017), the results of this series of SVM analyses were that the 80-100% boundary had the highest classification accuracy of all adjacent JOL confidence levels, and the 20-40% boundary had the lowest of all classification accuracies. Even the 20-40% boundary classification accuracy was significantly greater than chance, $p < .001$, so all classification accuracies for adjacent JOL confidence levels were greater than chance. Further, a proportion test revealed that even the difference in accuracy between the highest accuracy classifier (80-100%) and the lowest accuracy classifier (20-40%) did not reach statistical significance, $X^2(1) = 1.238$, $p = .266$. This indicates that all JOL SVMs were similarly accurate at classifying adjacent JOL confidence levels, a finding that could indicate two different processes, or differential uses of evidence, occurring at each level of confidence.

Similar to the JOL SVM results, all of the RCJ SVMs were quite accurate at classifying justifications into the correct confidence level when determining to which of two adjacent confidence levels said justification belonged. The highest classification accuracy was at the 60-80% confidence boundary, and the lowest classification accuracy was at the 20-40% confidence boundary. Even the lowest accuracy classifier was significantly more accurate than chance, $p < .001$, so all classification accuracies for adjacent RCJ confidence levels were greater than chance. A proportion test revealed that the difference in accuracy between the highest accuracy classifier (60-80%) and the lowest accuracy classifier (20-40%) was statistically significant, $X^2(1) = 5.710$, $p = .017$. However, neither the difference in accuracy between the highest and second highest accuracy classifiers, $X^2(1) = .983$, $p = .321$., the difference in accuracy between the lowest and second lowest accuracy classifiers, $X^2(1) = 1.785$, $p = .182$, nor the difference between the second highest and lowest accuracy classifiers, $X^2(1) = 2.30$, $p = .130$, were

statistically significant. This indicates that most RCJ SVMs were similarly accurate at classifying adjacent RCJ confidence levels. This once again suggests that different processes or differential use of evidence occurring between each level of confidence. Interestingly, the 20-40% classifier was the least accurate classifier in both the JOL and the RCJ condition, suggesting that perhaps these two levels of confidence are most similar to each other in terms of processes involved and evidence used within each MC judgment type.

Finally, the SVMs comparing JOLs and RCJs at the same level of confidence were once again very accurate. The lowest accuracy classifier was the 0% confidence level, at 71.96%. This SVM classified justifications correctly significantly greater than chance, $p < .001$. As such, all classifiers were significantly more accurate than chance. The next lowest accuracy classifier was the 100% confidence level, at 78.61%. Interestingly, although these two classifiers were greater than chance, this finding nonetheless suggests that RCJ and JOL confidence is most similar at the highest and lowest ends of the confidence scale in terms of processes involved and evidence used. The highest accuracy classifier was the 60% classifier, at 88.75%, followed by the 80% classifier (87.14%), the 40% confidence classifier (85.43%) and finally the 20% confidence classifier (84.63%). Proportion tests revealed that although there was no difference in accuracy between the 0% classifier and the 100% confidence classifier, $X^2(1) = 3.54$, $p = .06$, there was a difference in accuracy between the 0% confidence classifier and the 20% confidence classifier, $X^2(1) = 13.431$, $p < .001$. As such, all classifiers with higher accuracy than the 20% confidence classifier are also significantly more accurate than the 0% classifier. However, only the 60% confidence and 80% confidence classifiers were significantly more accurate than was the 100% confidence classifier, $X^2(1) = 5.438$, $p = .020$, and $X^2(1) = 7.005$, $p < .01$, respectively. Ultimately, all classifiers are accurate at higher than chance levels, suggesting that at all levels of

confidence there is something fundamentally different about how JOLs and RCJs operate, even if, as evidenced by the results of the LSA and N-gram analyses, both JOLs and RCJs make use of cue-familiarity and target access in the formation of confidence.

SVM Results – Using only bigrams and trigrams

The results of the JOL SVM analyses are *very* different from those of Jersakova et al. (2017), to the point that they are somewhat concerning. It is possible, however, that this discrepancy in results is not from an actual difference in how confidence was formed between studies, but instead a methodological difference in how the SVMs were constructed. Given the wording in the procedure of Jersakova et al. (2017), it is possible that instead of using every single word or n-gram contained in all justifications as variables in their SVM classification matrix, they used only those n-grams that were also used in their n-grams analysis. All SVMs were re-run on a much smaller dataset, containing only variables for the bigrams and trigrams that occurred at least 10 times across all justifications within a given MC judgment condition. It should be noted that this decreased the available features for the SVM to use in classification by approximately two orders of magnitude. The results of these analyses, both within and between the MC judgment conditions are detailed below in Tables 11, 12, and 13.

Table 11

Bivariate SVM classification accuracy results by for JOLs – Only N-grams.

Confidence Level	20%	40%	60%	80%	100%
0%	71.11%				
20%		58.68%			
40%			59.56%		
60%				55.25%	
80%					60.27%

Note: The results here are the percentage of test cases classified correctly by the SVM algorithm and indicate the extent to which the same confidence level for the two MC judgment types can be differentiated based on the content of their justifications. Additionally, only bigrams and trigrams with a frequency of at least 10 across all justifications were used as classification variables.

The pattern of results for these SVMs is considerably more in line with the findings of Jersakova et al. (2017). The SVMs of adjacent JOL confidence levels revealed that the highest classification accuracy was between 0% and 20% JOL confidence. The accuracy of the 0-20% JOL classifier was significantly greater than chance, $X^2(1) = 54.69$, $p < .001$. Additionally, the 0-20% JOL confidence classifier was significantly more accurate than the next most accurate classifier, the 80-100% JOL confidence classifier, $X^2(1) = 6.21$, $p < .05$. Unlike the Jersakova et al. (2017) findings, all adjacent JOL confidence classifiers, except for the 60-80% confidence classifier, were significantly more accurate than chance. Only the 0-20% classifier was significantly more accurate than the other classifiers.

Table 12

Bivariate SVM classification accuracy results by for RCJs – Only N-grams

Confidence Level	20%	40%	60%	80%	100%
0%	71.19%				
20%		62.78%			
40%			57.89%		
60%				60.02%	
80%					66.16%

Note: The results here are the percentage of test cases classified correctly by the SVM algorithm and indicate the extent to which the same confidence level for the two MC judgment types can be differentiated based on the content of their justifications. Additionally, only bigrams and trigrams with a frequency of at least 10 across all justifications were used as classification variables.

The pattern of results for the adjacent RCJ confidence level classifiers follow a similar pattern to the JOL classifiers. The 0-20% classifier was the most accurate, followed by the 80-100% classifier. The 0-20% RCJ confidence classifier was significantly more accurate than chance, $X^2(1) = 60.77$, $p < .001$; however, it was not significantly more accurate than the 80-100% classifier, $X^2(1) = 1.44$, $p = .23$. The 0-20% RCJ confidence classifier was significantly more accurate than the third most accurate classifier, the 20-40% confidence classifier, $X^2(1) = 3.93$, p

< .05, as well as the other, less accurate classifiers. The other classifiers did not differ significantly from each other, but all classifiers were more accurate than chance.

Table 13

Bivariate SVM classification accuracy results by MC Judgment Type – Only N-grams

Confidence	RCJ 0%	RCJ 20%	RCJ 40%	RCJ 60%	RCJ 80%	RCJ 100%
JOL 0%	58.96%					
JOL 20%		60.09%				
JOL 40%			55.43%			
JOL 60%				63.64%		
JOL 80%					57.06%	
JOL 100%						59.80%

Note: The results here are the percentage of test cases classified correctly by the SVM algorithm and indicate the extent to which the same confidence level for the two MC judgment types can be differentiated based on the content of their justifications. Additionally, only bigrams and trigrams with a frequency of at least 10 across all justifications were used as classification variables.

The SVMs comparing JOLs to RCJs within a given level of confidence also had noticeably different results when run only including the bigrams and trigrams used in the N-gram analysis as inputs. The four most accurate classifiers (60%, 20%, 100%, 0%) were all significantly more accurate than chance, whereas the two least accurate classifiers did not differ from chance, 40% confidence $X^2(1) = 1.96$, $p = .16$ and 80% confidence $X^2(1) = 3.72$, $p = .054$. However, none of the classifiers differed from each other in terms of accuracy, even when comparing the most and least accurate classifiers, $X^2(1) = 2.18$, $p = .14$. This finding does raise doubt over whether any of the classifiers for this set of SVMs are more accurate than chance and suggests that the two MC judgments largely used the same types of evidence at each level of confidence.

Ultimately, the results of the SVM analyses are mixed. The SVMs utilizing the full set of words and n-grams were extremely accurate at all levels of confidence both within and across MC judgment conditions. These findings are not in line with any hypothesis tested by this study

and cannot be easily explained, other than perhaps given sufficient features, any two categories can be effectively discriminated between. However, there is support for both JOLs and RCJs operating following the dual-process descriptive model (Jersakova et al., 2017; Metcalfe & Finn, 2008b), when only including the n-grams used in the n-gram analysis as features for the SVM to classify justifications on. When only the n-grams from the n-gram analysis were used classification accuracy was highest for the 0%-20% confidence boundary for both JOLs and RCJs, suggesting that there are two different processes involved in the formation of confidence for these two different categories. We already know from the LSA analysis that both of these levels of confidence refer more to the cue than the target. Further, we know from the n-gram analysis that the cue referencing in the 0% confidence level is an absence of cue familiarity, whereas the 20% confidence level is better characterized by the presence of cue familiarity, as well as target access failure. This is likely the distinction the SVM picked up on. Additionally, classification accuracy for the other confidence boundaries tended to be much lower than for the 0%-20% boundary – a finding that once again seems to support the dual-process descriptive model (Jersakova et al., 2017; Metcalfe & Finn, 2008b). Higher levels of confidence are best characterized by differing quantities of the same process, target accessibility, according to the model, and the results of the SVM appear to support this proposition. Additionally, the use of associations at the 100% RCJ confidence level appears to have been detected by the SVM as well and is shown via a classification accuracy at the 80% and 100% confidence boundary that is not quite as high as the 0%-20% boundary but is still noticeably higher than all other boundaries.

Discussion

General Discussion

Previous research on metamemorial confidence judgments has provided evidence that judgments of learning (JOLs) operate following a dual-process model (Dougherty et al., 2005;

Jersakova et al., 2017; Metcalfe & Finn, 2008b; Son & Metcalfe, 2005) and that JOLs and RCJs appear to use different processes, at least to some extent (Dougherty et al., 2005; Maki, 2008; Nguyen et al, 2017; Robey et al., 2017; Watier & Collin, 2011). On the other hand, there is also mixed evidence on RCJs also use a dual-process method (Gardiner et al., 1998; Hanczakowski et al., 2013; Selmeczy & Dobbins, 2014; Williams et al., 2013) or instead rely solely on target retrievability (Dougherty et al., 2005; Koriat et al., 1980). This study tested the dual-process model of JOLs and assessed whether RCJs use the same processes as do JOLs or instead based solely on target retrievability/accessibility. Written justifications for confidence judgments for JOLs and RCJs were compared and analyzed using three different text analysis techniques. Results indicate that: 1) participants successfully and effectively justified their metamemorial confidence judgments; 2) JOLs followed the dual-process descriptive model that differentially referenced cue and target information at different levels of confidence (Jersakova et al., 2017; Metcalfe & Finn, 2008b; Son & Metcalfe, 2005); 3) RCJs also generally followed the same dual-process descriptive model and differentially referenced cue and target information at different levels of confidence; and 4) although JOLs and RCJs generally operated in accordance with the descriptive dual-process model, RCJs actually made *more frequent* use of cue information, particularly at higher levels of confidence, than did JOLs. These findings are entirely inconsistent with the theory of RCJs being based solely on target accessibility (Dougherty et al., 2005; Koriat et al., 1980), and instead provide strong evidence that much like JOLs, RCJs involve a fast cue-familiarity check followed by a more deliberate target retrieval attempt.

Participants were able to justify their confidence judgments and did so by referring to both cue- and target- related semantic information, as well as inclusion of the cues and targets themselves, and in some cases, created idiosyncratic associations between the cue and target

terms. These justifications and associations occurred in the absence of specific instructions on how to commit to memory the cue-target pairs, and even in the face of many word pairs being *very* unrelated and therefore not easy to associate. This finding is in line with studies that have found that associations between cues and targets can arise naturally from study, even in the absence of specific instructions to form associations, and that said associations can affect confidence (Hertzog et al, 2014; Jersakova et al., 2017; Metcalfe & Finn, 2008b).

JOL results for the LSA and n-gram analyses were in accordance with the descriptive model of JOLs as proposed by Metcalfe & Finn (2008b). Zero percent and 20% JOL justifications had the strongest evidence in favor of cue familiarity. Further, whereas n-grams for 0% JOL justifications indicated an absence of cue familiarity, 20% JOL justifications were best characterized by some amount of cue-familiarity but an inability to retrieve the target. When using SVM analyses on only the n-grams used in the n-gram analysis, the 0-20% confidence boundary was the most successfully discriminated boundary by the SVM algorithm – and this boundary classifier was significantly more accurate than was any other JOL confidence boundary classifier. Together, these results all provide support for the dual-process descriptive model of JOLs (Metcalfe & Finn, 2008b), whereby the lowest levels of confidence are characterized by cue familiarity or lack thereof, and then increasing amounts of target accessibility as confidence increases and the cue is successfully deemed to be familiar. These results are also consistent with the findings of Jersakova et al. (2017).

RCJ results for the LSA and n-gram analyses were generally parallel to the JOL results and in line with the dual-process descriptive model of JOLs (Metcalfe & Finn, 2008b) – a set of findings that potentially extends the utility of the dual-process descriptive model to another type of MC judgment. However, there were some key differences in both the LSA and n-gram

analyses between JOLs and RCJs. First, the LSA results for RCJs indicated significantly more cue- than target- referencing and semantic information at more than just the lowest level of confidence. In fact, RCJs appear to include more cue than target information at four of the six possible confidence levels, the three lowest levels of confidence and the highest level of confidence. This contrasts with the JOL LSA results, which only had more cue than target information present at the lowest and second lowest levels of confidence. Despite this finding, the relative amount of cue/target information at any given level of confidence did not differ significantly between RCJs and JOLs - although according to the Bayesian analysis of these results, evidence was insensitive at the highest level of confidence for both cue and target cosine similarity. This result could be taken as indicating that neither the null hypothesis – that JOLs and RCJs contained the same amount of cue/target information – nor the alternative hypothesis – that JOLs and RCJs contained differing amounts of cue/target information – were supported. It is therefore possible that JOLs and RCJs do in fact differ in terms of cue/target information at the highest levels of confidence, but this study cannot support this claim with the evidence available.

The n-gram analyses for RCJs paint a similar picture to the LSA analyses. RCJs generally followed a similar pattern to JOLs – 0% confidence was best characterized by an absence of cue-familiarity, 20% RCJs contained indications of cue-familiarity and an inability to retrieve the target, and target information and certainty increased as confidence increased. There were, however, differences between the JOL and RCJ n-gram results. Most notably, the RCJ results for the highest level of confidence included associations between the cue and target, a finding that was less prevalent or even absent in the highest level of JOL confidence. Further, the 80% confidence level contained slightly stronger ratings of certainty in RCJs than it did for JOLs. Finally, RCJ justifications included significantly more frequent inclusion of the cue term itself, at

all levels of confidence, than did JOL justifications, whereas target inclusion did not differ at any level of confidence. These findings are a strong refutation of the hypothesis of RCJs as being solely comprised of target-accessibility (Dougherty et al., 2005; Koriat et al.,1980). Support for the target-accessibility hypothesis would instead have been not finding any references to the cue-term or n-grams indicating cue familiarity, at any level of confidence, and increasing amounts of target inclusion as confidence increased. As this was not the case, we can safely state that the RCJ n-gram analyses support the dual-process model of confidence.

The RCJ SVM analyses using only n-grams used in the n-gram analyses largely support the results of the LSA and n-gram analyses, and the dual-process descriptive model of MC judgments (Metcalfe & Finn, 2008b), with one notable exception – the classifier was most accurate at the lowest boundary of confidence and was less accurate at all other levels of confidence *except for the highest level of confidence*. This finding, however, does seem to support the earlier LSA result that the highest level of RCJ confidence had significantly higher cue- than target- cosine similarity. Furthermore, this finding can be explained at least partially by the n-gram results – the presence of associations between cue and target at the highest level of confidence would necessitate information semantically related to both the cue and the target, if not the cue and target terms themselves. This additional information is categorically different from the information solely based on the target and would therefore allow for this boundary to be more successfully classified when compared to a boundary that is only assessing different magnitudes of target accessibility.

The questions this study sought to answer were 1) whether JOLs use the dual-process model proposed by Son & Metcalfe (2005) and further expanded upon by Metcalfe & Finn (2008b); and 2) whether RCJs operate on the same dual-process model as do JOLs (Gardiner et

al., 1998; Hanczakowski et al., 2013; Selmeczy & Dobbins, 2014; Williams et al., 2013), or instead operate solely on target-retrievability/accessibility, as per Dougherty et al (2005) and Koriat et al. (1980). The results of this study provide a rather clear response to both questions. As far as the first question is concerned, all three text data analyses provided support for the dual process descriptive model of JOL confidence. Cue semantic similarity and referencing was found at the lowest levels of JOL confidence via the LSA. This cue similarity and referencing was then differentiated into absence of cue familiarity and presence of cue familiarity by the n-gram analysis – n-grams at the lowest level of confidence indicated not remembering the cue, whereas n-grams at the second lowest level of confidence indicated some vague memory of the cue but no memory of the target. Finally, the SVM successfully classified this boundary between the lowest and second lowest confidence ratings at a rate higher than all other boundary classifications. Taken together, these findings match exactly with the predictions derived from the dual-process descriptive model (Jersakova et al., 2017; Metcalfe & Finn, 2008b).

Regarding the second question – whether RCJs operate on the same dual-process model as do JOLs, or instead operate solely on target-accessibility - the results also provided a clear and compelling answer. The prediction of the target-retrievability/accessibility hypothesis was that RCJs would at no point utilize cue information to a greater extent than target information, if any cue information was utilized whatsoever. The results of this study not only do not support this prediction, but in fact directly contradict it. Evidence that RCJ justifications not only used cue-familiarity and cue-based information but did so to a greater extent than they made use of target-based information, was evident in all three text data analyses. The LSA results of RCJs indicated greater cue than target semantic similarity at four out of six possible levels of confidence. The n-gram analysis found evidence of cue referencing at the lowest, second lowest, and highest levels

of confidence, and that the lowest two levels of confidence followed the same pattern as did JOLs, with 0% confidence justifications indicating an absence of cue familiarity and 20% confidence justifications indicating the presence of cue-familiarity and a failure to retrieve the target. Additionally, the final n-gram analysis found inclusion of the cue term at significantly greater frequencies than inclusion of the target term at all levels of confidence. The SVM analyses for RCJs was able to successfully categorize the lowest and highest boundaries of confidence responses at a rate higher than chance, a finding that is inconsistent with what would be expected if only differing magnitudes of target access were used in the formation of confidence. Therefore, not only did the results provide support for RCJs using cue-familiarity in addition to target access, but they also indicated that RCJs make more extensive use of cue information than target information at many levels of confidence, and in some cases more so than do JOLs. This is strong support for RCJs using the dual-process descriptive model (Jersakova et al., 2017; Metcalfe & Finn, 2008b) access just as JOLs do (Gardiner et al., 1998; Hanczakowski et al., 2013; Selmeczy & Dobbins, 2014; Williams et al., 2013) and an equally strong refutation of the target accessibility model of RCJ confidence (Dougherty et al., 2005; Koriat et al., 1980).

Limitations and Future Directions

There are several limitations to the current study. The LSA required dropping words that were not included in the LSA corpus. This led to the exclusion of 53 observations from analyses. Future paired associates studies would do well to verify that all of the words used are included in the LSA corpus prior to data collection. Additionally, at least 11 confidence levels were underpowered for the LSA analyses - they did not meet the minimum of 220 observations needed for a power of 0.8 as per the a priori power analysis. These 11 confidence levels tended to be towards the middle of the confidence scale, suggesting that participants tended to make use

of the extremes of the confidence scale more so than using the full range of the scale. Perhaps inclusion of instructions requesting participants to make use of the full range of the scale would help to distribute confidence more evenly in future studies. Fortunately, none of the confidence levels that were key to testing the research questions of this study were underpowered.

Although the n-gram analyses went rather smoothly, the SVM analyses that made use of the same data structure as the n-gram analyses had quite a few issues. When using the full set of words and n-grams (with stop words removed), all SVM classifiers were very highly and similarly accurate. These results may tell us that there are categorical differences between each level of confidence, for each type of confidence judgment. These results, although potentially correct, are not supported by the results of the LSA (that also used all words and n-grams, more wholistically), or the n-gram analysis. Furthermore, for the SVMs that were able to be run with all words and n-grams *without* removing stop words, classification accuracy was even higher, closer to 100%, in all cases. This tells us is that with sufficient characteristics – somewhere between 4000 and 25000 features depending on the presence of stop words - it is possible to differentiate anything effectively – a finding that is not terribly helpful in answering the research questions at the heart of this study. Additional issues with the SVM analyses were largely related to constraints of available computing power. Whereas the n-gram analyses were able to be conducted without removing stop-words, the required computing power to run the SVM analyses without removing stop-words greatly exceeded what was available. Even making use of available graphical processing units in addition to central processing units for calculations and increasing the maximum pointer references to the highest value allowed by R were insufficient to run the SVMs when the feature dataset was in excess of 20000 variables. The removal of stop-words from the dataset used in the SVM analyses reduced the number of features to a much more

manageable ~9000 variables, allowing for the SVMs to run successfully. This did, however, mean that a different dataset was utilized for the n-gram and SVM analyses. Future studies could either 1) run both sets of analyses with stop-words removed, or 2) make use of much more powerful computers.

Beyond the issues with the text analysis procedures, there are limitations to the study itself. The study was set up to test whether JOLs and RCJs both operated on a dual-process model of confidence, comprised of a cue familiarity component and a target accessibility component. The study was able to fulfill this purpose effectively – both JOLs and RCJs appear to use the same dual-process model. However, there are certainly differences in *how* that dual-process model is used – RCJs appear to make more extensive use of cue information at more levels of confidence than do JOLs. Furthermore, as indicated by the SVM results when including all words and n-grams, there appear to be other differentiating factors at each level of confidence for the two confidence judgments. This study is not designed to explore or test what those differences are, and as such there are still unresolved questions about the exact differences between JOLs and RCJs as well as why RCJs tend to be more accurate than JOLS. It should be noted that although cue familiarity and target accessibility are suitable explanations for metamemorial confidence, they are certainly not the only ones. Cue familiarity is a limited factor in confidence, only explaining differences in the lowest levels of confidence – was the cue familiar or not? Target accessibility is a bit broader, it can consist of varying amounts of partial or even complete target information as well as the retrieval fluency associated with the retrieval attempt that resulted in said partial or complete target information. But there are *many* other sources of confidence that can be drawn from the inferential approach, and if the SVM results are correct, at least some of them are at play here. One such source of information that could

differentially affect JOLs and RCJs is the *memory for past test (MPT) heuristic*. The MPT heuristic (Ariel & Dunlosky, 2011; Finn & Metcalfe, 2007; 2008; Hertzog, Hines, and Touron, 2013; Serra & Ariel, 2014; Tauber & Rhodes, 2012) states that JOLs, when made following an explicit retrieval attempt or test, make use of the results of that test in the formation of confidence. Indeed, JOLs that can make use of the MPT heuristic tend to be more accurate than JOLs that occur prior to a test and therefore cannot use MPT. One thing to keep in mind – all RCJs, by definition, occur following an explicit retrieval attempt/test, and therefore always have MPT available as a source of information. The JOLs in this study were made prior to any explicit retrieval attempts/tests, and therefore could not have utilized the MPT heuristic as a source of confidence. It should be noted, however, that even including the MPT heuristic, JOLs are still often less accurate than are RCJs (Dougherty et al., 2005). That said, a possible future direction would be to replicate the procedure of this study but with a PRAM methodology (Nelson, Narens, and Dunlosky, 2004) for the JOL condition – this would allow participants in the JOL condition to make use of the MPT heuristic – allowing for a better parsing of the specific differences in process between JOLs and RCJs on a more even ground.

Generally, RCJs have notably superior calibration and resolution to JOLs, however, in this study, JOL and RCJ calibration and resolution were both very high. One of the traditional hallmarks of JOLs in a multi-test paradigm is that they are initially overconfident and then shift to underconfidence following explicit retrieval attempts/tests. The design of this study did not allow for JOLs to be observed at multiple time points – all participants made RCJs at the final test – so it was not possible to test for the so-called Underconfidence with practice (UPT) effect (Koriat, Sheffer, & Ma'ayan, 2002). Further, JOLs at the initial judgment were not overconfident - as would be predicted by the UPT effect - they were instead appropriately confident.

Participants in the JOL condition correctly recalled the target ~27% of the time, and average confidence at initial judgment for the JOL condition was 1.79 out of 6, which is approximately 29.8%. For comparison, participants in the RCJ condition correctly recalled the target ~33% of the time and had an average confidence at initial judgment of 2.07 out of 6, which is approximately 34.5%. Additionally, gamma correlations (i.e. resolution) for both JOLs and RCJs were in excess of .85 and can therefore be classified as very high. Given that calibration and resolution of confidence judgments in both conditions was very high – participants in both the JOL and RCJ conditions were generally able to tell when they were going to/had correctly retrieve(d) the target term and when they would not. This finding is at odds with previous findings in the field that RCJs demonstrate superior calibration and resolution to JOLs (Dougherty et al., 2005; Hines et al., 2009; Nelson & Dunlosky,1991; Nguyen et al., 2017; Perfect and Hollins, 1996; Robey et al., 2017; Ryal et al., 2016; Siedlecka et al., 2016; Wattier & Collins, 2011). It is not entirely clear why participants' JOLs were so accurate in this study. One possible explanation involving a third popular explanation for MC confidence, the anchoring explanation (Scheck et al., 2004; Scheck & Nelson, 2005), is that the task was difficult enough for participants to rely primarily on their psychological anchor point for confidence for the task, approximately 30% (Scheck et al., 2004; Scheck & Nelson, 2005), and then use other available information to adjust away from that anchor point. The task was presumably equally difficult for participants in the JOL and RCJ conditions, so the only differences in confidence observed would be the slight adjustments away from that anchor point as a result of cue-familiarity and target accessibility. Participants in the JOL condition did perform slightly worse on the recall task than did participants in the RCJ condition, indicating slightly lower target access. This would in turn mean there was slightly less adjustment away from the anchor point for

participants in the JOL condition and that said adjustment was likely in the opposite direction from participants in the RCJ condition.

As mentioned previously, a possible future direction for this line of research is to repeat this study but using a PRAM methodology (Nelson, Narens, and Dunlosky, 2004) – essentially just having participants in the JOL condition complete the initial recall test in the same manner the participants in the RCJ condition do, and then having them make a JOL as usual. This methodology would allow for participants in both conditions to have access to the same information – namely memory for their test performance – when making their confidence judgment.

An additional future direction would be to follow up on exactly why RCJs appeared to make more frequent use of cue information at higher levels of confidence than did JOLs. This result was not predicted by either the target-accessibility approach or the dual process descriptive model and is perhaps the most unique finding of the study. It is possible that repeating the study with a PRAM methodology would shed light on exactly what caused this finding, but other avenues of research should be considered as well.

A third possible future direction would be to add an additional condition in which participants are asked to engage in a perceptual discrimination task and then make a metacognitive confidence judgment instead of a memory task and a metamemorial confidence judgment. This would potentially allow for a parsing of the procedure involved in metamemorial as opposed to metacognitive confidence, and might provide insight on the role of inference as opposed to direct access in both types of confidence judgments.

One more future direction that might shed further light specifically on the difference between process in JOLs and RCJs can be derived the conceptual differences between JOLs and

RCJs. RCJs intrinsically have access to information not necessarily available to the individual when making JOLs – memory for test performance. Studies exploring the memory for past test heuristic (Ariel & Dunlosky, 2011; Finn & Metcalfe, 2007; 2008; Hertzog, Hines, & Touron, 2013; Serra & Ariel, 2014; Tauber & Rhodes, 2012) as well as the PRAM methodology (Nelson, Narens, and Dunlosky, 2004) have taken steps to equate the information available when making confidence judgments, however, there is also evidence that JOLs integrate information pertinent to new learning and forgetting (Ariel & Dunlosky, 2011). Further, there is one other major conceptual difference between JOLs and RCJs – JOLs are predictions of future memory performance, whereas RCJs are assessments of past memory performance. As such, new learning and forgetting might not be included in the calculus involved in making an RCJ. This difference *could* be explored by a somewhat unorthodox approach - reversing the temporal direction of a JOL – so that new learning and forgetting is also not necessarily included in the judgment process. This could be done by having participants make a JOL about hypothetical past memory performance – i.e. asking them about how they would have done on a memory test that they did not in fact do. This could help shed light on whether the differences observed in JOL and RCJ accuracy is more a result of differential information use or instead the different temporal direction of the judgments.

Conclusion

Ultimately, and despite numerous limitations, the findings of the current study do answer the research questions this study explored. Not only do JOL justifications indicate that people make use of both cue familiarity and target access in the formation of confidence, but RCJs do as well. Further, and in direct contrast with the target access hypothesis proposed by Dougherty et al., (2005) and Koriat et al. (1980), RCJ justifications actually made more extensive use of cue information than of target information at all levels of confidence. In some cases, RCJ

justifications even made more use of cue information than did JOL justifications. As such, it can safely be concluded that although JOLs and RCJs do historically demonstrate differences in calibration and resolution (albeit not in this study), these differences are likely not derived from differential utilization of cue-familiarity and target accessibility.

Appendix A:

Procedure Flowchart

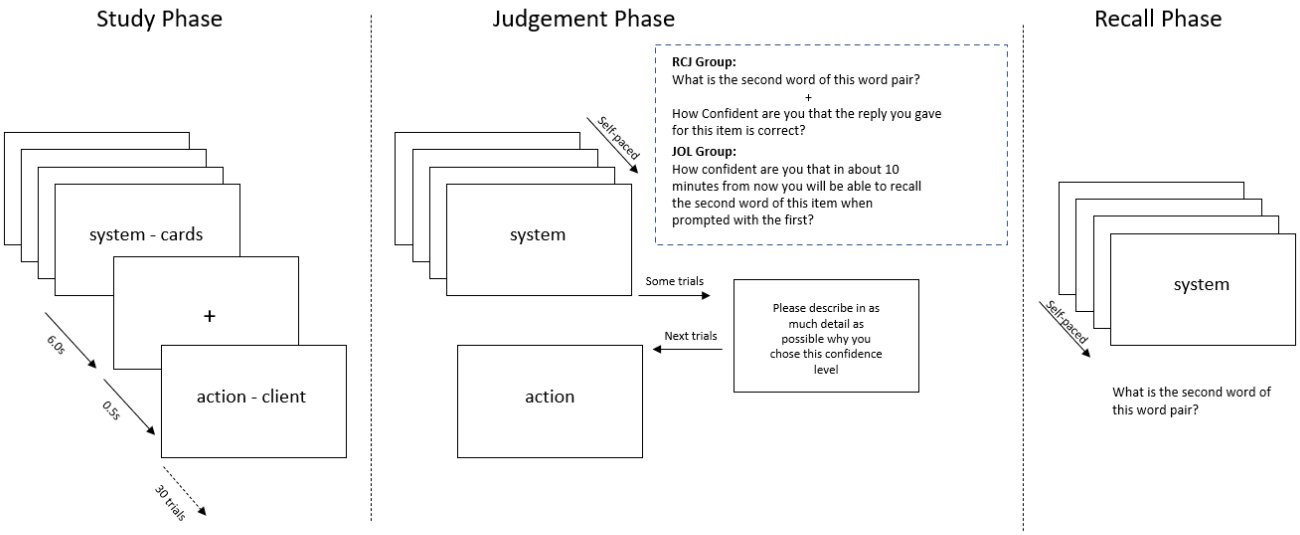

Fig. 1. *Experimental Procedure.* The three phases together are one experimental block, and participants will complete two blocks, each block consisting of different items. In the judgement phase, participants either 1) attempt to recall the second word of the word pair associated with the presented word, and then provide a retrospective confidence judgment on that recall attempt; or 2) provide a judgment of learning indicating how confident they are that in the future they will be able to recall the second word of the word pair associated with the presented word. Confidence judgments are on a 0-100% scale, in increments of 20%. On a subset of trials, participants will be asked to explain why they gave a particular confidence rating on the preceding trial.

Appendix B:

N-gram Frequency Analysis Tables

Table 3

JOL N-gram analysis results by confidence level

JOL Confidence	N-gram	Count	Total	Proportion	p
0%	10 minutes	8	21	0.3810	0.0156
	am not confident	5	11	0.4545	0.0245
	and i	11	30	0.3667	0.011
	at all	39	47	0.8298	<.001
	can not	6	11	0.5455	0.0046
	cannot remember	29	55	0.5273	<.001
	cannot remember the	11	15	0.7333	<.001
	did not	14	27	0.5185	<.001
	do not	188	317	0.5931	<.001
	do not recall	24	43	0.5581	<.001
	do not remember	130	169	0.7692	<.001
	first word	8	17	0.4706	0.0035
	have no	20	25	0.8000	<.001
	i cannot	37	78	0.4744	<.001
	i cannot recall	7	16	0.4375	0.0101
	i cannot remember	27	44	0.6136	<.001
	i did	8	17	0.4706	0.0035
	i did not	8	15	0.5333	0.0013
	i do	161	265	0.6075	<.001
	i do not	161	261	0.6169	<.001
	i have	31	94	0.3298	<.001
	i have no	18	23	0.7826	<.001
	i will	26	90	0.2889	0.0041
	i will be	6	11	0.5455	0.0046
	i will not	9	10	0.9000	<.001

JOL Confidence	N-gram	Count	Total	Proportion	p
	in 10	7	19	0.3684	0.0281
	in 10 minutes	7	17	0.4118	0.0147
	many words	9	10	0.9000	<.001
	matching word for	5	11	0.4545	0.0245
	no idea	8	10	0.8000	<.001
	no memory	9	11	0.8182	<.001
	not confident	6	13	0.4615	0.0127
	not know	11	33	0.3333	0.0172
	not recall	30	51	0.5882	<.001
	not remember	143	193	0.7409	<.001
	not remember it	16	19	0.8421	<.001
	not remember seeing	14	15	0.9333	<.001
	not remember the	58	70	0.8286	<.001

	not remember this	16	22	0.7273	<.001
	not think	11	35	0.3143	0.0367
	now so	8	19	0.4211	0.0079
	now so i	6	11	0.5455	0.0046
	of words	7	15	0.4667	0.0066
	other word	10	24	0.4167	0.0033
	paired with	6	15	0.4000	0.0274
	recall seeing	5	11	0.4545	0.0245
	recall the	12	36	0.3333	0.0125
	recall this	7	15	0.4667	0.0066
	remember it	29	117	0.2479	0.0248
	remember seeing	24	45	0.5333	<.001
	remember seeing the	9	21	0.4286	0.0042
	remember the	81	209	0.3876	<.001
	remember the pair	7	18	0.3889	0.0206
JOL Confidence	N-gram	Count	Total	Proportion	p
	remember the second	20	48	0.4167	<.001
	remember the word	36	95	0.3789	<.001
	remember this word	9	29	0.3103	0.0464
	second word	31	131	0.2366	0.035
	seeing the	12	27	0.4444	<.001
	seeing the word	9	23	0.3913	0.0086
	seeing this	9	16	0.5625	<.001
	seeing this word	8	15	0.5333	0.0013
	that word	11	15	0.7333	<.001
	the first	12	36	0.3333	0.0125
	the first word	7	13	0.5385	0.0024
	the other	12	30	0.4000	0.002
	the other word	9	20	0.4500	0.0028
	the pair	14	44	0.3182	0.0132
	the second	35	141	0.2482	0.0126
	the second word	30	127	0.2362	0.0423
	the word	61	255	0.2392	0.0031
	the word pair	11	27	0.4074	0.0026
	there were	7	10	0.7000	<.001
	this word	27	79	0.3418	<.001
	too many	11	13	0.8462	<.001
	went with	6	12	0.5000	0.0079
	what the	16	51	0.3137	0.0081
	what the second	8	20	0.4000	0.0113
	what word	9	16	0.5625	<.001
	will be	6	12	0.5000	0.0079
	will not	11	13	0.8462	<.001
	with it	6	16	0.3750	0.0378

JOL Confidence	N-gram	Count	Total	Proportion	p
	with this	7	15	0.4667	0.0066
	with this one	6	11	0.5455	0.0046
	word at	13	15	0.8667	<.001
	word at all	13	14	0.9286	<.001
	word now	8	18	0.4444	0.0053
	word pair	14	46	0.3043	0.0173
	word pairing	8	14	0.5714	<.001
20%	am not sure	15	48	0.3125	0.011
	but i	34	117	0.2906	<.001
	but i cannot	10	14	0.7143	<.001
	but i do	6	16	0.3750	0.0378
	but it	7	17	0.4118	0.0147
	but not	8	22	0.3636	0.021
	cannot recall	10	23	0.4348	0.0023
	cannot remember	18	55	0.3273	0.0032
	could not	7	15	0.4667	0.0066
	do not	79	317	0.2492	<.001
	do not recall	16	43	0.3721	0.0014
	do not think	14	30	0.4667	<.001
	i am	47	190	0.2474	0.0045
	i am not	23	92	0.2500	0.0361
	i cannot	25	78	0.3205	<.001
	i cannot recall	6	16	0.3750	0.0378
	i could	12	37	0.3243	0.0239
	i do	63	265	0.2377	0.0029
	i do not	62	261	0.2375	0.0035
	i remember seeing	9	17	0.5294	<.001
JOL Confidence	N-gram	Count	Total	Proportion	p
	if i	11	34	0.3235	0.0207
	is not	8	16	0.5000	0.0021
	not recall	18	51	0.3529	0.0011
	not sure	20	62	0.3226	0.003
	not sure if	8	17	0.4706	0.0035
	not think	15	35	0.4286	<.001
	not think i	12	22	0.5455	<.001
	of it	5	10	0.5000	0.0155
	really remember	5	12	0.4167	0.0364
	recall seeing	5	11	0.4545	0.0245
	remember seeing	13	45	0.2889	0.0422
	remembering the	5	10	0.5000	0.0155
	right now	9	16	0.5625	<.001
	second word but	5	11	0.4545	0.0245

	seeing the word	8	23	0.3478	0.0426
	sure if	8	19	0.4211	0.0079
	sure if i	5	10	0.5000	0.0155
	the word	59	255	0.2314	0.0071
	the word but	8	16	0.5000	0.0021
	think of	8	22	0.3636	0.021
	what it	8	18	0.4444	0.0053
	word but	15	35	0.4286	<.001
	word but i	9	20	0.4500	0.0028
40%	am not	26	94	0.2766	0.0078
	but i	31	117	0.2650	0.0086
	but i am	14	48	0.2917	0.0306
	do not really	6	13	0.4615	0.0127
JOL Confidence	N-gram	Count	Total	Proportion	p
	i am not	26	92	0.2826	0.0048
	i believe	6	16	0.3750	0.0378
	i may	9	23	0.3913	0.0086
	i might	12	38	0.3158	0.0259
	idea of what	6	13	0.4615	0.0127
	if i	13	34	0.3824	0.0022
	it but	7	15	0.4667	0.0066
	it might	5	12	0.4167	0.0364
	might be	10	16	0.6250	<.001
	not feel	5	10	0.5000	0.0155
	not really	6	14	0.4286	0.0191
	not sure	18	62	0.2903	0.0155
	of what	9	27	0.3333	0.0336
	of what the	5	12	0.4167	0.0364
60%	because of	5	10	0.5000	0.0155
	but i	28	117	0.2393	0.0459
	but i am	15	48	0.3125	0.011
	could remember	6	13	0.4615	0.0127
	i could remember	6	13	0.4615	0.0127
	i may	8	23	0.3478	0.0426
	i think	52	139	0.3741	<.001
	i think i	33	88	0.3750	<.001
	i think it	7	16	0.4375	0.0101
	i think the	6	10	0.6000	0.0024
	i will remember	13	44	0.2955	0.0395
	it in	8	24	0.3333	0.048
	it is	20	78	0.2564	0.0464
JOL Confidence	N-gram	Count	Total	Proportion	p

	the answer	5	12	0.4167	0.0364
	think i	33	112	0.2946	<.001
	think i know	7	15	0.4667	0.0066
	think i remember	15	45	0.3333	0.0075
	think it	9	22	0.4091	0.0061
	think the	6	12	0.5000	0.0079
	will remember	13	44	0.2955	0.0395
	with another	6	12	0.5000	0.0079
80%	can recall	7	14	0.5000	0.0041
	can remember	8	23	0.3478	0.0426
	i can	20	78	0.2564	0.0464
	i can recall	7	14	0.5000	0.0041
	i can remember	8	22	0.3636	0.021
	i remembered	8	19	0.4211	0.0079
	i think	33	139	0.2374	0.0302
	i think i	24	88	0.2727	0.0139
	memory visual	5	11	0.4545	0.0245
	think i know	6	15	0.4000	0.0274
	to the	7	13	0.5385	0.0024
100%	because i	24	74	0.3243	<.001
	because i remember	9	16	0.5625	<.001
	chose this	10	20	0.5000	<.001
	chose this confidence	8	12	0.6667	<.001
	confidence level	8	18	0.4444	0.0053
	confidence level because	8	16	0.5000	0.0021
	had a	10	18	0.5556	<.001

JOL Confidence	N-gram	Count	Total	Proportion	p
	i chose	10	26	0.3846	0.0066
	i chose this	10	20	0.5000	<.001
	i know	18	64	0.2813	0.0186
	i know the	10	26	0.3846	0.0066
	i made	10	21	0.4762	0.001
	i recall	5	12	0.4167	0.0364
	i remember	84	239	0.3515	<.001
	i remember it	10	27	0.3704	0.0089
	i remember the	27	70	0.3857	<.001
	i remember this	11	34	0.3235	0.0207
	i remembered	7	19	0.3684	0.0281
	i thought	9	24	0.3750	0.0118
	i thought of	5	11	0.4545	0.0245

	it was	18	58	0.3103	0.0071
	level because	8	18	0.4444	0.0053
	level because i	8	17	0.4706	0.0035
	me remember	5	10	0.5000	0.0155
	of a	12	26	0.4615	<.001
	one of	6	14	0.4286	0.0191
	one of the	6	14	0.4286	0.0191
	other word	9	24	0.3750	0.0118
	remember this	26	103	0.2524	0.0243
	remember this pair	5	12	0.4167	0.0364
	stuck out	7	11	0.6364	<.001
	the other word	8	20	0.4000	0.0113
	the same	6	12	0.5000	0.0079
	the two	8	18	0.4444	0.0053
	the two words	8	16	0.5000	0.0021
JOL Confidence	N-gram	Count	Total	Proportion	p
	the word is	9	24	0.3750	0.0118
	they are	5	10	0.5000	0.0155
	this confidence	8	14	0.5714	<.001
	this confidence level	8	14	0.5714	<.001
	this one	28	104	0.2692	0.008
	this was	6	10	0.6000	0.0024
	thought of	7	13	0.5385	0.0024
	two words	14	25	0.5600	<.001
	was one	6	10	0.6000	0.0024
	was one of	6	10	0.6000	0.0024
	word is	15	46	0.3261	0.0084

Note: The count of occurrences in justifications for a given confidence level is reported, along with the total number of occurrences across all confidence levels, the proportion of occurrences, and the p-value for the binomial test.

Table 4

RCJ N-gram analysis results by confidence level

RCJ Confidence	N-gram	Count	Total	Proportion	p
0%	a guess	6	11	0.5455	0.0046
	am sure	7	16	0.4375	0.0101
	associated with	9	28	0.3214	0.0393
	at all	60	68	0.8824	<.001
	because i	21	82	0.2561	0.0371
	because i do	8	11	0.7273	<.001
	came to	6	13	0.4615	0.0127
	came to mind	6	12	0.5000	0.0079
	cannot remember	37	58	0.6379	<.001
	cannot remember the	8	13	0.6154	<.001
	could not	19	31	0.6129	<.001
	could not remember	15	25	0.6000	<.001
	did not	23	32	0.7188	<.001
	did not remember	15	15	1.0000	<.001
	do not	202	310	0.6516	<.001
	do not recall	17	20	0.8500	<.001
	do not remember	152	206	0.7379	<.001
	first word	12	35	0.3429	0.0104
	from the	8	18	0.4444	0.0053
	have no	14	19	0.7368	<.001

RCJ Confidence	N-gram	Count	Total	Proportion	p
	have no idea	9	13	0.6923	<.001
	i am sure	7	16	0.4375	0.0101
	i can	9	28	0.3214	0.0393
	i cannot	32	70	0.4571	<.001
	i cannot remember	24	38	0.6316	<.001
	i could	18	39	0.4615	<.001
	i could not	15	24	0.6250	<.001
	i definitely	5	12	0.4167	0.0364
	i did	15	23	0.6522	<.001
	i did not	15	22	0.6818	<.001
	i do	161	264	0.6098	<.001
	i do not	161	256	0.6289	<.001
	i forgot	9	12	0.7500	<.001
	i had	10	24	0.4167	0.0033
	i have	18	48	0.3750	<.001
	i have no	11	16	0.6875	<.001
	i just	15	42	0.3571	0.0027
	i know	20	61	0.3279	0.0017
	i know that	5	10	0.5000	0.0155
	i put	15	33	0.4545	<.001
	i really	7	11	0.6364	<.001
	instead of	6	14	0.4286	0.0191

RCJ Confidence	N-gram	Count	Total	Proportion	p
	is not	11	19	0.5789	<.001
	know that	6	12	0.5000	0.0079
	know the	6	15	0.4000	0.0274
	no idea	10	14	0.7143	<.001
	not confident	7	20	0.3500	0.0371
	not know	8	24	0.3333	0.048
	not recall	18	22	0.8182	<.001
	not recall the	9	12	0.7500	<.001
	not remember	186	251	0.7410	<.001
	not remember seeing	11	12	0.9167	<.001
	not remember that	12	12	1.0000	<.001
	not remember the	73	93	0.7849	<.001
	not remember this	34	43	0.7907	<.001
	not remember what	10	17	0.5882	<.001
	not the	8	14	0.5714	<.001
	not think	12	36	0.3333	0.0125
	paired with	9	27	0.3333	0.0336
	recall the	13	17	0.7647	<.001
	recall this	6	10	0.6000	0.0024
	remember at	9	11	0.8182	<.001
	remember at all	9	11	0.8182	<.001
	remember seeing	27	69	0.3913	<.001

RCJ Confidence	N-gram	Count	Total	Proportion	p
	remember seeing the	9	23	0.3913	0.0086
	remember so	7	10	0.7000	<.001
	remember that	16	38	0.4211	<.001
	remember that word	10	10	1.0000	<.001
	remember the	93	205	0.4537	<.001
	remember the pair	7	11	0.6364	<.001
	remember the second	28	41	0.6829	<.001
	remember the word	41	106	0.3868	<.001
	remember this	37	144	0.2569	0.0068
	remember this one	18	55	0.3273	0.0032
	remember this word	6	16	0.3750	0.0378
	remember what	18	30	0.6000	<.001
	remember what the	9	13	0.6923	<.001
	second word	48	92	0.5217	<.001
	seeing the	11	31	0.3548	0.0124
	seeing the word	10	22	0.4545	0.0015
	so i	28	106	0.2642	0.0124
	that i	14	48	0.2917	0.0306
	that word	16	24	0.6667	<.001
	the first	17	56	0.3036	0.0108
	the first word	11	32	0.3438	0.0145
	the pair	15	44	0.3409	0.0041

RCJ Confidence	N-gram	Count	Total	Proportion	p
	the pairing	7	16	0.4375	0.0101
	the pairs	5	11	0.4545	0.0245
	the second	51	96	0.5313	<.001
	the second word	45	85	0.5294	<.001
	the word	98	322	0.3043	<.001
	the word pairing	7	10	0.7000	<.001
	this word	19	56	0.3393	0.0017
	to mind	8	21	0.3810	0.0156
	was associated	5	12	0.4167	0.0364
	was associated with	5	11	0.4545	0.0245
	was not	13	29	0.4483	<.001
	was paired	8	16	0.5000	0.0021
	was paired with	7	14	0.5000	0.0041
	what the	22	37	0.5946	<.001
	what the second	10	13	0.7692	<.001
	what the word	6	16	0.3750	0.0378
	word at	15	15	1.0000	<.001
	word at all	15	15	1.0000	<.001
	word is	8	16	0.5000	0.0021
	word pairing	9	16	0.5625	<.001
	word that	14	40	0.3500	0.0045

RCJ Confidence	N-gram	Count	Total	Proportion	p
20%	a word	11	34	0.3235	0.0207
	am not	38	99	0.3838	<.001
	am not confident	8	14	0.5714	<.001
	am not sure	17	37	0.4595	<.001
	be correct	5	10	0.5000	0.0155
	but i	58	173	0.3353	<.001
	but i am	16	53	0.3019	0.0148
	but i cannot	13	27	0.4815	<.001
	but i think	5	12	0.4167	0.0364
	came to mind	5	12	0.4167	0.0364
	confident in	5	12	0.4167	0.0364
	could be	7	16	0.4375	0.0101
	do not	67	310	0.2161	0.0222
	do not think	15	34	0.4412	<.001
	feel like	12	35	0.3429	0.0104
	have been	7	15	0.4667	0.0066
	i am	70	271	0.2583	<.001
	i am not	35	94	0.3723	<.001
	i cannot	19	70	0.2714	0.0242
	i do	61	264	0.2311	0.0064
	i do not	58	256	0.2266	0.0147
	i feel	17	46	0.3696	0.001

RCJ Confidence	N-gram	Count	Total	Proportion	p
	i feel like	10	30	0.3333	0.0239
	i guessed	7	12	0.5833	0.0013
	i put was	5	12	0.4167	0.0364
	i remember seeing	15	41	0.3659	0.0023
	i remember that	6	15	0.4000	0.0274
	is a	11	34	0.3235	0.0207
	is the	13	45	0.2889	0.0422
	it could	7	18	0.3889	0.0206
	it could be	6	12	0.5000	0.0079
	it is	24	96	0.2500	0.0386
	like it	5	10	0.5000	0.0155
	might be	7	16	0.4375	0.0101
	not confident	11	20	0.5500	<.001
	not really	6	10	0.6000	0.0024
	not sure	27	79	0.3418	<.001
	not think	15	36	0.4167	<.001
	not think it	10	18	0.5556	<.001
	not too	7	17	0.4118	0.0147
	not too sure	7	13	0.5385	0.0024
	not very	6	12	0.5000	0.0079
	on the	5	11	0.4545	0.0245
	put was	5	12	0.4167	0.0364

RCJ Confidence	N-gram	Count	Total	Proportion	p
	sort of	5	12	0.4167	0.0364
	that is	10	21	0.4762	0.001
	that the word	6	15	0.4000	0.0274
	that was	8	21	0.3810	0.0156
	the word	80	322	0.2484	<.001
	the word i	11	23	0.4783	<.001
	there is	9	17	0.5294	<.001
	there is a	9	14	0.6429	<.001
	think i	17	62	0.2742	0.0384
	think it	16	48	0.3333	0.0053
	think it is	9	20	0.4500	0.0028
	think this	5	12	0.4167	0.0364
	this is	14	45	0.3111	0.015
	to mind	9	21	0.4286	0.0042
	too sure	7	13	0.5385	0.0024
	went with	9	21	0.4286	0.0042
	word but	18	38	0.4737	<.001
	word but i	12	24	0.5000	<.001
	word i	18	40	0.4500	<.001
	word i put	6	13	0.4615	0.0127
	word that	15	40	0.3750	0.002

RCJ Confidence	N-gram	Count	Total	Proportion	p
40%	a pair	5	10	0.5000	0.0155
	am not sure	11	37	0.2973	0.0445
	am unsure	6	14	0.4286	0.0191
	but i	47	173	0.2717	<.001
	but i am	18	53	0.3396	0.0024
	but not	8	20	0.4000	0.0113
	could be	7	16	0.4375	0.0101
	i am unsure	6	13	0.4615	0.0127
	i believe	11	28	0.3929	0.0037
	it could be	5	12	0.4167	0.0364
	not 100	5	12	0.4167	0.0364
	not sure	22	79	0.2785	0.0145
	something to	5	12	0.4167	0.0364
	something to do	5	12	0.4167	0.0364
	what i	5	11	0.4545	0.0245
60%	i am not	24	94	0.2553	0.0264
	i think	36	143	0.2517	0.0094
	it being	5	10	0.5000	0.0155
	it was a	6	15	0.4000	0.0274
	think i remember	9	28	0.3214	0.0393
	was a	8	24	0.3333	0.048

RCJ Confidence	N-gram	Count	Total	Proportion	p
80%	about this	9	25	0.3600	0.0262
	am pretty	15	31	0.4839	<.001
	am pretty sure	13	25	0.5200	<.001
	i am pretty	15	31	0.4839	<.001
	i remember thinking	11	24	0.4583	<.001
	if it	10	25	0.4000	0.0047
	if it was	8	19	0.4211	0.0079
	imagined a	5	12	0.4167	0.0364
	is right	7	14	0.5000	0.0041
	it was	33	133	0.2481	0.0192
	pretty sure	14	29	0.4828	<.001
	remember thinking	12	26	0.4615	<.001
	thinking about	6	12	0.5000	0.0079
	thinking of	6	14	0.4286	0.0191
	to the	7	20	0.3500	0.0371
100%	a connection	5	12	0.4167	0.0364
	a lot	5	10	0.5000	0.0155
	a mental	9	10	0.9000	<.001
	about a	5	10	0.5000	0.0155
	am sure	6	16	0.3750	0.0378
	as a	7	10	0.7000	<.001

RCJ Confidence	N-gram	Count	Total	Proportion	p
	because i	27	82	0.3293	<.001
	because i remember	6	16	0.3750	0.0378
	because it	7	14	0.5000	0.0041
	because the	7	10	0.7000	<.001
	between the	5	11	0.4545	0.0245
	confidence level	5	11	0.4545	0.0245
	easy to	13	20	0.6500	<.001
	easy to remember	10	16	0.6250	<.001
	i am sure	6	16	0.3750	0.0378
	i imagined	11	14	0.7857	<.001
	i know this	5	10	0.5000	0.0155
	i made	12	25	0.4800	<.001
	i made a	6	11	0.5455	0.0046
	i read	8	13	0.6154	<.001
	i remember	125	346	0.3613	<.001
	i remember the	17	65	0.2615	0.0459
	i remember thinking	8	24	0.3333	0.048
	i remember this	52	70	0.7429	<.001
	i remembered	29	62	0.4677	<.001
	i remembered the	5	10	0.5000	0.0155
	i remembered this	11	18	0.6111	<.001
	i saw	8	24	0.3333	0.048

RCJ Confidence	N-gram	Count	Total	Proportion	p
	i thought	32	60	0.5333	<.001
	i thought about	8	12	0.6667	<.001
	i thought of	15	30	0.5000	<.001
	i was	13	40	0.3250	0.0167
	imagined a	7	12	0.5833	0.0013
	in a	7	16	0.4375	0.0101
	is a	11	34	0.3235	0.0207
	know this	6	13	0.4615	0.0127
	like a	5	10	0.5000	0.0155
	made a	6	12	0.5000	0.0079
	me of	8	19	0.4211	0.0079
	of a	22	48	0.4583	<.001
	of my	5	11	0.4545	0.0245
	of the	26	90	0.2889	0.0041
	of the words	5	12	0.4167	0.0364
	pair i	5	10	0.5000	0.0155
	remember because	7	10	0.7000	<.001
	remember it	13	46	0.2826	0.0458
	remember thinking	9	26	0.3462	0.0293
	remember this	69	144	0.4792	<.001
	remember this one	25	55	0.4545	<.001
	remember this pair	16	28	0.5714	<.001

RCJ Confidence	N-gram	Count	Total	Proportion	p
	remember this pairing	10	12	0.8333	<.001
	remembered the	6	11	0.5455	0.0046
	remembered this	13	20	0.6500	<.001
	remembered this one	7	10	0.7000	<.001
	reminded me	5	12	0.4167	0.0364
	reminded me of	5	12	0.4167	0.0364
	saw the	6	12	0.5000	0.0079
	so it	8	18	0.4444	0.0053
	the last	6	14	0.4286	0.0191
	the same	7	18	0.3889	0.0206
	the two	13	20	0.6500	<.001
	the two words	10	15	0.6667	<.001
	the words	16	48	0.3333	0.0053
	this one	49	144	0.3403	<.001
	this pair	23	49	0.4694	<.001
	this pairing	17	36	0.4722	<.001
	thought about	9	14	0.6429	<.001
	thought of	18	40	0.4500	<.001
	to make	5	12	0.4167	0.0364
	to remember	22	51	0.4314	<.001
	tried to	5	11	0.4545	0.0245
	two words	13	22	0.5909	<.001

RCJ Confidence	N-gram	Count	Total	Proportion	p
	was easy	13	16	0.8125	<.001
	was easy to	8	11	0.7273	<.001
	when i	20	40	0.5000	<.001
	when i saw	7	10	0.7000	<.001
	with the	14	48	0.2917	0.0306
	word pair	15	44	0.3409	0.0041
	words i	7	11	0.6364	<.001

Note: The count of occurrences in justifications for a given confidence level is reported, along with the total number of occurrences across all confidence levels, the proportion of occurrences, and the p-value for the binomial test.

Table 5

JOL vs RCJ N-gram analysis results by confidence level

Confidence Level	N-gram	Count JOL	Count RCJ	Total	Proportion JOL	Proportion RCJ	p
0%	not remember this	16	34	50	0.3200	0.6800	0.0153
	remember it	29	7	36	0.8056	0.1944	<.001
	to remember	17	5	22	0.7727	0.2273	0.0169
20%	am not confident	1	8	9	0.1111	0.8889	0.0391
	do not recall	16	3	19	0.8421	0.1579	0.0044
	if i	11	1	12	0.9167	0.0833	0.0063
	not recall	18	4	22	0.8182	0.1818	0.0043
	the first word	1	9	10	0.1000	0.9000	0.0215
	to remember	18	6	24	0.7500	0.2500	0.0227
40%	have a	11	2	13	0.8462	0.1538	0.0225
	i can	14	1	15	0.9333	0.0667	0.001
	i have	15	3	18	0.8333	0.1667	0.0075

	if i	13	4	17	0.7647	0.2353	0.049
	might be	10	1	11	0.9091	0.0909	0.0117
	remember it	20	6	26	0.7692	0.2308	0.0094
	second word	22	8	30	0.7333	0.2667	0.0161
	the second	24	9	33	0.7273	0.2727	0.0135
	the second word	22	8	30	0.7333	0.2667	0.0161
	what the	10	2	12	0.8333	0.1667	0.0386
Confidence Level	N-gram	Count JOL	Count RCJ	Total	Proportion JOL	Proportion RCJ	p
60%	have a	11	2	13	0.8462	0.1538	0.0225
	i can	13	3	16	0.8125	0.1875	0.0213
	i have	14	4	18	0.7778	0.2222	0.0309
	i think i	33	13	46	0.7174	0.2826	0.0045
	the second	15	5	20	0.7500	0.2500	0.0414
	think i	33	13	46	0.7174	0.2826	0.0045
80%	am pretty sure	2	13	15	0.1333	0.8667	0.0074
	i am pretty	4	15	19	0.2105	0.7895	0.0192
	i can	20	4	24	0.8333	0.1667	0.0015
	i know	15	5	20	0.7500	0.2500	0.0414

	i think i	24	11	35	0.6857	0.3143	0.041
	remember it	17	5	22	0.7727	0.2273	0.0169
	remember the word	4	15	19	0.2105	0.7895	0.0192
	second word	22	7	29	0.7586	0.2414	0.0081
	the second	21	8	29	0.7241	0.2759	0.0241
	the second word	20	7	27	0.7407	0.2593	0.0192
	think i	25	12	37	0.6757	0.3243	0.047
100%	had a	10	1	11	0.9091	0.0909	0.0117
	i remember this	11	52	63	0.1746	0.8254	<.001
	i thought of	5	15	20	0.2500	0.7500	0.0414
	remember this one	11	25	36	0.3056	0.6944	0.0288
	remember this pair	5	16	21	0.2381	0.7619	0.0266
	the other	8	1	9	0.8889	0.1111	0.0391

Note: The count of occurrences in justifications for a given confidence level for a given MC judgment type is reported, along with the total number of occurrences across all confidence levels across both MC

judgment types, the proportion of occurrences for each MC judgment type, and the p-value for the binomial test.

Table 6

N-gram that only appeared in one MC condition by confidence level

Confidence Level	N-gram	Count	Judgment Type
0%	could not remember	15	RCJ
	did not remember	15	RCJ
	i could not	15	RCJ
	i put	15	RCJ
	i will	26	JOL
	not remember it	16	JOL
	not remember that	12	RCJ
	too many	11	JOL
	was not	13	RCJ
	will not	11	JOL
20%	a word	11	RCJ
	i might	11	JOL
	i will	16	JOL
Confidence Level	N-gram	Count	Judgment Type
	not think i	12	JOL
	the word i	11	RCJ
40%	i might	12	JOL
	i will	15	JOL

60%	i might	11	JOL
	i remembered	12	RCJ
	i will	16	JOL
	i will remember	13	JOL
	will remember	13	JOL
80%	i remember thinking	11	RCJ
	remember thinking	12	RCJ
100%	easy to	13	RCJ
	i imagined	11	RCJ
	i remembered this	11	RCJ
	remembered this	13	RCJ
	this pairing	17	RCJ
	was easy	13	RCJ
	word is	15	JOL

Appendix C:

Sample SVM Results Tables

Table 14

Bivariate SVM classification accuracy results by for JOLs

Confidence Level	20%	40%	60%	80%	100%
0%	84.14%	93.88%	99.63%	99.63%	98.36%
20%		87.55%	91.44%	91.96%	94.62%
40%			89.53%	89.12%	91.70%
60%				85.16%	91.74%
80%					88.64%

Note: The results here are the percentage of test cases classified correctly by the SVM algorithm and indicate the extent to which two levels of JOL confidence can be differentiated based on the content of their justifications.

Table 15

Bivariate SVM classification accuracy results by for RCJs

Confidence Level	20%	40%	60%	80%	100%
0%	88.44%	95.16%	96.23%	98.39%	96.85%
20%		83.86%	94.25%	96.31%	95.56%
40%			89.35%	91.44%	93.41%
60%				93.16%	92.72%
80%					90.68%

Note: The results here are the percentage of test cases classified correctly by the SVM algorithm and indicate the extent to which two levels of JOL confidence can be differentiated based on the content of their justifications.

Table 16

Bivariate SVM classification accuracy results by MC Judgment Type

Confidence	RCJ 0%	RCJ 20%	RCJ 40%	RCJ 60%	RCJ 80%	RCJ 100%
JOL 0%	75.46%					
JOL 20%		86.03%				
JOL 40%			82.07%			
JOL 60%				85.23%		
JOL 80%					88.27%	
JOL 100%						80.20%

Note: The results here are the percentage of test cases classified correctly by the SVM algorithm and indicate the extent to which the same confidence level for the two MC judgment types can be differentiated based on the content of their justifications.

Appendix D:

List of Stop-words

Stop-words included the following: "i", "the", "it", "this", "a", "to", "is", "of", "that", "but", "was", "in", "with", "and", "so", "because", "what", "my", "for", "me", "as", "at", "if", "an", "on", "there", "about", "out", "or", "too", "by"

References

1. Abbott, E. E. (1909). On the analysis of the factor of recall in the learning process. The Psychological Review: Monograph Supplements, 11(1), 159–177. https://doi.org/10.1037/h0093018

2. Ackerman, R., & Goldsmith, M. (2008). Control over grain size in memory reporting—with and without satisficing knowledge. Journal of Experimental Psychology: Learning, Memory and Cognition, 34, 1224–1245

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

4. Ariel, R., & Dunlosky, J. (2011). The sensitivity of judgment-of-learning resolution to past test performance, new learning, and forgetting. Memory & Cognition, 39(1), 171-184.

5. Azevedo, R., and Hadwin, A. F. (2005). Scaffolding self-regulated learning and metacognition - Implications for the design of computer-based scaffolds. Instructional Science 33: 367-379. DOI 10.1007/s11251-005-1272-9

6. Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. Behavior Research Methods, 39, 445-459.

7. Benjamin, A.S., Bjork, R.A., & Schwartz, B.L., (1998). The mismeasure of memory: when retrieval fluency is misleading as a metamnemonic index. Journal of Experimental Psychology: General *127*(1), 55-68.

8. Bol, L., Hacker, D.J., O'Shea, P., and Allen, D., (2010). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. The Journal of Experimental Psychology 73(4), 269-290. DOI: 10.3200/JEXE.73.4.269-290

9. Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, Word Frequency and Negative Recognition. Quarterly Journal of Experimental Psychology, 29(3), 461–473. https://doi.org/10.1080/14640747708400622

10. Costermans, J., Lories, G., & Ansay, C. (1992). Confidence level and feeling of knowing in question answering: The weight of inferential processes. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18, 142–150. doi: 10.1037/0278-7393.18.1.142http://dx.doi.org/10.1037/0278-7393.18.1.142

11. Cowan N. (2008). What are the differences between long-term, short-term, and working memory? Progress in brain research, 169, 323–338. https://doi.org/10.1016/S0079-6123(07)00020-9

12. Cowan, N. (2017) The many faces of working memory and short-term storage. Psychonomic Bulletin & Review, *24*(4):1158-1170. doi: 10.3758/s13423-016-1191-6.

13. Cutting, J.E. (1975). Orienting tasks affect recall performance more than subjective impressions of ability to recall. Psychological Reports. 36, 155-158.

14. Dennis, S. (2007). How to use the LSA web site. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), Handbook of latent semantic analysis (pp. 57–70). Lawrence Erlbaum Associates Publishers.

15. Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. Frontiers in Psychology, 5.doi: 10.3389/fpsyg.2014.00781

16. Dougherty, M. R., Scheck, P., Nelson, T. O., & Narens, L. (2005). Using the past to predict the future. Memory & Cognition, 33(6), 1096–1115. https://doi.org/10.3758/BF03193216

17. Dunlosky, J., & Bjork, R. A. (2008). The integrated nature of metamemory and memory. In J. Dunlosky and R. A . Bjork (Eds.) Handbook of Metamemory and Memory

18. Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOLs) and the delayed-JOL effect. Memory & Cognition, 20, 373-380

19. Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(1), 238.

20. Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. Journal of Memory and Language, 58(1), 19–34.

21. Flavell, J.H. (1979). Metacognition and Cognitive Monitoring; A new Area of Cognitive-Developmental Inquiry. American Psychologist, 34(10), 906-911

22. Gardiner, J. M, Ramponi, C., & Richardson-Klavehn, A. (1998). Experiences of remembering, knowing, and guessing. Consciousness and Cognition: An International Journal, 7(1), 1–26. https://doi.org/10.1006/ccog.1997.0321

23. Goldsmith, M., Koriat, A., & Weinberg-Eliezer, A. (2002). Strategic regulation of grain size in memory reporting. Journal of Experimental Psychology: General, 131, 73–95. DOI:10.1037//0096-3445.131.1.73

24. Hamel, L. H. (2009). Knowledge discovery with support vector machines. Hoboken, NJ: Wiley. doi:10.1002/9780470503065

25. Hanczakowski, M., Pasek, T., Zawadzka, K., & Mazzoni, G. (2013). Cue familiarity and 'don't know' responding in episodic memory tasks. Journal of Memory and Language, 69(3), 368–383. https://doi.org/10.1016/j.jml.2013.04.005

26. Hart, J. T. (1965). Memory and the feeling-of-knowing experience. Journal of Educational Psychology, 56(4), 208–216. https://doi.org/10.1037/h0022263

27. Hart, J. T. (1967). Memory and the memory-monitoring process. Journal of Verbal Learning and Verbal Behavior, 6, 685-691

28. Hertzog, C., Dunlosky, J., Robinson, A.E., & Kidder, D.P., (2003). Encoding Fluency is a cue used for judgments about learning. Journal of Experimental Psychology: Learning, Memory, and Cognition 29(1). 22-34. doi: 10.1037/0278-7393.29.1.22

29. Hertzog, C., Hines, J.C., and Touron, D.R. (2013). Judgments of learning are influenced by multiple cues in addition to memory for past test accuracy. Archives of Scientific Psychology 1: 23-32. DOI: http://dx.doi.org/10.1037/arc0000003

30. Hertzog, C., Kidder, D. P., Powell-Moman, A., & Dunlosky, J. (2002). Aging and monitoring associative learning: Is monitoring accuracy spared or impaired? Psychology and Aging, 17, 209–225. doi:10.1037/0882-7974.17.2.209

31. Hines, J. C., Touron, D. R., & Hertzog, C. (2009). Metacognitive influences on study time allocation in an associative recognition task: An analysis of adult age differences. Psychology and Aging, 24(2), 462–475. https://doi.org/10.1037/a0014417

32. Jersakova, R., Allen, R.J., Booth, J., Souchay, C, and O'Connor, A.R. (2017). Understanding Metacognitive confidence: Insights from judgment-of-learning justifications. Journal of Memory and Language 97, 187-207. http://dx.doi.org/10.1016/j.jml.2017.08.002

33. Jurka, T.P., Collingwood, L., Boydstun, A. E., Grossman, E., and van Atteveldt, W. (2012). RTextTools: Automatic Text Classification via Supervised Learning. R package version 1.3.9. http://CRAN.R-project.org/package=RTextTools

34. Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. Journal of Memory and Language, 32, 1-24.

35. Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. Psychological Review, 100, 609–639. doi: 10.1037/0033-295X.100.4.609http://dx.doi.org/10.1037/0033-295X.100.4.609

36. 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. https://doi.org/10.1037/0096-3445.126.4.349

37. Koriat, A. (2012). The Self-Consistency Model of Subjective Confidence. Psychological Review 119(1) 80-113. DOI: 10.1037/a0025648

38. Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. Psychological Review, 103, 490–517. DOI:10.1037/0033-295X.103.3.490

39. Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. Journal of Experimental Psychology: Human Learning and Memory, 6, 107-118

40. Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. Journal of Memory and Language, 52, 478–492. doi:10.1016/j.jml.2005.01.001

41. Koriat, A., Sheffer, L., and Ma'ayan, H., (2002). Comparing objective and subjective learning curves; Judgments of learning exhibit increased Underconfidence with practice. Journal of Experimental Psychology: General, 131(2), 147–162. https://doi.org/10.1037/0096-3445.131.2.147

42. Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The Ease-of-Processing Heuristic and the Stability Bias: Dissociating Memory, Memory Beliefs, and Memory

Judgments. Psychological Science, 22(6), 787–794. https://doi.org/10.1177/0956797611407929

43. Kuhn, M. (2022). caret: Classification and Regression Training. R package version 6.0-91. https://CRAN.R-project.org/package=caret

44. Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

45. Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25(2-3), 259–284. https://doi.org/10.1080/01638539809545028

46. Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259-284

47. Lateef, Z. (2020, May 15). Support Vector Machine in R: Using SVM to predict heart diseases. Support Vector Machine In R: Using SVM To Predict Heart Diseases. Retrieved April 7, 2022, from https://www.edureka.co/blog/support-vector-machine-in-r/

48. Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory?Journal of Experimental Psychology: Learning, Memory, and Cognition, 16, 464–470. doi: 10.1037/0278-7393.16.3.464http://dx.doi.org/10.1037/0278-7393.16.3.464

49. Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982) Calibration of probabilities: the state of the art to 1980. In: Kahneman, D., Slovic, P., & Tversky, A., (eds.) Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge, UK, pp. 306-334. ISBN 9780521284141

50. Luttrell, A., Briñol, P., Petty, R.E., Cunningham, W., & Díaz, D (2013) Metacognitive confidence: A neuroscience approach, Revista de Psicología Social, 28:3, 317-332, DOI: 10.1174/021347413807719148

51. Maki, Ruth H. , "Privileged Access for General Knowledge and Newly Learned Text Material" , in Handbook of Metamemory and Memory ed. John Dunlosky and Robert A. Bjork (Abingdon: Routledge, 28 May 2008), Routledge Handbooks Online.

52. Maniscalco, B., Peters, M.A.K., & Lau, H., (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. Attention Perception and Psychophysics. 78:923-937. DOI 10.3758/s13414-016-1059-x

53. McCabe, D. P., Geraci, L., Boman, J. K., Sensenig, A. E., & Rhodes, M. G. (2011). On the validity of remember–know judgments: Evidence from think aloud protocols. Consciousness and Cognition: An International Journal, 20(4), 1625–1633. https://doi.org/10.1016/j.concog.2011.08.012

54. Metcalfe, J. (1993). Novelty monitoring, metacognition, and control in a composite holographic associative recall model: Implications for Korsakoff amnesia. Psychological Review, 100(1), 3–22. https://doi.org/10.1037/0033-295X.100.1.3

55. Metcalfe, J., and Finn, B., (2008a). Evidence that judgments of learning are causally related to study choice. Psychonomic Bulletin & Review. 15(1), 174-179. doi: 10.3758/PBR.15.1.174

56. Metcalfe, J., & Finn, B. (2008b). Familiarity and retrieval processes in delayed judgments of learning. Journal of Experimental Psychology. Learning, Memory, and Cognition, 34(5), 1084–1097. http://dx.doi.org/10.1037/a0012580

57. Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue familiarity heuristic in metacognition. Journal of Experimental Psychology: Learning, Memory, & Cognition, 19, 851-861.

58. Miller, G. A., Galanter, E., & Pribram, K. H. (1960). Plans and the structure of behavior. Henry Holt and Co. https://doi.org/10.1037/10039-000

59. Morey, R.D., & Rouder, J.N. (2021). BayesFactor: Computation of bayes factors for common designs. R package version 0.9.12-4.3. https://CRAN.R-project.org/package=BayesFactor

60. Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. Psychonomic Bulletin & Review, 20, 378–384. doi:10.3758/s13423-012-0343-6

61. Mullen LA, Benoit K, Keyes O, Selivanov D, Arnold J (2018). "Fast, Consistent Tokenization of Natural Language Text." Journal of Open Source Software, 3, 655. doi:10.21105/joss.00655, https://doi.org/10.21105/joss.00655.

62. Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. Psychological Bulletin, 95, 109–133.

63. 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(4), 267–270. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x

64. Nelson T.O., & Narens L (1990) Metamemory: a theoretical framework and new findings. In: The psychology of learning and motivation (Bower GH, ed), pp 1–45. New York: Academic.

65. Nelson, T. O., Narens, L., & Dunlosky, J. (2004). A Revised Methodology for Research on Metamemory: Pre-judgment Recall And Monitoring (PRAM). Psychological Methods, 9(1), 53–69. https://doi.org/10.1037/1082-989X.9.1.53

66. Nguyen, T. Abed, E, & Pezdek, K. (2018). Postdictive confidence (but not predictive confidence) predicts eyewitness memory accuray. Cognitive Research: Principles and Implications 3:32 https://doi.org/10.1186/s41235-018-0125-4

67. Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. Behavior Research Methods. 10.3758/s13428-018-01193-y

68. Perfect, T. J., & Hollins, T. S. (1996). Predictive feeling of knowing judgements and postdictive confidence judgements in eyewitness memory and general knowledge. Applied Cognitive Psychology, 10(5), 371–382. https://doi.org/10.1002/(SICI)1099-0720(199610)10:5<371::AID-ACP389>3.0.CO;2-O

69. R Core Team. (2012). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

70. Rajaram, S., Hamilton, M., Bolton, A. (2002) Distinguishing states of awareness from confidence during retrieval: evidence from amnesia. Cognitive, Affective, & Behavioral Neuroscience 2(3):227-35. doi: 10.3758/cabn.2.3.227. PMID: 12775187.

71. Robey, A.M., Dougherty, M.R., and Buttaccio, D.R., (2017). Making Retrospective Confidence Judgments Improves Learners' Ability to Decide What Not to Study. Psychological Science 28(11). 1683-1693. https://journals.sagepub.com/doi/10.1177/0956797617718800

72. Ryals, A., Rogers, L., Gross, E., Polnaszek, K., & Voss, J. (2015). Associative Recognition Memory Awareness Improved by Theta-Burst Stimulation of Frontopolar Cortex. Cerebral cortex (New York, N.Y. : 1991). 26. 10.1093/cercor/bhu311.

73. 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.

74. Scheck, P. & Nelson, T. O. (2005). Lack of pervasiveness of the under confidence with-practice effect: Boundary conditions and an explanation via anchoring. Journal of Experimental Psychology: General, 134, 124–128.

75. Schwartz, B.L. (1994). Sources of information in metamemory: judgments of learning and feelings of knowing. Psychonomic Bulletin & Review. 1,357–375.doi:10.3758/BF03213977

76. Schwartz, B. L., & Metcalfe, J. (1992). Cue familiarity but not target retrievability enhances feeling-of-knowing judgments. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18(5), 1074–1083. https://doi.org/10.1037/0278-7393.18.5.1074

77. Selmeczy, D., & Dobbins, I. G. (2014). Relating the content and confidence of recognition judgments. Journal of Experimental Psychology: Learning, Memory, and Cognition, 40, 66–85.

78. Serra, M.J., and Ariel, R. (2014). People use the memory for past-test heuristic as an explicit cue for Judgments of learning. Memory and Cognition 42: 1260-1272. DOI 10.3758/s13421-014-0431-0

79. Siedlecka, M., Paulewicz, B., & Wierzchoń, M. (2016). But I was so sure! Metacognitive judgments are less accurate given prospectively than retrospectively. Frontiers in Psychology, 7, Article 218. https://doi.org/10.3389/fpsyg.2016.00218

80. Son, L. K., & Metcalfe, J. (2005). Judgments of learning: Evidence for a two-stage process. Memory & Cognition, 3, 1116–1129.

81. Susser, J. A., Panitz, J., Buchin, Z., & Mulligan, N. W. (2017). The motoric fluency effect on metamemory. Journal of Memory and Language, 95, 116–123. https://doi.org/10.1016/j.jml.2017.03.002

82. Tauber, S. K., & Rhodes, M. G. (2012). Multiple bases for young and older adults' judgments of learning in multitrial learning. Psychology and Aging, 27(2), 474–483. https://doi.org/10.1037/a0025246

83. Tullis, J.G., & Fraundorf, S.H., (2017). Predicting others' memory performance: The accuracy and bases of social metacognition. Journal of Memory and Language, 95, 124-137. http://dx.doi.org/10.1016/j.jml.2017.03.003

84. Undorf, M., and Erfelder, E. (2015). The Relatedness effect on judgments of learning: A closer look at the contribution of processing fluency. Memory & Cognition 43: 647-658. DOI 10.3758/s13421-014-0479-x

85. Wandmacher, T., Ovchinnikova, E., & Alexandrov, T. (2008). Does latent semantic analysis reflect human associations? In Proceedings of the lexical semantics workshop et ESSLLI'08. Hamburg, Germany.

86. Watier, N. W., & Collins, C. A. (2011). Metamemory for faces, names, and common nouns. Acta Psychologica, 138, 143–154.

87. Weber, N., & Brewer, N. (2008). Eyewitness recall: Regulation of grain size and the role of confidence. Journal of Experimental Psychology. Applied, 14, 50–60

88. Williams, H. L., Conway, M. A., & Moulin, C. J. A. (2013). Remembering and knowing: Using another's subjective report to make inferences about memory strength and subjective experience. Consciousness and Cognition: An International Journal, 22(2), 572–588. https://doi.org/10.1016/j.concog.2013.03.009

89. Yaniv, I., & Foster, D. P. (1995). Graininess of judgement under uncertainty: An accuracy-informativeness trade-off. Journal of Experimental Psychology: General, 124, 424–432. DOI:10.1037/0096-3445.124.4.424.

90. Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20, 1341–1354. doi: 10.1037/0278-7393.20.6.1341

91. Zhang, Z., & Yuan, K.-H. (2018). Practical Statistical Power Analysis Using WebPower and R (Eds). Granger, IN: ISDSA Press. [https://webpower.psychstat.org]