Understanding Metacognitive Confidence: Insights from Judgment-of-Learning Justifications

Radka Jersakova[1], Richard J. Allen[1], Jonathan Booth[2], Céline Souchay[3] and Akira R. O'Connor[2]

[1] School of Psychology, University of Leeds , Leeds, LS2 9JT, UK

[2] School of Psychology and Neuroscience, University of St Andrews, St Andrews, KY16 9JP, UK

[3] Laboratoire de Psychologie et Neurocognition CNRS UMR 5105, Université Grenoble Alpes, Grenoble, 38040, France

Corresponding author:

Radka Jersakova, School of Psychology, University of Leeds, Leeds, LS2 9JT, UK.

Email: r.jersakova@hotmail.com

Abstract

This study employed the delayed judgment-of-learning (JOL) paradigm to investigate the content of metacognitive judgments; after studying cue-target word-pairs, participants predicted their ability to remember targets on a future memory test (cued recognition in Experiments 1 and 2 and cued recall in Experiment 3). In Experiment 1 and the confidence JOL group of Experiment 3, participants used a commonly employed 6-point numeric confidence JOL scale (0-20-40-60-80-100%). In Experiment 2 and the binary JOL group of Experiment 3 participants first made a binary *yes/no* JOL prediction followed by a 3-point verbal confidence judgment (*sure-maybe-guess*). In all experiments, on a subset of trials, participants gave a written justification of why they gave that specific JOL response. We used natural language processing techniques (latent semantic analysis and word frequency [*n*-gram] analysis) to characterize the content of the written justifications and to capture what types of evidence evaluation uniquely separate one JOL response type from others. We also used a machine learning classification algorithm (support vector machine [SVM]) to quantify the extent to which any two JOL responses differed from each other. We found that: (i) participants can justify and explain their JOLs; (ii) these justifications reference cue familiarity and target accessibility and so are particularly consistent with the two-stage metacognitive model; and (iii) JOL confidence judgements do not correspond to *yes*/*no* responses in the manner typically assumed within the literature (i.e. 0-40% interpreted as *no* predictions).

Keywords: metacognition, judgments-of-learning, episodic memory, confidence, linguistics

Understanding Metacognitive Confidence: Insights from Judgment-of-Learning Justifications

Cognitive processes are accompanied by states of awareness that guide evaluation of their function and content (Fleming, Dolan, & Frith, 2012; Nelson & Narens, 1990; Overgaard & Sandberg, 2012). This metacognitive awareness (or monitoring) is understood as an inferential process, relying on cues derived from the task at hand to construct judgments about performance (Koriat, 2000), that has behavioral consequences (Koriat, Ma'ayan, & Nussinson, 2006; Metcalfe & Finn, 2008a). As such, understanding the basis on which these metacognitive judgments are made is crucial. While there have been numerous paradigms developed for the study of metacognition, subjective report from participants remains a vital method for tapping into metacognitive and related processes (Jersakova, Moulin, & O'Connor, 2016; Overgaard & Fazekas, 2016). Confidence in particular is the hallmark of metacognitive judgments and the most commonly used paradigm for investigating metacognition across domains, ranging from decision making and reasoning (Ackerman & Thompson, 2014; Fletcher & Carruthers, 2012; Yeung & Summerfield, 2012) to perceptual judgments (Fleming et al., 2015; Peters & Lau, 2015; Rahnev, Koizumi, McCurdy, D'Esposito, & Lau, 2015) and memory evaluations (Dunlosky, Serra, Matvey, & Rawson, 2005; Finn & Metcalfe, 2007; Koriat & Levy-Sadot, 2001).

Metacognitive confidence is often interpreted as corresponding to quantity and quality of some (internal) evidence gathered toward the judgment being made (e.g. ease of reading as evidence that an item has been sufficiently learned and will later be remembered; Rhodes & Castel, 2008) and reflecting the probability that the given judgment is correct (Kepecs & Mainen, 2012). Whereas metacognitive research has tended to focus on examining which variables lead to general shifts in confidence (e.g. Alban & Kelley, 2013; Castel, McCabe, & Roediger, 2007; Koriat & Levy-Sadot, 2001; Rhodes & Castel, 2008), there is less understanding of what expressed metacognitive confidence *means.* This includes considerations of what differentiates one confidence level (e.g. 40% confidence) from another (e.g. 60% confidence) and whether confidence judgments simply rank items against each other or whether they can be further interpreted (e.g. as *yes/no* predictions). Understanding this has implications for both theory and practice and our ability to interpret participant behavior in the laboratory. In this study we focused on metacognitive judgments made about memory (metamemory) to investigate what expressed metacognitive confidence represents.

We employed the delayed judgments-of-learning (JOL) paradigm; a prediction of whether recently learned information would be successfully retrieved in the future (Nelson & Dunlosky, 1991). In a typical delayed JOL experiment, participants study cue-target word pairs following which they are again presented with the studied cues and asked to make a prediction about whether they think they would retrieve the target on the subsequent memory test. These predictions are usually made on a numeric confidence scale expressed as percentages; e.g. 0%-20-40-60-80-100%. This study evaluated how participants construct and justify their delayed JOLs by asking them to provide written reports alongside their JOLs. Participants were given no instructions on how to write their justifications, as we wanted to see what features would be referenced spontaneously. We used natural language processing techniques to investigate the type of information and explanation that characterizes each JOL and differentiates one JOL from another (e.g. 20% vs. 40%), as well as to quantify the extent to which any two JOLs are justified with reference to different types of evidence.

The experiments presented here draw on research investigating retrospective confidence in contents of memory retrieval, which has established that probing participants for explanations and justifications of their answers is a powerful tool for characterizing processes underlying cognition and metacognition. For example, Koriat et al. (1980) asked participants to list reasons for and against their chosen answer to a general knowledge question. They observed that confidence was influenced by the amount of evidence accessed in support of the given answer, lending support to the idea that confidence is a result of a process of evaluation of different sources of evidence. More recently, Selmeczy and Dobbins (2014) asked participants to justify their confidence in recognition judgements. Analysis of these justifications showed a pattern of results consistent with dual-process accounts of recognition memory (see Yonelinas, 2002); for example, the presence of 'remembering' characterized high confidence *old* responses and its absence corresponded to high confidence *new* responses. In other words, this quantitative analysis of subjective reports lent support to one side of an on-going debate in recognition memory.

Furthermore, these results were obtained without explicit instructions or theory-laden manipulations from the experimenters, who did not highlight specific experiences or types of evidence for participants to focus on. This is in contrast to classic metamemory research, which relies largely on explicitly asking participants about access to specific types of information relating to the studied items (e.g. the degree to which they can remember partial characteristics, such as the first letter, of the target item, Koriat, 1993). Such an approach allows for the evaluation of how access to specific features of the studied items influences confidence judgments. However, it leaves open the question whether participants would rely on this type of information in their judgments if their attention was not drawn to it by asking (see Hertzog, Fulton, Sinclair, & Dunlosky, 2014). Overall, studies that have asked participants to justify their responses (see also Gardiner, Ramponi, & Richardson-Klavehn, 1998; Urquhart & O'Connor, 2014; Williams, Conway, & Moulin, 2013) indicate that much can be learned from the relatively infrequent practice of asking participants to explain their metacognitive judgments. In the present study we adopted and developed the analytical approach to participant justifications pioneered by Selmeczy and Dobbins (2014) to gain insight into processes underlying JOL confidence.

Turning specifically to the theoretical issues that could be informed by this approach, there is a debate about how numeric confidence JOL responses relate to a binary (*yes/no*) sub-classification of the scale. The idea that low confidence JOL predictions should equate to a rejection of future retrieval makes sense probabilistically (i.e. a 40% predicted success rate should correspond to a 60% predicted failure rate). Correspondingly, it is common practice to interpret confidence as representing success probabilities as evidenced by the use of calibration measures (e.g. Finn & Metcalfe, 2007; Koriat, Sheffer, & Ma'ayan, 2002; Serra & England, 2012). Similarly, in cases where binary data is required for analysis purposes, confidence responses are commonly split equally into a binary (*yes*/*no*) sub-classification (e.g. Hanczakowski, Zawadzka, Pasek & Higham 2013, Masson & Rotello, 2009).

There is some theoretical support for the idea that participants explicitly make a *yes/no* sub-classification in their interpretation of the confidence scale, which has been suggested for a range of metacognitive tasks (Dunlosky et al., 2005; Liu, Su, Xu, & Chan, 2007). For example, Dunlosky et al. (2005) observed that when participants were asked to make a confidence judgment about the accuracy of their JOL prediction (a second-order judgment; SOJ), a plot of the SOJ magnitude against JOL confidence yielded a U-shaped function – participants were most confident in the predictions that lay on the extremes of the JOL scale (with least confidence at the mid-range of the scale). Dunlosky et al. (2005) interpreted the SOJ function minimum as the point where *yes* and *no* predictions diverge and suggested that one possible interpretation is that JOLs could be viewed as a two-step process that consists first of a *yes* or *no* judgment, directly followed by an assignment of confidence.

While the binary *yes/no* sub-classification seems intuitive and plausible, it has not yet been directly tested. Further, its relationship to confidence, such as whether we can split the numerical scale into equal proportion of *yes* and *no* responses, is poorly understood. This is crucial since it is a theory-laden interpretation of JOLs that is commonly employed in the literature when analyzing confidence data and yet one that lacks explicit support. This absence of verification of widely held interpretations of how participants respond highlights the need to understand better how confidence and binary *yes/no* judgments relate to each other.

An alternative approach to understanding JOL confidence has been to investigate the underlying processes that shape the formation of JOLs. The early literature focused on explaining JOLs as a result of single process (target retrieval) evaluations (e.g. Nelson & Dunlosky, 1991). In this view, it was assumed that participants accrue one type of evidence (the degree to which the target is accessible) toward their JOL—the more evidence they collect, the higher their JOL. According to this view, different JOLs (e.g. 60% as compared to 80%) express different degrees of access to the target. An alternative two-stage view has proposed a quick pre-retrieval stage driven by cue-familiarity followed by an effortful memory search (target accessibility evaluation;

Benjamin, 2005). Metcalfe & Finn (2008b) further elaborated this view, suggesting the first stage can result in: (i) a quick *don't know* decision driven by lack of cue familiarity (expressed as responding with the lowest point on the JOL scale); or (ii) the initiation of the second effortful retrieval stage. In this case, there should be qualitatively different processes that underlie the lowest confidence JOL (i.e. 0%) and distinguish it from all others. More specifically, it should be a cue-driven evaluation as compared to a target-based judgment. If this holds, we would expect participants to refer to these different types of evidence in their justifications and to observe a qualitative difference in the evidence favored at different levels of the JOL scale.

Thus, there are two modes of understanding JOL confidence; an interpretative model proposing what confidence represents (a binary *yes* or *no* judgment) and a descriptive model defining what determines a given confidence judgment (e.g. level of access to the target). These two views are not irreconcilable; they both suggest there is an underlying point of divergence in the JOL scale either side of which the scale is characterized by different processes. The descriptive model (Metcalfe and Finn, 2008b) places that point on the lowest end of the scale and describe it in terms of the information evaluation processes that change at that point such that a 0% JOL is a decision made as the result of a different process (cue familiarity) than the process that characterizes JOLs on the rest of the scale (target accessibility). In some ways that could also be interpreted as a *yes/no* divergence with lack of cue familiarity leading to a 0% (or a *no*) judgment and the rest of the scale representing different degrees of a *yes* judgment corresponding to different levels of target access. Consistent with this view is a study by Dougherty, Scheck, Nelson and Narens (2005) who evaluated JOL accuracy at predicting subsequent item retrieval for all possible JOL dyad comparisons (e.g. 0% JOL vs. 20% JOL, 0% JOL vs. 40% JOL and so on for all possible pairings). Dougherty et al. observed that JOLs predicted which items were subsequently remembered most accurately when they compared 0% JOLs against all other judgments (20%, 40%, 60%, 80% and 100%) in contrast to any other dyad comparisons. This is consistent with the idea that 0% JOL represents a rejection of future retrieval whereas all other JOL responses correspond to a prediction of the item being subsequently retrieved. There is thus a lot of scope for combining the two views. However, as previously stated, this position is not reconcilable with the common practice of re-interpreting confidence responses in binary terms and assigning *yes* and *no* labels by splitting the confidence scale down the middle (the interpretative model, see e.g. Mason & Rotello, 2009).

The aim of the present study was to shed light on what JOL confidence represents by drawing on and contrasting the interpretative and descriptive models of delayed JOL confidence. Across all experiments, participants completed a standard delayed JOL task with cue-target word pairs. In Experiment 1 participants made JOL predictions on a 6-point numeric confidence scale (0-20-40-60-80-100%) whereas in Experiment 2 participants made first a binary *yes*/*no* JOL prediction followed by a 3-point verbal confidence judgment made about that prediction (*sure*-*maybe*-*guess*). Thus, in both experiments there were a total of six JOL response options. In Experiment 3 participants were similarly randomly assigned to one of two experimental conditions, the confidence JOL condition (same response format as in Experiment 1) and the binary JOL condition (same response format as in Experiment 2). Whereas in Experiment 1 and 2 participants predicted recognition performance when making their JOLs, in Experiment 3 they predicted recall. This allowed us to test the generalizability of our findings across different experimental contexts.

The key novel feature of the present study was that participants provided written justifications on a subset of their JOLs, which were then analyzed using natural language processing techniques. Participants were not given any instructions on how to write their justifications nor did we manipulate any additional variables known to influence JOL confidence. Three methods of text data analyses (described in more detail in the Methods section) were used to examine in detail the content of these justifications as well as evaluate differences between justifications for different JOL responses. Latent Semantic Analysis (LSA) allowed us to evaluate whether the justifications were more likely to refer to the cue or the target term. This was followed by an *n*-gram analysis, which isolates unique phrases that are significantly more likely to occur in justifications for one JOL category (e.g. 0%) as compared to all others. We examined these for references to processes such as familiarity and remembering. Lastly, we used Support Vector Machines (SVMs) to quantify the extent to which two sets of JOL responses (e.g. *no-guess* and *yes-guess*) differed from each other. If different types of evidence are referenced (e.g. cue familiarity vs. target accessibility) as compared to levels of access to the same evidence (e.g. different degrees of target access), then accuracy of SVM classification between them should be high. Metcalfe and Finn (2008b) make a clear prediction that 0% JOLs should reference the cue (and its lack of familiarity) whereas other JOL responses should, with increasing confidence, increasingly focus on the target. This should also lead to high SVM classification accuracy between justifications written for 0% vs all other confidence JOLs. It is not clear whether we could expect a similar pattern of results for binary *yes/no* JOLs. Altogether, these three analyses allowed us to characterize the content of JOL justifications, with a focus on the role of the cue and the target, and to also explore whether confidence and binary JOLs directly map onto each other (which would be reflected in a similar pattern of justifications).

In summary, we assessed how participants arrive at JOL confidence independently and spontaneously without experimentally making any one source of information (e.g. cue familiarity) more salient than others. We also assessed whether and how confidence and binary *yes/no* JOLs compare with each other in terms of the underlying influences that participants reference in their justifications. This was done in the context of participants predicting future recognition performance when making their JOLs (Experiment 1 and 2) and participants predicting recall performance (Experiment 3). The general procedure and majority of methods adopted to analyse the text data were modeled on Selmeczy & Dobbins (2014). Within each experiment we thus examined: (i) how participants justify their JOLs and (ii) to what extent are such justifications characterized by cue and target references and whether this is consistent with the descriptive model. Comparing the pattern of justifications for confidence and binary JOLs allowed us to investigate (iii) whether there is an underlying *yes*/*no* sub-classification in numeric JOL confidence responses and, if yes, where it lies. Investigating both recognition and recall JOL predictions provided an indication of the replicability and generalizability of our findings to different experimental contexts.

Experiment 1 and 2

Method

Participants

All participants were native English speakers affiliated with the University of Leeds (students and staff) with 54 participants in Experiment 1 (13 men; mean age = 23.4; *SD*

= 7.4) and 73 participants (12 men; mean age = 27.5, *SD* = 10.7) in Experiment 2.[1] In Experiment 1, two participants were excluded, both for not following instructions (one for using only 0% and 100% judgements, the other because their written responses referred to multiple cue-target pairs instead of the pair preceding the written report). This left 52 participants in the analysis for Experiment 1 (13 men, mean age = 22.5, *SD* = 6.2). In both experiments, participants either received course credit or £5 as reimbursement. The study was granted ethical approval by the School of Psychology Ethics Committee, University of Leeds, UK.

Stimuli

The stimuli used in both experiments were selected from a list of 628 common, singular English nouns (5-6 letters long) taken from the English Lexicon Project (minimum log Hyperspace Analogue to Language frequency 8.02; Balota et al., 2007). For each participant, an algorithm randomly selected words from the list and formed them into cue-target word-pairs. Each participant was thus exposed to a unique set of 90 cue-target pairs (45 in each of the two experimental blocks). This meant that we did not control for associative strength between the cue and the target but also that the observed effects would not be specific to and limited by the nature of the word-pairs studied.

Procedure

The study was programmed using PsychoPy (Peirce, 2007), with all participants completing the delayed JOL task individually on a computer, in the presence of the

[1] The primary data of interest were the written justifications. In both experiments each participant could provide at most 3 justifications per JOL type with most participants providing fewer justifications than the allowed maximum. Further, participants used some JOL responses more commonly than others. In Experiment 2 this was especially pronounced with, for example, the *yes-guess* JOL response being used relatively infrequently by all participants. At the outset of Experiment 2 we decided that we wanted to collect at least 100 justifications per JOL response type to make the dataset comparable to Experiment 1. This condition and the less evenly distributed nature of JOL responding in Experiment 2 meant that the sample size of Experiment 2 was necessarily larger than that of Experiment 1.

experimenter. The JOL task is constrained by the number of word-pairs a participant can be expected to memorize in one session so to collect a sufficient number of JOL justifications, participants completed the whole task twice in two identical blocks consisting of three consecutive phases (with no breaks or delays between them, see Figure 1). In each block participants: (i) studied 45 cue-target pairs-presented for 6000ms each with a fixation cross (500ms) between all trials; (ii) were presented with the cue of each pair, and gave a JOL predicting performance for the target on the subsequent recognition memory test[2]; and (iii) completed a forced choice recognition test where, on presentation of each cue, they selected the cue-matched target from two words (both options were targets from the study and each target appeared on the recognition test twice, once as a lure and once as the correct response). All responses were made by pressing a key corresponding to the confidence response or target. The tasks were completed consecutively without any intervening breaks. The order in which items were presented in each phase of each block was randomized. In each block, participants were exposed to a different set of 45 cue-target pairs (90 in the whole experiment).

The only difference between the two experiments was in the JOL stage (part ii of the procedure). In Experiment 1, participants gave their JOLs on a 6-point numeric confidence scale (0-20-40-60-80-100%). In Experiment 2, participants first gave a binary y*es/no* response indicating whether they would recognize the target, followed by a 3-point verbal confidence judgment (*sure-maybe-guess*) relating to the *yes*/*no* response. Thus, in both experiments, there were 6 distinct JOL response options participants could give.

[2] The nature of the associative recognition task was made clear to participants as part of the instructions. Participants therefore knew what the upcoming memory test was and what they were predicting performance for.

On a subset of the judgment trials, immediately after giving a JOL, participants justified the previously rendered JOL using a written, keyboard-entered response. Over the two blocks participants could give a maximum of 18 justifications (9 per block)—3 per JOL response option. More specifically, no questions were asked on the first five trials of either block. After that, requests for written justifications were spread out throughout the judgment task as follows. If the maximum number of justifications was reached for a given JOL response type, no more justifications were asked for that response option. Participants would not be asked for any written responses for the two trials following a justification, though this enforced gap reduced over the course of the block (there was no enforced justification gap for the last 10 trials). Some participants therefore gave fewer judgments than others, especially since some participants used some JOL responses less than others. On average participants gave 15.4 justifications in Experiment 1 (*SD* = 1.9) and 12.7 justifications in Experiment 2 (*SD* = 2.7).

<Figure 1>

Text analysis methods

Text data pre-processing. Before any text analysis was carried out, we corrected spelling mistakes in the text and removed articles (*a* and *the*). We also removed justifications where participants explicitly indicated they wanted to change the previously rendered JOL response (in total, three justifications in Experiment 1, six justifications in Experiment 2). In all reported analyses, we aggregated the descriptive reports for each JOL confidence level and response type across participants for comparison. A minimum of 102 justifications was collected per JOL type (see Table 1 for number of justifications collected per JOL response category).

Latent Semantic Analysis (LSA). LSA is a technique by which one can evaluate the semantic relationship between a single term and a text document. Drawing on singular value decomposition (closely related to factor analysis), LSA creates a mathematical matrix representation of a large body of text, mapping the semantic relationships between single words and sets of words. This mapping relies on frequency of co-occurrence but also on a weighting function that takes into account the 'importance' of a term to a given text (see Landauer, Foltz & Laham, 1998 for more detail). LSA that has been trained on a relevant corpus of texts (e.g. general or subject specific) to create this representation, also called semantic space, can then be applied to new examples to compute their semantic relationship. The subsequent classification of semantic similarities between new examples very closely imitates humans (e.g. Laham, 1997).

The online LSA tool (available at http://lsa.colorado.edu/) offers a semantic space that has been trained on 'general reading' corpus with 300 factors (Dennis, 2006). We used this to classify the semantic similarity between each justification and the cue-target pair it was written in response to. The toolkit returns a cosine value for each comparison; as such the range of output values is -1 to 1, with 0 or lower interpreted as no semantic relationship. Following Wandmacher, Ovchinnikova, and Alexandrov (2008), we set negative LSA values to 0 since in this context we could not interpret a justification and a studied item (cue or target) as being more dissimilar than 'not similar at all'.

Specifically, we computed an LSA score between the cue and the justification and compared it against the LSA score computed between the target and the justification. For example, if a justification for a given JOL response type is more likely to refer to the cue than the target (e.g. "I cannot remember studying the word truth" where 'truth' is the cue) then the LSA value should be higher for the cue-justification as compared to the target-justification comparison. This enabled us to assess whether any JOL category was characterized by referring more to the cue or the target, as predicted by Metcalfe and Finn's (2008b) two-stage JOL account.

Word frequency analysis (n-grams). An *n*-gram is a continuous series of words found to occur within a text (*n* = 1, 2, 3 are referred to as *uni*-grams, *bi*-grams and *tri*-grams respectively). To compare sets of texts (in this case, justifications) the frequency of occurrence of each *n*-gram is counted across all justification texts. To account for some participants writing more than others (and possibly repeating themselves), we restricted the analysis so that each JOL justification could contribute a maximum of 1 to any given *n*-gram count. For any given *n*-gram (e.g. "do not remember") we could thus compute the total number of justifications that contained it for each JOL category.

In previous experiments analysing *n*-grams (Selmeczy & Dobbins, 2014; Urquhart & O'Connor, 2014), only two categories with equal probability of occurrence were ever compared against each other. This was done using a binomial test, computing a *p*-value for the proportion of occurrence of the given *n*-gram under one response category assuming a binomial distribution with the *p*-parameter of 0.5. This allowed for the examination of whether the *n*-gram was significantly more likely to appear in justifications for one response category or whether the probability of it occurring in texts justifying either response category was equal. Here we contrasted each JOL category (e.g. 0%) against all other JOL categories (e.g. 20%-100%). As such we set the *p*-parameter as the number of justifications written in the given category divided by the total number of justification written within the whole experiment as this better reflected the probability of a given *n*-gram occurring equally likely in any of the collected justifications. For example, in the case of the 0% JOL, the *p*-parameter was set to 127/796. In other words, for each JOL category, we computed whether the proportion of occurrence (out of all occurrences) of any *n*-gram was significantly higher than that *n*-gram having equal probability of occurrence in justifications of all JOL response categories.

This analysis allowed the isolation of simple phrases that were most likely to be used in justifying one JOL response type as compared to all others. Where LSA focused on semantic similarity between the studied items (cue and target) and the justification texts, *n*-gram analysis examined whether different phrases (e.g. relating to familiarity as compared to retrieval success) would differentiate different JOL response categories. Rather than analysing information specific to each trial (i.e. whether participants named or referred to the studied items), this analysis enabled the extraction of general phrases that held true across trials, irrespective of what the studied cue or target were. In this way the *n*-gram analysis complemented, and helped to further explicate, the LSA results.

Classification analysis (Support Vector Machine [SVM]). SVM is a machine-learning algorithm commonly used in text classification. Here we employed it as a tool for quantifying the extent to which different JOL responses differed from each other. If there are highly distinct features that separate one category from another (such as references characteristic of different processes), then the SVM would pick up on this and classification of future examples would be highly accurate. On the other hand, if the differences were merely of degree (e.g. different levels of target access), then the classification of future examples would be low.

To carry out SVM analysis, we represented each written justification as a vector where each vector component corresponded to a *uni*-gram, *bi*-gram or *tri*-gram, with 0 denoting its absence in the given justification text and 1 denoting its presence. We included all *n*-grams as this allowed us to account for individual word usage as well as word combinations, which carry specific semantic meaning. For example, the *uni*-grams 'not', 'remember' and 'confident' could only be coded as present once which would mark the texts 'I am confident I will not remember' and 'I am not confident but might remember' as the same while including the bigrams 'not remember' and 'not confident' avoided this problem. Each *n*-gram thus constituted an input feature and each text was represented as a vector of features while the output was the JOL category the given vector belonged to (e.g. 0%). In principle, an SVM looks for a 'decision boundary' or a line that separates the two sets of data being compared so that the distance between the boundary and any point of any class is the biggest it can possibly be—that is why it is called a maximum-margin classifier (Hamel, 2009). Once an SVM has been trained it can be used to classify new data which will be assigned either of the categories the SVM has been trained on, based on which side of the margin it falls.

The SVM analysis was implemented with scikit-learn, an open source toolkit developed for Python (Pedregosa et al., 2011). To compare two JOL response categories (e.g. 0% vs. 20% JOL), the justification responses for both were labeled and combined. We trained the classifier on a randomly selected half of the combined data with a linear kernel and a cost value of 0.10 and tested it on the other half. Once the classifier was trained, it was then used to classify the remaining half of the data, and its performance was evaluated by its ability to correctly distinguish the JOL for which a given text was written.

The interpretative and descriptive JOL confidence models described in the Introduction both speculate a divergence on the confidence scale with regards to the processes that drive the judgment. A difference in processes relied upon (i.e. a qualitative difference) should lead to high classification accuracy whereas differences merely of degree (i.e. quantitative differences) should lead to low classification accuracy due to low likelihood of distinct, differentiating features.

In summary, the LSA allowed us to investigate whether different JOL responses were more likely to *semantically* reference the cue or the target. The *n*-gram analysis on the other hand allowed us to isolate unique phrases that were significantly more likely to be used for a JOL response category as compared to all others (e.g. familiarity or remembering references). Together, these analyses thus allowed us to describe what types of evidence feed into and differentiate different JOL responses. The SVM analysis allowed us to quantify the extent to which justifications for any two JOL response categories differed from each other. If different types of evidence are referenced (e.g. cue familiarity vs. target accessibility), then classification should be high.

We compared these results against the descriptive model (Metcalfe & Finn, 2008b) which predicts high classification accuracy between the responses at the low end of the numeric JOL scale (0% vs 20%) with the lowest confidence responses characterized primarily by cue familiarity and the remaining responses characterized by differing levels of target access. Contrasting the pattern of responses across the two response formats allowed us to evaluate the interpretative model which predicts that the classification accuracy at the boundary between *yes* and *no* predictions should be high and that the justifications for binary JOLs should directly map onto justifications for the numeric JOL scale in content and character.

Results

Memory and JOL responses

In Experiment 1, participants correctly recognized 84.7% (*SD* = 11.6) targets on the final memory test. In Experiment 2 they correctly recognized 86.2% (*SD* = 12.2) of targets. Memory performance did not differ between the two experiments, $t(123) = 0.67$, $p = .507$, $d = 0.12$.

To examine JOL prediction accuracy, in Experiment 1, average JOL confidence expressed for recognized vs. unrecognized targets was compared. Participants indicated higher JOL confidence for items they recognized (*M* = 46.84, *SD* = 13.82), compared to items they did not recognize (*M* = 24.96, *SD* = 15.88), $t(49) = 11.62$, $p < .001$, $d = 1.46$. To assess JOL accuracy in Experiment 2, percentage of *yes* JOL predictions was compared for recognized vs. unrecognized targets. The results revealed a higher percentage of *yes* JOLs for recognized (*M* = 57.12%, *SD* = 21.99) as compared to not recognized (*M* = 28.88%, *SD* = 28.47) items, $t(69) = 10.26$, $p < .001$, $d = 1.08$. Across both experiments, overall JOL predictions accurately predicted subsequent memory performance. See Figure 2 for the mean proportion of trials each JOL category was used and Table 1 for the number of written justifications collected per JOL category.

<Figure 2>

<Table 1>

Latent semantic analysis (LSA)

Metcalfe & Finn (2008b) proposed that the lowest point on the JOL confidence scale should reflect the result of a cue-evaluation stage whereas all other JOL levels should correspond to target access evaluations. We used LSA to evaluate whether for each JOL response type, participants were more likely to refer semantically to the cue or the target in their justifications (or neither). For each trial with a JOL justification, we computed an LSA value between the cue and the written justification and compared it against the LSA value computed between the target and the justification. Because the written justifications refer to specific memories, one could expect that overall the semantic similarity scores would be relatively low. However, if participants refer specifically to the cue or the target term (or information relating to them) this would increase the score. Additionally, because LSA has been shown to successfully map to meaning (Laham, 1997), an increase in the LSA score should be observed even when participants did not directly refer to the cue or the target but, for example, reported partial semantic information about them. We used paired-samples *t*-tests to compare the cue-justification and target-justification LSA scores for each JOL response category (e.g. 0% JOL confidence) to analyze whether the JOL justifications were more likely to refer to the cue or the target term. The LSA scores range from 0 (no relationship) to 1 (high semantic relationship). This analysis was done for both Experiment 1 and 2 separately with the results reported in Table 2.

<Table 2>

The results of the LSA revealed that in Experiment 1, the 0% and 20% JOL confidence level justifications were significantly more likely to semantically refer to the cue than the target. On the other hand, the 100% level was more significantly likely to refer to the target than the cue. The pattern of results of Experiment 2 showed it was the *guess* responses (for both *no* and *yes* predictions) that were significantly more likely to refer semantically to the cue rather than the target term. These results demonstrate that participants rely on both cue and target related information in justifying their JOLs and that these two types of processes provide a useful framework for differentiating different types of JOL predictions. To understand more precisely whether the cue-references were the same or differed between the different JOL responses we turned to word-frequency analysis.

Word-frequency analysis

The next step in the analyses was the examination of unique phrases that differentiated one JOL response from all others. This allowed us to determine whether the cue references in JOL justifications were of the same character (e.g. expressing lack of cue familiarity) or whether they relied on the cue term differently (e.g. cue familiarity characterizing 20% whereas its absence characterizing 0% JOL). Further, whereas LSA only tracked semantic similarity, participants could express lack of cue familiarity without naming the cue itself (e.g. "This cue is not familiar"). Compared to LSA, *n*-gram analysis thus allowed us to capture these types of phrases and extract meaningful patterns of expression across trials that were significantly more likely to occur for one type of JOL response as compared to others. For example, we expected to see an increase in recollection-specific terminology with increases in JOL confidence as well as greater use of intensity modifiers indicating greater certainty of access.

To constrain the number of *n*-grams analysed, we focused only on *bi*-grams and *tri*-grams with a minimum total occurrence of 10 (stricter than previous analyses which have included *uni*-grams and used lower median occurrences). We only reported *tri*-grams and *bi*-grams reaching significance at $p < .05$ (Table 3 reports *n*-gram analysis results for Experiment 1, Table 4 for Experiment 2). For each JOL, the analysis extracted phrases that occurred significantly more often than would be expected if the phrase was used equally across all JOL responses. Notably, this does not preclude the possibility that certain phrases might have significantly higher proportion of occurrence for two JOL category responses (e.g. if they never occurred for any other response) and thus allows for extraction of similarities (e.g. are there certain phrases that characterize *no* predictions that are never employed in *yes* predictions) as well as the expected characterization of differences.

<Table 3>

The *n*-gram analysis results presented in Table 3 show the 0% JOL confidence level was characterized by an inability to remember ("do not remember") and could be interpreted as expressing lack of cue familiarity as participants indicated they cannot even remember having seen the presented word at study ("not remember seeing"). The 20% JOL confidence level on the other hand was characterized by a vague sense of cue familiarity ("vaguely remember seeing [word]") accompanied by a lack of recollection for the target term ("but cannot remember"… "what it was"). While the LSA results revealed that the 0% and 20% JOL confidence levels were more likely to refer to the cue than the target term semantically, the *n*-gram analysis showed they nevertheless differed from each other in whether the cue term was said to be remembered. The 40% JOL also referenced cue familiarity ("I remember seeing") suggesting the role of the cue in JOLs isn't isolated to lowest confidence responses when it isn't familiar but can in itself provide a degree of evidence when the target cannot be accessed. Indeed, justifications for the 40% and 60% JOL confidence levels expressed feelings of possible target access ("I think I"… "could recognise"… "but cannot recall") whereas the 80% JOL confidence level started bringing in language of certainty ("pretty sure") and memory for associations ("I associated"). Unsurprisingly, the 100% JOL expressed memory for the target term ("I can remember"). All in all, this pattern of descriptions fits with Metcalfe and Finn's (2008b) suggestions that a lack of cue familiarity leads to a 0% JOL confidence response whereas, when the cue is recognized, the JOL confidence increases with increase in target access. The results further demonstrated that the role of the cue does not stop after that initial stage and is carried as evidence through to the target access stage.

<Table 4>

As seen in Table 4, the types of descriptions for the highest confidence *no* and *yes* responses respectively correspond to 0% (e.g. "cannot remember") and 100% (e.g. "I remember") responses of Experiment 1. It is noteworthy that the high numeric confidence JOLs and *yes* JOL predictions refer to not just the target, but also memory for the "word association" or "link between" the items. This supports recent findings that memory for associations made between the cue and the target at study influences metacognitive confidence (Hertzog et al., 2014) and demonstrates that this is true even when participants are not instructed to use any specific memory techniques in learning the cue-target pairs.

The *guess* responses (for both *yes* and *no* JOL predictions), were relatively low on unique *n*-gram use compared to most of the other JOLs. The LSA results revealed that participants were more likely to reference the cue than the target for these responses but the *n*-gram results are not clear as to which way this was. However, the *tri*-gram "not remember seeing" occurred 10 times in justifications for the *no-guess* responses (with further 12 occurrences for *no* – *sure* and 9 occurrences for *no* – *maybe* out of a total of 32 occurrences). Altogether, this shows that references to lack of cue familiarity were primarily reserved for *no* JOL predictions. Consequently, it seems likely that if there is a distinction between *yes* and *no* predictions, it is in whether the cue feels familiar or not.

Nevertheless, the results indicate a less clearly defined distinction between *yes* and *no* responses than some (e.g. Dunlosky et al., 2005) would predict. *Guess* predictions (which here capture low magnitude SOJs) might just be what the term suggests—instances where participants do not feel strongly predisposed toward a *yes* or a *no* prediction and rather the evidence available to them (or its lack) makes them uncertain about the future retrieval status of the items they are evaluating. If anything, this highlights the usefulness of allowing participants to express uncertainty. If one were to interpret the character of the *yes/no* sub-classification, it is the closest to the differentiation between 0 and 20% JOL.

Lastly, some phrases were almost equally likely for all *no* predictions. Namely "I do not", "do not remember", "cannot remember" and "not remember seeing". This indicates that participants were less clear on how to differentiate the three *no* response types from each other and were inclined towards using similar responses across all three confidence levels associated with *no* predictions. Together with the results from Experiment 1, these results suggest that if there is an underlying *yes/no* sub-classification in the JOL confidence scale, it is likely located at the low-end of a numeric scale, with most of the scale above this point consistent with use of *yes* predictions. This is consistent with framing effects which suggest that participants primarily accrue evidence toward a *yes* prediction as indicated by their judgments being swayed by whether the question is phrased in terms of forgetting or remembering (Finn, 2008; Koriat, Bjork, Sheffer, & Bar, 2004; Serra & England, 2012).

Support vector machine (SVM) analysis

Our final analysis was to evaluate the extent to which the written justifications for any two JOL response types were quantifiably distinct. Within each experiment, we trained SVM classifiers to compare each JOL category against all other JOL categories. If two JOL categories were justified by referring to different types of evidence, then classification accuracy for distinguishing the two categories would be good. The results are reported in Table 5, which presents overall SVM classifier performance for all JOL categories expressed as percentage of examples classified correctly.[3]

<Table 5>

Examining all adjacent JOL confidence levels, Experiment 1 revealed that the 0% and 20% JOLs were classified with above chance accuracy (50%; $X^2(1) = 20.04$, $p < .001$) and with the highest degree of accuracy of all adjacent levels. Indeed, this performance was significantly higher than the classification performance comparing the next numeric categories, the 20% vs 40% JOL comparison ($X^2(1) = 9.38$, $p = .003$). This would agree with the proposal of the descriptive model of JOL confidence that if there is a divergence in processes relied on in making the judgments, it is located between the lowest two points on the scale (Metcalfe & Finn, 2008b). All other JOL confidence levels would appear to be graded variations of a similar process (the highest classification accuracy between all the other adjacent responses of 60.32% was not significantly different from chance performance of 50%; $X^2(1) = 2.31$, $p = .129$).

In Experiment 2, the highest adjacent classification accuracy was between *yes-maybe and yes-sure* predictions, which was significantly higher than the classification

[3] We also used SVMs to compare justifications in the first and second block of each experiment. If the SVM classifier performed significantly well in classifying responses as either belonging to block 1 or block 2 of an experiment, this could suggest the justifications were substantially different across the two blocks. It is more than likely that there were some differences between the two blocks as participants completing the second block had experience with the entire experimental task. Using SVMs allowed us to quantify this difference. We compared the classifier performance to chance (50%) and found that in Experiment 1 classification accuracy (56.7%) was not above chance, $X^2(1) = 3.41$, $p = .065$. In Experiment 2 classification accuracy for comparing justifications between the two blocks was similarly low (57.3%) although this time the comparison was statistically above chance, $X^2(1) = 4.62$, $p = .032$. Given the relatively low classification accuracy however, we do not believe this to be a problem, especially as all text analyses collapsed data across blocks.

accuracy between the *yes* and *no* prediction boundary (i.e. the *guess* responses); $X^2(1) = 11.87$, $p < .001$. This is consistent with the *n*-gram results which showed there were few distinct features (*bi*-grams and *tri*-grams) characterizing the *guess* responses but contrary to the prediction that *yes* vs. *no* predictions should be distinct and so highly classifiable (e.g. Dunlosky et al., 2005).

In contrast to Experiment 1, in Experiment 2 classification accuracy for all adjacent JOL responses was above chance even when relatively low (e.g. 62.4%, $X^2(1) = 4.16$, $p < .05$). Markedly, the highest classification accuracies were at the terminal ends of the scale contrasting high confidence response (*yes* and *no*) against their adjacent medium confidence responses. This is also in contrast to Experiment 1 where we observed high classification accuracy only at the low end of numeric JOL scale. It is clear the response profiles between the JOL scales are different. Whereas the numeric confidence scale does not lead to clearly defined adjacent response boundaries (except for at the lowest end of the scale), the binary *yes/no* sub-classification scale (Experiment 2) leads to more clearly defined categories of responses. If participants treated most (if not all) of the numeric JOL confidence scale in Experiment 1 as accumulation of evidence toward a *yes* prediction then it follows that the JOL confidence levels were more clearly defined when there were fewer options provided for a positive prediction as was the case in Experiment 2. This is in line with other research (e.g. Finn, 2008; Koriat, Bjork, Sheffer, & Bar, 2004) which has shown that participants need to be asked to predict their own forgetting to treat the confidence scale as also expressing the degree to which they might forget (i.e. a *no* prediction) as compared to only the degree to which they might remember (or what we would classify as a *yes* prediction).

Indeed, the *no* responses of Experiment 2 were less clearly demarcated (as compared to the *yes* predictions). The average classification accuracy between all the *no* responses (68%) was significantly lower than the average classification accuracy between all the *yes* responses (78%), $X^2(1) = 13.35$, $p < .001$). As we saw from the *n*-gram analysis, there was a great deal of overlap between the *n*-grams participants used as a way of classifying their *no* predictions. Overall, it seems that in a paradigm where participants aim to predict their remembering, they struggle to differentiate between different levels of forgetting. This is again consistent with the idea that participants would primarily focus on the familiarity of the cue as a way of rejecting future target memory. Cue familiarity is a less varied type of signal than the more heterogeneous nature of different levels and types of target access that would be thought to characterize the unique *yes* JOL predictions.

Most relevant in regards to the current study, the classification pattern for the two response formats is clearly different. This suggests that while there is a distinction in the types of processes driving JOL confidence responses, it might be troublesome trying to assign them a discrete *no* vs *yes* prediction status. Rather, the two response formats might encourage related but nevertheless different modes of evaluation.

Experiment 3

In Experiment 3 we examined how the findings of Experiment 1 and 2 compared to a JOL task where participants predicted recall rather than recognition. Delayed JOL tasks commonly employ a cued-recall rather than a cued-recognition task and as such the question of the generalizability of Experiment 1 and 2 findings to recall JOL predictions is particularly pertinent. Further, while memory research has long established that performance on these two memory tasks can substantially differ

(MacDougall, 1904) with recognition performance generally superior to recall performance (although see Tulving & Thomson, 1973), the metacognitive literature has not truly examined the extent to which participants are sensitive to these differences when making their metacognitive judgments.

The key question of Experiment 3 was whether the patterns of metacognitive justifications for the two JOL response formats established using the recognition JOL task of Experiments 1 and 2 would persist during recall JOL predictions. Experiment 3 thus allowed us to examine the generalizability of our findings to other contexts. In contrast to Experiments 1 and 2, Experiment 3 consisted of two groups; the numeric confidence JOL group using the same JOL response method as in Experiment 1 and the binary JOL group using the same JOL response method as Experiment 2. We expected that the pattern of results in Experiment 3 would be in line with those of Experiments 1 and 2 given the only key difference in method was that participants were predicting recall rather than recognition.

Method

Participants

All 64 participants were native English speakers and randomly assigned to one of two experimental conditions. Six participants were excluded from data analysis because they did not follow the instructions when writing their justifications (e.g. referring to multiple items rather than only to the JOL given on the last trial. This left 58 participants (19 men, mean age = 21.3, *SD* = 2.7). Of those, 28 were in the numeric confidence JOL condition and 30 in the binary JOL condition.[4] The participants were all students at the University of St Andrews and received £5 as reimbursement for taking part in the study. The study was granted ethical approval by the University Teaching and Research Ethics Committee at the University of St Andrews.

Stimuli and Procedure

The stimuli were the same as those used in Experiment 1 and 2. The procedure was also mostly identical to that used in the previous experiments. Half of the participants gave their JOLs as numeric confidence (as in Experiment 1), and the other half gave their JOLs as binary (*yes/no*) predictions followed by verbal confidence (as in Experiment 2). The key difference in Experiment 3 was that participants predicted and were tested on memory recall as compared to recognition. We also increased the maximum number of justifications a participant could provide given that participants in Experiment 1 and 2 ended up providing significantly fewer justifications than the set maximum number. In Experiment 3 participants, on average, provided 19.8 justifications in the numeric confidence JOL group and 19.9 justifications in the binary JOL group.

Results

Memory and JOL performance

Participants correctly recalled 37.7% (*SD* = 21.4) of items in the numeric confidence JOL condition and 36.5% (*SD* = 21.2) of targets in the binary JOL condition.

[4] In Experiment 1 and 2 participants on average provided fewer justifications than the allowed maximum, leading to a need for larger sample sizes. We compensated for this in Experiment 3 by increasing the cap on the maximum justifications any one given participant could provide, thus allowing for faster data collection with fewer participants.

Memory performance did not differ between the two groups, $t(56) = 0.22$, $p = .828$, $d = 0.06$.

To verify that JOL predictions were accurate we compared JOLs for items that were recalled with items that were not recalled. In the numeric confidence JOL group, participants gave higher confidence JOLs (*M* = 84.75%, *SD* = 13.44) when they subsequently recalled the target as compared to when they did not (*M* = 21.58%, *SD* = 14.75). Similarly, in the binary JOL group participants gave a higher percentage of *yes* JOL predictions to items that were subsequently recalled (*M* = 88.54%, *SD* = 21.52) as compared to items that were not recalled (*M* = 24.76%, *SD* = 19.84). A group (numeric confidence JOL, binary JOL) x target recall (recalled, unrecalled) ANOVA confirmed that while JOLs were higher for recalled as compared to unrecalled targets, $F(1, 56) = 557.86$, $p < .001$, $\eta_p^2 = .91$, this did not differ between groups, $F(1, 56) = 0.82$, $p = .368$, $\eta_p^2 = .02$, and there was no interaction between the two factors, $F(1, 56) = 0.01$, $p = .911$, $\eta_p^2 < .001$.

See Figure 2 for the mean proportion of trials each JOL category was used, and Table 6 for the number of written justifications collected per JOL category. Figure 2 clearly shows that when making recall predictions in Experiment 3, participants were more likely to use JOL responses at the extreme ends of the scales (i.e. 0% and 100% JOLs and *no-sure* and *yes-sure* JOLs) than all other responses. This was not the case in Experiments 1 and 2 (where participants predicted recognition), where the JOL responses were more evenly distributed across all available options.

<Table 6>

Latent Semantic Analysis (LSA)

See Table 7 for results of LSA applied to Experiment 3 justifications.

<Table 7>

Starting with the confidence JOL responses, as in Experiment 1 participants referred to the cue more than the target in the 0% and 20% JOL justifications. In Experiment 3, the participant responses also exhibited this pattern in the 40% JOL justifications, which was not the case in Experiment 1. Note that in Experiment 1 participants also referenced the target more than the cue in justifications of 100% JOL responses, which was not the case for the confidence JOL group in Experiment 3.

In contrast, turning to the binary JOL group, there were substantial differences between LSA results for this group and LSA results of Experiment 2. In Experiment 2 participants referenced the cue more than the target in justifications of *no-guess* and *yes-guess* JOL responses. In Experiment 3 we observed this pattern for the *no-sure* and *no-maybe* justifications. As such the justifications of the binary JOL group in Experiment 3 matched quite closely those of the confidence JOL group of Experiment 3 and Experiment 1 justifications in the semantic references made to the cue term.

Word-frequency analysis

See Table 8 for *n-gram* results for the confidence JOL group and Table 9 for *n-gram* results of the binary JOL group.

<Table 8>

As in Experiment 1, the *n-gram* results for the confidence JOL group showed that for 0% JOLs participants were more likely to indicate they do "not remember seeing [the cue]" whereas for the 20% JOLs they were more likely to write "I remember seeing [the cue]". That they were referring to the cue is supported by the LSA results which showed that participants were more likely to semantically reference the cue than the target in their justifications of 0% and 20% JOLs. There is thus again an indication that the demarcation of these two responses was primarily in cue familiarity. All other levels were less clear but seemed to reference possible target access (e.g. "think I remember", "remember paired word") and were primarily differentiated by words indicating certainty (e.g. "vague" for 40% vs. "clearly" for 100% JOL justifications).

<Table 9>

As in Experiment 2, the *no-sure* responses in the binary JOL group echoed the 0% JOL responses ("not remember seeing [the cue]") while the *yes-sure* justifications were similar to the 100% justifications (using words such as "clearly", "associated", "imagined"). However, the reference to lack of cue familiarity ("not remember seeing") was almost exclusive to the *no-sure* responses in the binary JOL group in Experiment 3, as compared to being shared by all *no* JOL justifications, as was the case in Experiment 2. The *no-maybe* and *no-guess* responses focused on inaccessibility of the target ("but I cannot"…"remember word"). The *yes* responses, in contrast, referenced partial accessibility of the target such as a general awareness of what it was about and that it might be possible to access it.

Support vector machine (SVM) analysis

See Table 5 for results of the SVM analyses carried out on the confidence JOL and binary JOL group responses.[5] Overall, the pattern of classification performance was lower than that observed in Experiments 1 and 2 and there were no categories of JOL justification responses that were classified with above 90% accuracy. Despite that, the data from Experiment 3 showed the same general pattern of results as observed in Experiments 1 and 2. The pattern of classifications between the two JOL response formats differed while the confidence JOL group's justifications mapped onto those of Experiment 1 and the justifications of the binary JOL group mapped onto those of Experiment 2.

The SVM classification results for the confidence JOL group were very similar to those of Experiment 1 which also employed the numerical confidence JOL scale for participants' predictions. The 0% JOL justifications were most clearly demarcated in contrast to all other JOL justifications and the 0% vs. 20% JOL classification was the highest from among all adjacent level JOL justification classifications. The 77.06% classification accuracy was above chance (50%), $X^2(1) = 16.15$, $p < .001$. As in Experiment 1, the next highest classification accuracy for adjacent JOL levels (comparing 80% and 100% JOL justifications) was not above chance, $X^2(1) = 3.06$, $p = .080$.

[5] As in previous experiments, we also trained an SVM classifier to compare justifications written for block 1 and block 2 of the experiment. We carried out this comparison for each experimental group separately. This allowed us to evaluate whether the justifications written between the two blocks were quantitatively different. This would be reflected in significantly above chance (50%) classifier performance. In the numeric confidence JOL group, the SVM classification accuracy (56.48%) for justifications written in the two blocks was not significantly above chance, $X^2(1) = 2.48$, $p = .115$. The same was true for classification accuracy of the binary JOL group responses (52.5%), $X^2(1) = 0.26$, $p = .612$

In contrast, the binary JOL justifications were mostly defined for *yes-sure* as compared to all other JOL categories, with the *yes-sure* vs. *yes-maybe* classification accuracy being the highest of all comparisons made for JOLs adjacent to each other (again corresponding to Experiment 2 results). While this classification performance was also numerically higher than that between *yes-guess* and *no-guess* responses, this time the comparison of classification accuracy between *yes-sure* vs. *yes-maybe* and *yes-guess* vs. *no-guess* responses was not statistically significantly different, $X^2(1) = 3.23$, $p = .072$. However, the classification accuracy for comparisons of *yes-guess* and *no-guess* responses was not above chance (50%), $X^2(1) = 0.47$, $p = .49$. Lastly, as in Experiment 2, justifications for all the *yes* JOLs were overall more clearly defined between each other than justifications for the *no* JOLs, as indicated by a higher overall classification accuracy (69.93% vs. 57.20%), $X^2(1) = 9.76$, $p = .002$.

Summary

The only key methodological difference between Experiment 3 and Experiments 1 and 2 was that when making a JOL, participants predicted future recall rather than recognition performance. Other than this change in the JOL task, procedures within the confidence and binary JOL groups in Experiment 3 were designed to replicate the procedures in Experiments 1 and 2 respectively. We were primarily interested in observing whether the pattern of results of Experiments 1 and 2, speaking to the lack of direct correspondence between the two JOL response formats, would generalize from a JOL task using recognition to a JOL task using cued recall.

The pattern of results from the confidence JOL group matched that observed in Experiment 1. There were some minor exceptions, the most notable being within the LSA results, which showed that participants in the confidence JOL group in Experiment

3 were more likely to reference the cue as compared to the target when expressing 40% JOL confidence, which was not the case in Experiment 1. Other than this slight difference, which is not in conflict with the descriptive model, it appears that participants predicting recall performance used the confidence scale in a comparable manner to participants predicting recognition.

There were more differences observed between the binary JOL group here and in Experiment 2. This was especially true for the LSA results. In Experiment 2 participants were more likely to reference the cue (as compared to the target) in justifications of *yes-guess* and *no-guess* JOLs. In the binary JOL group in Experiment 3 we observed this pattern of results for *no-sure* and *no-maybe* responses, aligning these binary results with those of the confidence JOL responses.

Nevertheless, the remainder of the results from this final experiment confirmed that the two JOL response formats were not used analogously, establishing that for both recognition and recall JOL predictions, binary JOLs did not directly map onto numeric confidence JOL responses. For example, the SVM results showed repeatedly that for confidence results the clearest demarcation was between 0% JOL confidence and all other responses whereas for binary responses it was *yes*-sure JOL (what in confidence terms could be interpreted as 100% JOL confidence) that seems to be most distinct from all other justifications, as reflected in high classification accuracy. Confidence judgments have a clear distinction whereby 0% represents lack of cue familiarity whilst most of the scale represents degrees of target access, such that the high confidence responses are in fact relatively similar. Binary JOLs on the other hand seem to have more clearly demarcated *yes* JOL categories than *no* JOL categories, with clearer differences between different levels of confidence assigned to these *yes* JOL predictions (along the lines of *sure-maybe-guess* used in this series of experiments).

General Discussion

Within any metacognitive paradigm, aspects of the task are manipulated so that specific information is made salient to participants; in metamemory tasks this is usually through encoding or retrieval instructions. The question that arises is whether the information that is shown to influence metacognitive judgments in such paradigms remains relevant in other contexts (see for example Hertzog et al., 2014). This study asked: what information do participants consider relevant to their JOLs in the absence of any such manipulation and how does this information map onto the interpretative and descriptive models of delayed JOLs? More specifically, we investigated written justifications for numeric confidence and binary (*yes/no*) JOL predictions. Participants completed a standard JOL task and on some trials were asked to justify their predictions, which were subsequently analysed using a range of natural language processing techniques. The results showed that (i) participants could justify their metacognitive judgments, (ii) numeric confidence JOL justifications mapped broadly onto the descriptive two-stage model (Metcalfe & Finn, 2008b) as they referenced both cue and target related information, (iii) numeric confidence JOLs had different characteristics to binary *yes/no* JOLs, thus challenging the interpretative model assuming that numeric confidence JOLs can consistently be sub-classified into equal proportion of *yes/no* judgments.

Overall, participants could justify their JOLs and did so with reference to both cue- and target-related information as well as with reference to associations they made between them. This was the case even though we did not manipulate these factors or instruct participants in any way as to how they should learn the items and what information they should focus on when making their JOLs. The results thus complement studies which have shown that emphasis on cue, target and associative information shifts metacognitive confidence (Benjamin, 2005; Hertzog et al., 2014; Metcalfe & Finn, 2008b) and support the heuristics view of metacognitive judgments as based on evidence accumulation processes (Brewer, Marsh, Clark-Foos, & Meeks, 2010; Koriat, 2000).

The results of numeric confidence JOLs were consistent with the predictions of Metcalfe and Finn (2008b). The 0% and 20% JOL responses were the most divergent of any adjacent JOL confidence levels as indicated by highest classification accuracy. The content analyses supported the idea that, whereas the 0% JOLs corresponded to a lack of cue familiarity, the 20% JOLs were given to items whose cue was familiar but whose target was not accessible. All other JOL confidence levels reflected an increase in target accessibility. We therefore provide support for a descriptive model of JOL confidence as resulting from a two-stage evaluation, with interrogation of different evidence characterizing each stage.

The results of Experiment 2 suggested that participants referred to the cue to distinguish between a *no* and some degree of a *yes* response as well as to characterize high confidence *no* responses from all other responses. This would map onto the differences between 0% and 20%, suggesting that if there is an underlying *yes/no* sub-classification in the numeric JOL confidence scale, it is a differentiation of the lowest confidence responses only. Consistent with this, there was also an indication that participants struggled to differentiate the three degrees of *no* confidence prediction from each other, at least when framed in terms of remembering. The degrees of *yes* prediction were more clearly demarcated. However, the overarching distinction between *yes* and *no* predictions was less clear-cut than predicted by the interpretative model and it remains questionable whether binary *yes/no* sub-classification reflects how participants approach the JOL confidence scale.

The overall pattern of results was confirmed in Experiment 3 where participants predicted future recall rather than recognition when making their JOLs. There were also some minor differences between Experiment 2 and the binary JOL group in Experiment 3, especially in the LSA results where the binary JOL group resembled, at least in some aspects, results of the confidence JOL group. It is very likely that the exact interpretations of the confidence scale and the use of the two JOL response formats will differ somewhat between different tasks and even between two studies using the same task. This was particularly clear in Figure 2 which showed the average proportion of trials on which each JOL response (e.g. 0% or *yes-maybe*) was used. In Experiment 3, when making recall JOL predictions, both numeric confidence and binary JOL responses favoured the extreme ends of the scale (i.e. 0% and 100% JOL and *no-sure* and *yes-sure* JOLs). This was not the case in Experiments 1 and 2 where participants predicted recognition and, on average, used the JOL responses available to them more evenly. Altogether, results of the experiments presented here suggest that the use of these response formats is unlikely to be completely fixed and will depend on experimental design and context.

It is clear that analogous points on numeric confidence and binary sub-classification 6-point scales are not always equivalent—we cannot treat the numeric JOL confidence scale as evenly corresponding to the sub-classifications within *yes* and *no* responses. It is possible that this might be true in some cases but it first needs to be established, it cannot be assumed. The experiments presented here showed that across the two scales, the terminal points corresponded, i.e. a 0% JOL was equivalent to a high confidence *no* and a 100% JOL was equivalent to a high confidence *yes*. It is unsurprising that our understanding of the extremes of the scale might be correct. However, these extremes differed in how they related to the mid-range responses and this is where we observed the most differences. The overall different pattern of JOL justifications across the two response formats highlights that participants do not use all points of the two scales in the same way. For example, asking participants to first give a *yes/no* judgment followed by a confidence assignment, seems to have led to more clearly demarcated categories of responses than was the case with the numeric confidence scale. This contradicts the idea that participants interpret the numeric confidence scale in terms of a binary *yes/no* sub-classification.

This lack of equivalence is worth highlighting, especially as there is an underlying assumption in much metacognitive research that confidence judgments are probabilistic. It is common, for example, to interpret 0%, 20% and 40% as *no* predictions. This is seen particularly in assessments of metacognitive accuracy in terms of calibration; an assessment of whether metacognitive judgments correspond exactly to performance (perfect calibration would be if items given 60% JOLs were recognized at a rate of 60% in subsequent memory tests etc.). Considerable research has focused on what drives poor calibration which is observed across domains (see for example Finn & Metcalfe, 2007; Gigerenzer, Hoffrage, & Kleinbolting, 1991; Koriat, Lichtenstein, & Fischhoff, 1980; Koriat, Ma'ayan, Sheffer, & Bjork, 2006; Kornell & Bjork, 2009). However, recently Hanczakowski et al. (2013; see also Zawadzka & Higham, 2015) showed that the common observation that participants tend to display underconfidence in terms of calibration (i.e. lower average confidence JOL than overall memory performance) was not observed with a *yes/no* response format and when the proportion of *yes* responses was used to assess calibration. Hanczakowski et al. interpreted this as indicating that participants were not truly underconfident as has been previously assumed and that the results could rather be explained as driven by misunderstanding of how participants treat the JOL confidence scale. This was observed using an immediate JOL task where predictions are made during study rather than after all items have been learned as is the case with delayed JOLs employed here. Nevertheless, the finding is consistent with the suggestion from the current study that participants are treating most of the JOL confidence scale as a *yes* prediction. In most likelihood, the anchoring of a confidence scale shifts between participants and across tasks.

This further relates to findings that question format influences how participants respond in both metacognitive (Finn, 2008; Serra & England, 2012) and recognition memory tasks (Mill & O'Connor, 2014). For example, participants anchored their JOLs lower on the JOL confidence scale when judging future remembering as compared to forgetting (Serra & England, 2012). Similarly, recognition judgments for whether an item has been studied or is seen for the first time have been shown to be influenced by whether the question is termed in terms of judging 'oldness' or 'novelty' (Mill & O'Connor, 2014). More specifically, participants shifted their response bias to more likely disconfirm the question asked (more likely to respond 'new' when asked 'old?'). This study adds to a newly growing literature demonstrating that, in addition to question format, response format also influences participant responding in metacognitive tasks (Jersakova et al., 2016; Overgaard & Sandberg, 2012). Taken together, the evidence indicates that the methods used to assess cognitive and metacognitive phenomena are of theoretical importance, with direct consequences for the inferences we draw from our data.

Lastly, we acknowledge that it is possible that asking participants to justify their responses might alter the nature of the JOL task. Indeed, this is a problem present throughout metacognition studies, which often require participants to make explicit reports of the processes under investigation (e.g. by responding to questions such as 'Can you remember the first letter of an unrecalled word?'), thus always leaving the question open whether the same results would be obtained if participants were not asked to reflect on their retrieval experience. The strength of the experiments presented here is that the findings are in line with existing literature. The current study was possible because it built on extensive published behavioral literature and was able to confirm, with new means, existing conclusions made with more traditional methods and data (e.g. Metcalfe & Finn, 2008b). We strongly believe that in this way subjective reports can be used to complement and develop existing findings. Given the inherently subjective nature of the processes under investigation in the metacognitive literature, it is clear that there are invaluable insights we can gain from probing for additional, subjective information from participants. If such approaches should go hand-in-hand with other methods of experimental design and data analysis, the field as a whole should benefit as a consequence.

For example, future work could investigate how the present results would compare to the immediate JOL paradigm, in which participants render judgment immediately after study, with both the cue and the target present. Behaviourally, immediate JOLs have been demonstrated to differ from delayed JOLs and to rely on different types of evidence (Koriat & Bjork, 2006; Rhodes & Tauber, 2011). Whereas delayed JOLs require an evaluation of access to information in long-term memory, immediate JOLs rely primarily on information held in short-term memory.

Consequently, one would expect to observe different patterns of responses and distinct content in justifications for immediate as compared to delayed JOLs; for example, participants do not need to attempt to retrieve the target which is present in immediate JOLs and they might instead focus on the level of association between the cue and the target (Koriat & Bjork, 2005). Based on the current findings, our prediction is that participants should be able to produce justifications of their immediate JOLs and these would reference the relationship between the cue and the target. A similar paradigm employed in other types of metacognitive tasks is likely to confirm that metacognition can collectively be considered an evidence aggregation and evaluation process. Conversely, it would be of interest to investigate which types of influences participants might not be aware of through failing to account for them in their justifications.

In summary, we provide evidence for metacognitive confidence judgements as resulting from evaluative processes that weigh the degree of evidence toward the decision framed by the response-eliciting question (in this case, 'will this item be remembered?'). The present study demonstrates that participants have at least a degree of access into this process and can justify the JOLs they are making. What is more, they do so with reference to processes observed to influence JOL magnitude in the literature. Importantly, the results demonstrate that widely used numeric confidence JOLs are unlikely to have an underlying *yes/no* direct mapping. At the very least, this distinction is unlikely to be couched in probabilistic terms (e.g. 40% interpreted as a rejection of future retrieval). This finding should guide future assessment and interpretation of metacognitive confidence judgments.

Acknowledgements

This research was supported by the Economic and Social Research Council Studentship awarded to Radka Jersakova [ES/J500215/1]. We want to thank Ellina Knudsen and Katherine Habel for their help with data collection.

References

Ackerman, R., & Thompson, V. (2014). Meta-Reasoning: What Can We Learn from Meta-Memory? In A. Feeney & V. Thompson (Eds.), *Reasoning as Memory*. Hove, UK: Psychology Press.

Alban, M. W., & Kelley, C. M. (2013). Embodiment meets metamemory: weight as a cue for metacognitive judgments. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(5), 1628–34. http://doi.org/10.1037/a0032420

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459.

Benjamin, A. S. (2005). Response speeding mediates the contributions of cue familiarity and target retrievability to metamnemonic judgments. *Psychonomic Bulletin & Review*, *12*(5), 874–879.

Brewer, G. A., Marsh, R. L., Clark-Foos, A., & Meeks, J. T. (2010). Noncriterial recollection influences metacognitive monitoring and control processes. *Quarterly Journal of Experimental Psychology (2006)*, *63*(10), 1936–42. http://doi.org/10.1080/17470210903551638

Castel, A. D., McCabe, D. P., & Roediger, H. L. (2007). Illusions of competence and overestimation of associative memory for identical items: evidence from judgments of learning. *Psychonomic Bulletin & Review*, *14*(1), 107–11. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17546739

Dennis, S. (2006). 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). London: Routledge.

Dougherty, M. R., Scheck, P., Nelson, T. O., & Narens, L. (2005). Using the past to predict the future. *Memory & Cognition*, *33*(6), 1096–1115.

Dunlosky, J., Serra, M. J., Matvey, G., & Rawson, K. A. (2005). Second-Order Judgments About Judgments of Learning. *The Journal of General Psychology*, *132*(4), 335–346. http://doi.org/10.3200/GENP.132.4.335-346

Finn, B. (2008). Framing effects on metacognitive monitoring and control. *Memory & Cognition*, *36*(4), 813–821. http://doi.org/10.3758/MC.36.4.813

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), 238–244. http://doi.org/10.1037/0278-7393.33.1.238

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: computation, biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1280–6. http://doi.org/10.1098/rstb.2012.0021

Fleming, S. M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., & Lau, H. (2015). Action-Specific Disruption of Perceptual Confidence. *Psychological Science*, *26*(1), 89–98. http://doi.org/10.1177/0956797614557697

Fletcher, L., & Carruthers, P. (2012). Metacognition and reasoning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 1366–1378. http://doi.org/10.1098/rstb.2011.0413

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

Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic Mental Models : A Brunswikian Theory of Confidence. *Psychological Review*, *98*(4), 506–528.

Hamel, L. H. (2009). *Knowledge discovery with support vector machine*. Hoboken, NJ: Wiley.

Hertzog, C., Fulton, E. K., Sinclair, S. M., & Dunlosky, J. (2014). Recalled aspects of original encoding strategies influence episodic feelings of knowing. *Memory & Cognition*, *42*, 126–40. http://doi.org/10.3758/s13421-013-0348-z

Jersakova, R., Moulin, C. J. A., & O'Connor, A. R. (2016). Investigating the role of assessment method on reports of déjà vu and tip-of-the-tongue states during standard recognition tests. *PloS ONE*, *11*(4), e0154334.

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 1322–1337. http://doi.org/10.1098/rstb.2012.0037

Koriat, A. (2000). The feeling of knowing: some metatheoretical implications for consciousness and control. *Consciousness and Cognition*, *9*, 149–171. http://doi.org/10.1006/ccog.2000.0433

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*(2), 187–94. http://doi.org/10.1037/0278-7393.31.2.187

Koriat, A., & Bjork, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory & Cognition*, *34*(5), 959–72. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17128596

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: the role of experience-based and theory-based processes. *Journal of Experimental Psychology. General*, *133*(4), 643–56. http://doi.org/10.1037/0096-3445.133.4.643

Koriat, A., & Levy-Sadot, R. (2001). The combined contributions of cue familiarity and accessibility heuristics to feeling of knowing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27*(1), 34–53.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for Confidence. *Journal of Experimental Psychology : Human Learning and Memory*, *6*(2), 107–118.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The Intricate Relationships Between Monitoring and Control in Metacognition : Lessons for the Cause-and-Effect Relation Between Subjective Experience and Behavior. *Journal of Experimental*

Psychology: General, *135*(1), 36–69.

Koriat, A., Ma'ayan, H., Sheffer, L., & Bjork, R. A. (2006). Exploring a Mnemonic Debiasing Account of the Underconfidence-With-Practice Effect. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *32*(3), 595–608. http://doi.org/10.1037/0278-7393.32.3.595

Koriat, A., Sheffer, L., & 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. http://doi.org/10.1037//0096-3445.131.2.147

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General*, *138*(4), 449–468. http://doi.org/10.1037/a0017350

Liu, Y., Su, Y., Xu, G., & Chan, R. C. K. (2007). Two dissociable aspects of feeling-of-knowing: knowing that you know and knowing that you do not know. *Quarterly Journal of Experimental Psychology (2006)*, *60*(5), 672–80. http://doi.org/10.1080/17470210601184039

MacDougall, R. (1904). Recognition and Recall. *Journal of Philosophy*, *1*(9), 229–233.

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

Metcalfe, J., & Finn, B. (2008b). Familiarity and retrieval processes in delayed judgments of learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *34*(5), 1084–97. http://doi.org/10.1037/a0012580

Nelson, T., & Narens, L. (1990). Metamemory: a theoretical framework and new findings. *The Psychology of Learning and Motivation*, *26*, 125–173.

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), 267–271.

Overgaard, M., & Fazekas, P. (2016). Can No-Report Paradigms Extract True Correlates of Consciousness ? *Trends in Cognitive Sciences*, *20*(4), 241–242. http://doi.org/10.1016/j.tics.2016.01.004

Overgaard, M., & Sandberg, K. (2012). Kinds of access : different methods for report reveal different kinds of metacognitive access. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 1287–1296. http://doi.org/10.1098/rstb.2011.0425

Pedregosa, F., Varoquaux, G., Gramfort, A., Vincent, M., Bertrand, T., Grisel, O., … Duchesnay, E. (2011). Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peirce, J. W. (2007). PsychoPy--Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13. http://doi.org/10.1016/j.jneumeth.2006.11.017

Peters, M. A. K., & Lau, H. (2015). Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. *eLife*, 9651. http://doi.org/10.7554/eLife.09651

Rahnev, D., Koizumi, A., McCurdy, L. Y., D'Esposito, M., & Lau, H. (2015). Confidence Leak in Perceptual Decision Making. *Psychological Science*, *26*(11), 1664–80. http://doi.org/10.1177/0956797615595037

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: evidence for metacognitive illusions. *Journal of Experimental Psychology. General*, *137*(4), 615–25. http://doi.org/10.1037/a0013684

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: a meta-analytic review. *Psychological Bulletin*, *137*(1), 131–48. http://doi.org/10.1037/a0021705

Selmeczy, D., & Dobbins, I. G. (2014). Relating the content and confidence of recognition judgments. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *40*(1), 66–85. http://doi.org/10.1037/a0034059

Serra, M. J., & England, B. D. (2012). Magnitude and accuracy differences between judgements of remembering and forgetting. *Quarterly Journal of Experimental Psychology (2006)*, *65*(11), 2231–57. http://doi.org/10.1080/17470218.2012.685081

Tulving, E., & Thomson, D. M. (1973). Encoding Specificity and Retrieval Processes in Episodic Memory. *Psychological Review*, *80*(5), 352–373.

Urquhart, J. A., & O'Connor, A. R. (2014). The awareness of novelty for strangely familiar words: a laboratory analogue of the déjà vu experience. *PeerJ*, *2*, e666. http://doi.org/10.7717/peerj.666

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.

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*, *22*(2), 572–588. http://doi.org/10.1016/j.concog.2013.03.009

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making : confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 1310–1321. http://doi.org/10.1098/rstb.2011.0416

Yonelinas, A. P. (2002). The Nature of Recollection and Familiarity: A Review of 30 Years of Research. *Journal of Memory and Language*, *46*(3), 441–517. http://doi.org/10.1006/jmla.2002.2864

Table 1

Number of justifications collected in each JOL category by experiment

Exp1	0%	20%	40%	60%	80%	100%
	127	146	134	120	132	137
Exp2	No - Sure	No - Maybe	No - Guess	Yes - Guess	Yes - Maybe	Yes – Sure
	102	177	137	102	195	205

Table 2

Cue-justification and target-justification LSA scores by category and experiment

Experiment	JOL category	Cue LSA score	Target LSA score	*t*-value	df	*p*-value	*d*
1	0%	.21 (.14)	.17 (.12)	2.29	120	.024*	0.30
	20%	.20 (.13)	.17 (.10)	2.42	143	.017*	0.27
	40%	.20 (.11)	.19 (.11)	0.55	133	.581	0.07
	60%	.17 (.12)	.19 (.13)	1.58	119	.118	0.19
	80%	.20 (.13)	.21 (.14)	0.39	129	.700	0.05
	100%	.21 (.12)	.24 (.14)	2.09	134	.039*	0.22
2	No - Sure	.08 (.12)	.06 (.11)	1.14	98	.259	0.12
	No - Maybe	.08 (.12)	.08 (.13)	0.38	171	.704	0.03
	No - Guess	.11 (.14)	.08 (.13)	2.79	133	.006*	0.25
	Yes - Guess	.14 (.17)	.09 (.13)	2.64	101	.010*	0.31
	Yes - Maybe	.10 (.13)	.09 (.13)	1.03	192	.302	0.08
	Yes - Sure	.14 (.17)	.16 (.19)	1.37	202	.172	0.11

Note. Cue and target LSA score descriptives expressed as: mean (standard deviation). Results of paired-samples t-tests comparing the cue and target LSA scores within each JOL category are also reported. *s indicate significance at an alpha threshold of .05.

Table 3

N-gram analysis results for Experiment 1

JOL	*n*-gram	Count	Total	Proportion	*p*
0%	not remember this	8	11	.73	<.001
	remember seeing this	13	30	.43	<.001
	remember what word	6	11	.55	.004
	do not remember	39	66	.59	<.001
	seeing this word	13	31	.42	<.001
	I do not	43	79	.54	<.001
	remember this word	10	26	.38	.005
	I cannot remember	17	45	.38	<.001
	not remember seeing	25	32	.78	<.001
	cannot remember what	9	19	.47	.001
	cannot remember word	5	11	.45	.021
	do not	58	114	.51	<.001
	not remember	42	73	.58	<.001
	that word	6	16	.38	.031
	have no	7	11	.64	<.001
	this word	31	99	.31	<.001
	word at	6	10	0.6	.002
	I do	43	81	.53	<.001
	seeing this	14	33	.42	<.001
	at all	18	23	.78	<.001
	remember seeing	35	105	.33	<.001
	cannot remember	34	88	.39	<.001
	I cannot	22	78	.28	.007
20%	seeing word but	7	10	.70	<.001
	be able to	12	37	.32	.034
	do not think	8	14	.57	.001
	not think I	7	13	.54	.004
	vaguely remember seeing	8	10	.80	<.001
	but I cannot	6	14	.43	.030
	but cannot remember	7	17	.41	.025
	I am not	11	30	.37	.016
	do not really	7	10	.70	<.001
	what it was	7	17	.41	.025
	remember seeing word	14	42	.33	.026
	what it	9	23	.39	.026
	not confident	6	12	0.5	.013
	be able	12	37	.32	.034
	not really	8	11	.73	<.001
	not think	9	15	.60	<.001
	am not	11	33	.33	.034
	I cannot	24	78	.31	.008
	word so	5	11	.45	.036
	really remember	10	15	.67	<.001
	seeing word	17	48	.35	.005
	vaguely remember	11	14	.79	<.001
	might be	6	10	.60	.004

	with it	6	15	.40	.042
	I do	26	81	.32	.004
	do not	39	114	.34	<.001
	cannot remember	24	88	.27	.038
	remember seeing	34	105	.32	<.001
	but I	20	71	.28	.044
40%	think I remember	6	16	.38	.040
	word but cannot	7	11	.64	<.001
	word but I	7	13	.54	.003
	if I saw	7	14	.50	.004
	I remember seeing	17	39	.44	<.001
	think I could	6	13	.46	.013
	remember word but	5	11	.45	.026
	I think I	15	52	.29	.026
	word and	9	29	.31	.048
	word it	5	12	.42	.038
	word I	7	18	.39	.022
	second word	8	22	.36	.022
	to recognise	7	18	.39	.022
	I could	20	63	.32	.004
	but I	20	71	.28	.016
	if I	13	29	.45	<.001
	word but	25	58	.43	<.001
	I may	8	19	.42	.008
	but cannot	12	29	.41	.002
	cannot recall	8	19	.42	.008
	I think	27	85	.32	<.001
	recognise it	10	28	.36	.018
	think I	26	82	.32	<.001
	I remember	41	176	.23	.026
	remember seeing	26	105	.25	.036
60%	but I am	7	18	.39	.012
	I feel like	6	18	.33	.043
	I think I	21	52	.40	<.001
	I am not	9	30	.30	.036
	remember making	7	13	.54	.001
	could recognise	5	10	.50	.010
	I might	6	16	.38	.025
	I feel	13	35	.37	.001
	and I	9	20	.45	.001
	I am	19	71	.27	.011
	think I	27	82	.33	<.001
	feel I	6	13	.46	.008
	I can	13	45	.29	.019
	pair word	8	15	.53	<.001
	it but	6	15	.40	.017
	that I	11	38	.29	.023
	its pair	5	14	.36	.048
	between two	5	13	.38	.035
	remember pair	6	17	.35	.033

	I think	27	85	.32	<.001
	but I	19	71	.27	.011
80%	I remember word	10	28	.36	.018
	I am pretty	9	14	.64	<.001
	in my head	7	16	.44	.010
	one of	5	11	.45	.025
	am pretty	9	14	.64	<.001
	pretty sure	5	10	.50	.015
	it was	19	70	.27	.024
	I remember	43	176	.24	.008
	in my	10	31	.32	.028
	I associated	6	14	.43	.019
	my head	7	19	.37	.027
100%	I can remember	11	26	.42	<.001
	link between	8	18	.44	.007
	as I	7	19	.37	.033
	thought of	5	10	.50	.018
	it is	9	25	.36	.028
	can remember	12	31	.39	.003
	I can	14	45	.31	.027
	I made	10	27	.37	.017

Note. A count of occurrences of each *n*-gram in justifications for the corresponding JOL category are reported along with total number of occurrences, proportion of occurrence and *p*-value computed using the binomial test.

Table 4

N-gram analysis results for Experiment 2

JOL	*n*-gram	Count	Total	Proportion	*p*
No-Sure	do not remember	28	78	.36	<.001
	I cannot remember	16	78	.21	.017
	I do not	37	93	.40	<.001
	do not think	4	13	.31	.048
	that I will	6	18	.33	.011
	remember this word	11	31	.35	<.001
	cannot remember seeing	7	12	.58	<.001
	remember seeing this	9	33	.27	.008
	seeing this word	11	38	.39	.002
	not remember this	9	12	.75	<.001
	remember this word	11	38	.29	.002
	word at all	10	21	.48	<.001
	not remember seeing	12	32	.38	<.001
	do not even	11	12	.92	<.001
	not even remember	11	11	1.00	<.001
	do not	53	137	.39	<.001
	this word	32	120	.27	<.001
	not remember	33	89	.37	<.001
	even remember	11	11	1	<.001
	not recognise	5	11	.45	.004
	I do	37	99	.37	<.001
	have no	5	12	.42	.007
	not think	5	15	.33	.020
	no idea	6	11	.55	<.001
	word at	10	25	.4	<.001
	not even	11	13	.85	<.001
	at all	16	34	.47	<.001
	cannot remember	24	144	.17	.045
	remember seeing	24	90	.27	<.001
	seeing this	11	39	.28	.002
	seeing word	13	48	.27	.002
	was paired	6	22	.27	.029
	not know	4	12	.33	.036
	first word	6	21	.29	.023
No-Maybe	able to recognise	6	13	.46	.025
	might be able	8	15	.53	.003
	not sure if	5	11	.45	.044
	be able to	23	66	.35	.003
	I cannot remember	23	78	.29	.030
	that I would	5	10	.50	.028
	if I saw	6	13	.46	.025
	I might be	7	15	.47	.015
	may be able	7	11	.64	.002
	I would remember	7	10	.70	<.001
	cannot remember what	8	19	.42	.019
	cannot remember	41	144	.28	.008

	of two	6	14	.43	.037
	now but	5	10	.50	.028
	be able	23	66	.35	.003
	would remember	7	10	.70	<.001
	recognise it	13	37	.35	.021
	to pair	5	11	.45	.044
	what word	9	23	.39	.029
	for this	6	14	.43	.037
	I feel	10	24	.42	.016
	sure if	5	11	.45	.044
	to me	8	12	.67	<.001
	if I	29	49	.59	<.001
	it if	7	13	.54	.006
	to mind	6	14	.43	.037
	I may	13	21	.62	<.001
	would recognise	8	21	.38	.047
	I would	27	74	.36	<.001
	I might	14	41	.34	.026
	may be	8	18	.44	.014
	able to	23	67	.34	.005
	it but	10	22	.45	.005
	remember what	12	34	.35	.027
	might be	10	27	.37	.027
	but I	26	92	.28	.034
No-Guess	to guess	6	16	.38	.023
	word so	9	24	.38	.006
	do not remember	25	78	.32	<.001
	not remember seeing	10	32	.31	.021
	do not recall	5	13	.38	.033
	cannot remember word	6	18	.33	.041
	I did not	6	17	.35	.031
	I do not	27	93	.29	<.001
	seeing word	15	48	.31	.004
	be guess	9	14	.64	<.001
	cannot remember	41	144	.28	<.001
	not remember	29	89	.33	<.001
	I do	28	99	.28	<.001
	at all	11	34	.32	.012
	have to	4	10	.40	.049
	I did	6	18	.33	.041
Yes-Guess	think I would	10	21	.48	<.001
	I think I	16	47	.34	<.001
	but I cannot	9	25	.36	<.001
	I am sure	5	14	.36	.014
	I recall	5	10	.50	.002
	but cannot	12	36	.33	<.001
	think I	21	77	.27	<.001
	but I	23	92	.25	<.001
	I could	9	35	.26	.012
	I would	17	74	.23	.004

	again I	4	11	.36	.026
	I think	25	88	.28	<.001
	it when	5	13	.38	.010
	I remember	24	139	.17	.030
	word but	16	72	.22	.007
	cannot remember	24	144	.17	.045
	am not	13	47	.28	.001
	but I	23	92	.25	<.001
Yes-Maybe	think I remember	7	13	.54	.010
	I think it	7	13	.54	.010
	when I see	14	23	.61	<.001
	think I will	6	13	.46	.040
	not hundred percent	10	15	.67	<.001
	I see it	14	17	.82	<.001
	presented with it	6	10	.60	.009
	hundred percent sure	9	13	.69	<.001
	to do with	11	19	.58	<.001
	but not sure	7	15	.47	.025
	word but not	9	18	.50	.006
	something to do	11	16	.69	<.001
	I think I	17	47	.36	.019
	word and	15	37	.41	.008
	I know	9	18	.50	.006
	see it	14	18	.78	<.001
	tried to	8	10	.80	<.001
	percent sure	9	13	.69	<.001
	I see	18	34	.53	<.001
	when I	15	35	.43	.006
	think I	28	77	.36	.002
	and think	9	10	.90	<.001
	it when	6	13	.46	.040
	word when	7	13	.54	.010
	something to	11	17	.65	<.001
	word but	23	72	.32	.031
	but not	20	40	.50	<.001
	to do	11	19	.58	<.001
	am not	16	47	.34	.047
	it I	5	10	.50	.042
	not hundred	10	15	.67	<.001
	I think	35	88	.40	<.001
	hundred percent	12	18	.67	<.001
	do with	11	19	.58	<.001
	will recognise	7	14	.50	.016
	think it	8	15	.53	.006
	with it	13	32	.41	.015
Yes-Sure	I remember word	10	24	.42	.045
	in my head	13	20	.65	<.001
	because I	14	31	.45	.004
	my head	13	27	.48	.004
	two words	15	38	.39	.018

I remembered	14	20	.70	<.001
remember that	8	16	.50	.014
I remember	43	139	.31	.019
I imagined	6	10	.60	.011
association between	7	13	.54	.013
thought of	8	13	.62	.003
in my	23	55	.42	.002
can remember	16	40	.40	.012
I can	24	50	.48	<.001
I made	11	24	.46	.011

Note. A count of occurrences of each *n*-gram in justifications for the corresponding JOL category are reported along with total number of occurrences, proportion of occurrence and *p*-value computed using the binomial test.

Table 5

Bivariate SVM classification accuracy results by experiment and JOL response.

Experiment 1 (Recognition): Numeric confidence JOL

	20%	40%	60%	80%	100%
0%	75.91	81.68	86.29	93.08	94.74
20%		57.86	73.69	79.86	84.51
40%			59.06	69.92	80.15
60%				60.32	68.99
80%					53.33

Experiment 3 (Recall): Numeric confidence JOL

	20%	40%	60%	80%	100%
0%	77.06	77.42	79.17	86.96	84.17
20%		54.65	55.06	65.88	69.03
40%			50.68	65.22	71.13
60%				50.0	65.0
80%					63.54

Experiment 2 (Recognition): Binary JOL

	No-Maybe	No-Guess	Yes-Guess	Yes-Maybe	Yes-Sure
No-Sure	76.43	64.17	85.29	91.28	92.21
No-Maybe		62.66	65.71	71.12	91.15
No-Guess			67.5	80.24	90.11
Yes-Guess				62.4	86.36
Yes-Maybe					84.57

Experiment 3 (Recall): Binary JOL

	No-Maybe	No-Guess	Yes-Guess	Yes-Maybe	Yes-Sure
No-Sure	62.62	60.44	69.13	73.73	89.52
No-Maybe		47.87	66.67	61.76	82.68
No-Guess			57.35	60.47	87.39
Yes-Guess				63.16	73.24
Yes-Maybe					71.43

50-60
60-70
70-80
80-90
90-100

Note. The results express percentage of test cases classified accurately and reflect the degree to which two JOL responses could be said to differ in how they were justified.

Table 6

Number of justifications collected in each JOL category by group in Experiment 3

Confidence JOL	0%	20%	40%	60%	80%	100%
	115	102	69	75	67	124
Binary JOL	No - Sure	No - Maybe	No - Guess	Yes - Guess	Yes - Maybe	Yes – Sure
	103	110	78	57	93	144

Table 7

Cue-justification and target-justification LSA scores by category and group in Experiment 3

Experiment	JOL category	Cue LSA score	Target LSA score	*t*-value	df	*p*-value	*d*
1	0%	.09 (.17)	.04 (.08)	2.65	113	.009*	0.25
	20%	.09 (.16)	.03 (.06)	3.58	100	<.001*	0.36
	40%	.09 (.17)	.04 (.07)	2.60	66	.011*	0.32
	60%	.13 (.17)	.11 (.20)	0.69	74	.494	0.08
	80%	.17 (.22)	.14 (.18)	1.00	65	.322	0.12
	100%	.14 (.20)	.15 (.21)	0.64	122	.523	0.06
2	No - Sure	.10 (.18)	.03 (.07)	3.72	100	<.001*	0.37
	No - Maybe	.09 (.15)	.03 (.06)	4.01	109	<.001*	0.38
	No - Guess	.05 (.09)	.04 (.09)	1.65	74	.104	0.19
	Yes - Guess	.09 (.15)	.08 (.16)	0.52	55	.607	0.07
	Yes - Maybe	.10 (.24)	.07 (.16)	1.09	89	.279	0.12
	Yes - Sure	.13 (.20)	.21 (.26)	2.90	143	.004	0.24

Note. Cue and target LSA score descriptives expressed as: mean (standard deviation). Results of paired-samples t-tests comparing the cue and target LSA scores within each JOL category are also reported. *s indicate significance at an alpha threshold of .05.

Table 8

N-gram analysis results for the numeric confidence JOL group of Experiment 3

JOL	*n*-gram	Count	Total	Proportion	*p*
0%	not remember this	10	11	.91	<.001
	not remember seeing	25	25	1	<.001
	remember seeing word	6	11	.55	.014
	I do not	42	60	.70	<.001
	do not remember	63	72	.88	<.001
	not remember	63	72	.88	<.001
	seeing word	6	13	.46	.035
	this word	17	41	.41	.003
	I do	43	62	.69	<.001
	seeing this	6	10	.60	.008
	at all	12	15	.80	<.001
	cannot recall	7	11	.64	.002
	do not	76	102	.75	<.001
	remember seeing	30	45	.67	<.001
	do	76	111	.68	<.001
	even	8	11	.73	<.001
	seeing	32	51	.63	<.001
	remember	76	272	.28	.004
	not	84	195	.43	<.001
	at	13	35	.37	.022
	all	15	19	.79	<.001
	no	13	15	.87	<.001
	now	9	21	.43	.026
	any	8	12	.67	.001
20%	I cannot remember	7	11	.64	.001
	I remember seeing	7	10	.70	.001
	cannot remember	12	23	.52	<.001
	but cannot	8	15	.53	.002
	word but	10	17	.59	<.001
	I cannot	11	22	.50	.001
	but not	11	27	.41	.006
	come	8	16	.50	.004
	other	10	28	.36	.026
	but	43	118	.36	<.001
	what	14	42	.33	.026
	it	40	160	.25	.040
	cannot	22	50	.44	<.001
	later	6	10	.60	.004
40%	second word	5	12	.42	0.011
	trying to	7	10	.70	<.001
	not sure	5	16	.31	0.04
	have vague	6	10	.60	<.001
	word and	6	22	.27	0.047
	of my	4	10	.40	0.027
	really	4	11	.36	0.038
	more	8	21	.38	0.003

	trying	7	10	.70	<.001
	be	9	34	.26	0.031
	out	8	23	.35	0.005
	vague	6	10	.60	<.001
60%	think I remember	4	11	.36	.050
	think it	7	17	.41	.005
	if I	4	10	.40	.036
	I think	15	63	.24	.025
	it is	12	28	.43	<.001
	sure	9	35	.26	.045
	can	8	28	.29	.045
	if	7	21	.33	.017
	think	26	108	.24	.003
	is	23	84	.27	.001
80%	I think I	8	24	.33	.006
	remember paired word	4	12	.33	.047
	I am	13	47	.28	.003
	paired word	8	31	.26	.028
	think I	11	38	.29	.004
	they	6	13	.46	.002
	paired	9	39	.23	.046
	or	11	32	.34	.001
	as	11	42	.26	.014
	pretty	5	13	.38	.014
	am	16	53	.30	<.001
100%	in my	11	26	.42	.030
	in	29	89	.33	.029
	very	6	12	.50	.033
	thought	6	12	.50	.033
	clearly	6	10	.60	.011
	imagined	6	10	.60	.011
	both	9	14	.64	.001
	made	15	38	.39	.018
	my	21	58	.36	.017
	because	16	37	.43	.005
	are	6	10	.60	.011
	and	39	126	.31	.024

Note. A count of occurrences of each *n*-gram in justifications for the corresponding JOL category are reported along with total number of occurrences, proportion of occurrence and *p*-value computed using the binomial test.

Table 9

N-gram analysis results for the binary JOL group of Experiment 3

JOL	*n*-gram	Count	Total	Proportion	*p*
No-Sure	not remember seeing	9	11	.82	<.001
	I do not	29	69	.42	<.001
	do not know	12	21	.57	<.001
	remember this word	6	10	.60	.003
	do not remember	22	48	.46	<.001
	right now	5	10	.50	.018
	I have	10	29	.34	.024
	seeing word	6	16	.38	.044
	sure I	6	13	.46	.015
	at all	8	14	.57	.001
	not remember	25	58	.43	<.001
	I do	29	74	.39	<.001
	not know	12	26	.46	.001
	seeing this	5	10	.50	.018
	this word	13	30	.43	.001
	know it	5	12	.42	.042
	no idea	13	20	.65	<.001
	do not	48	106	.45	<.001
	remember seeing	13	41	.32	.022
	do	48	123	.39	<.001
	recollect	5	12	.42	.042
	even	10	12	.83	<.001
	seeing	19	51	.37	.001
	idea	13	28	.46	<.001
	no	18	37	.49	<.001
	not	62	232	.27	<.001
	all	11	19	.58	<.001
No-Maybe	but I do	5	11	.45	.037
	but I cannot	7	14	.50	.007
	did not	8	19	.42	.015
	word but	10	19	.53	.001
	if I	7	18	.39	.035
	word	8	13	.62	.001
	not think	6	15	.40	.043
	remember word	10	26	.38	.018
	second word	6	10	.60	.004
	but I	26	70	.37	<.001
	remember what	9	15	.60	<.001
	maybe	11	25	.44	.003
	word	56	218	.26	.009
	did	9	20	.45	.006
	come	8	20	.40	.020
	more	7	13	.54	.004
	unsure	7	15	.47	.012
	remember	67	254	.26	.002
	but	51	151	.34	<.001

	what	16	50	.32	.026
	second	8	15	.53	.002
	cannot	24	70	.34	.002
	time	15	28	.54	<.001
	may	9	24	.38	.030
	back	7	17	.41	.025
	seen	5	11	.45	.037
	connection	7	18	.39	.035
No-Guess	I cannot remember	9	22	.41	.001
	but it	5	14	.36	.028
	cannot remember	15	47	.32	.001
	feel like	5	11	.45	.009
	I cannot	13	36	.36	<.001
	partner word	4	10	.40	.032
	partner	4	11	.36	.045
	guess	7	20	.35	.011
Yes-Guess	to do with	5	11	.45	.002
	I thought about	5	13	.38	.005
	I can remember	5	14	.36	.008
	something to do	5	11	.45	.002
	thought about	5	13	.38	.005
	do with	5	11	.45	.002
	can remember	5	14	.36	.008
	there is	5	12	.42	.004
	I could	8	25	.32	.002
	to do	5	14	.36	.008
	was something	5	10	.50	.001
	might be	4	10	.40	.011
	I was	4	10	.40	.011
	something to	5	11	.45	.002
	really	5	17	.29	.019
	feel	6	20	.30	.009
	get	5	11	.45	.002
Yes-Maybe	be able to	7	20	.35	.027
	am not sure	8	21	.38	.011
	I think I	11	21	.52	<.001
	but I am	9	17	.53	<.001
	I think it	7	17	.41	.010
	I am not	16	36	.44	<.001
	I am	21	75	.28	.006
	think it	8	20	.40	.008
	think I	12	36	.33	.009
	be able	7	20	.35	.027
	am not	17	38	.45	<.001
	not sure	13	36	.36	.002
	I think	23	57	.40	<.001
	I may	7	14	.50	.003
	but not	7	16	.44	.007
	remember it	8	25	.32	.046
	it is	12	39	.31	.014

	to	30	136	.22	.044
	similar	5	11	.45	.019
	something	16	50	.32	.003
	think	27	86	.31	<.001
	it	43	195	.22	.017
	forget	5	10	.50	.012
	be	11	37	.30	.037
	am	22	79	.28	.005
Yes-Sure	in my head	6	11	.55	.029
	in my mind	8	15	.53	.014
	I remembered	14	17	.82	<.001
	in my	17	31	.55	<.001
	I know	11	24	.46	.028
	in	27	69	.39	.007
	clearly	9	13	.69	.001
	words	19	52	.37	.050
	associated	20	49	.41	.011
	remembered	18	28	.64	<.001
	my	24	51	.47	<.001

Note. A count of occurrences of each *n*-gram in justifications for the corresponding JOL category are reported along with total number of occurrences, proportion of occurrence and *p*-value computed using the binomial test.

1

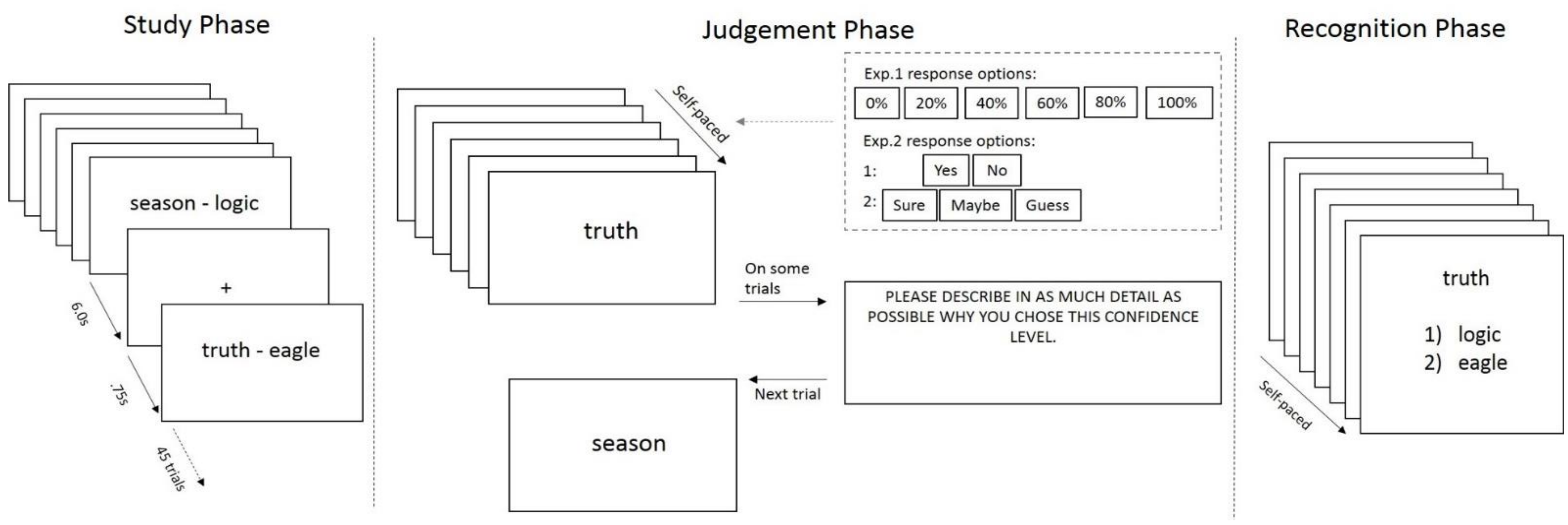

Figure 1: Schematic of experimental procedure. The three phases together constitute one experimental block. Participants completed two blocks, with a new set of items in each. In the judgment phase, participants gave a JOL with variation in response format across experiments. In Experiment 1, participants indicated their numeric confidence in one response. In Experiment 2, participants gave a binary judgment (*yes*/*no*) before indicating their verbal confidence in this judgement. On a subset of trials participants were asked to explain why they gave the particular JOL prediction on the preceding trial.

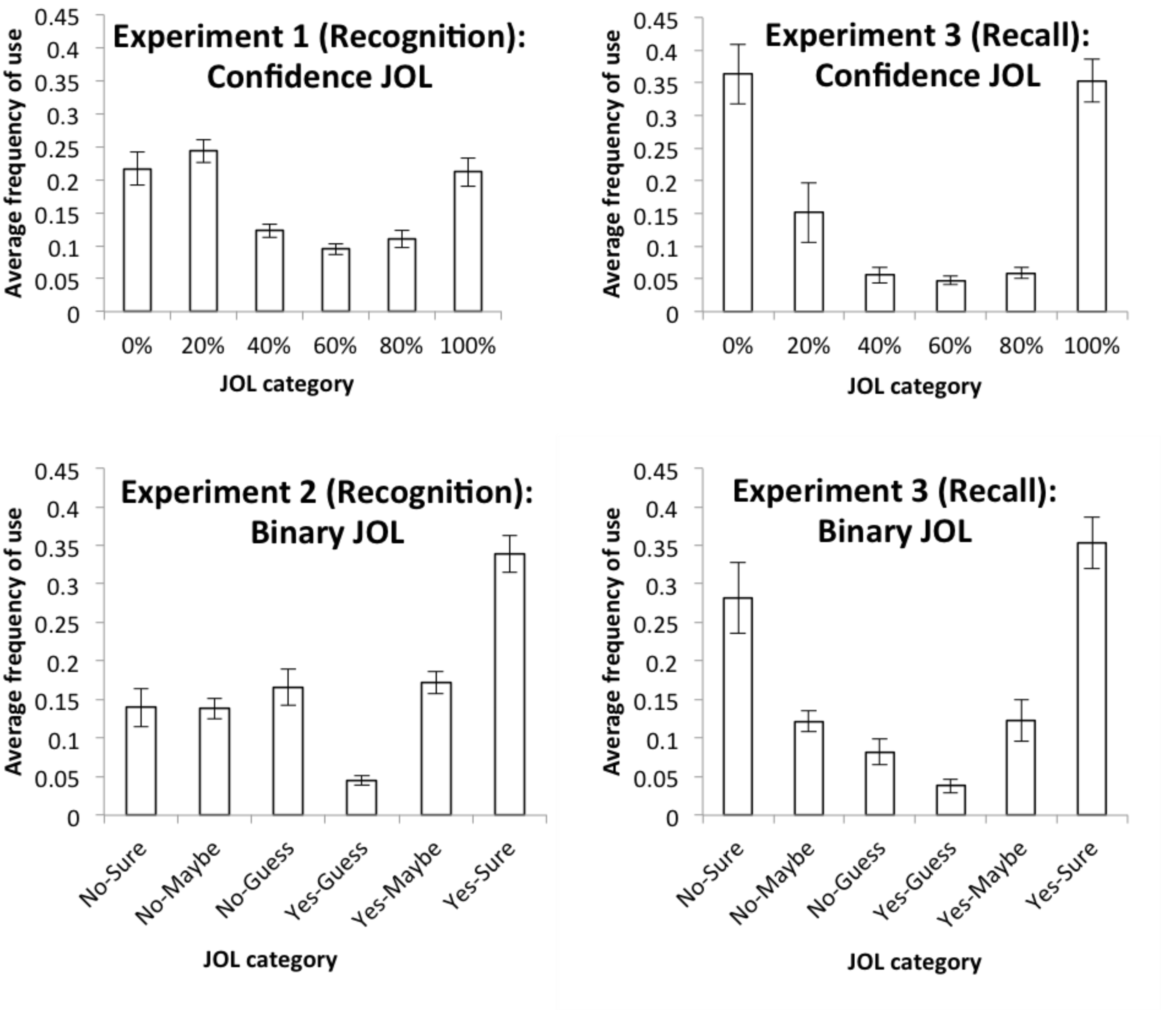

Figure 2: Mean proportion of trials in each JOL category by experiment. Error bars indicate standard error of the mean.