Published in final edited form as:
J Exp Psychol Learn Mem Cogn. 2014 January ; 40(1): 66–85. doi:10.1037/a0034059.

Relating the Content and Confidence of Recognition Judgments

Diana Selmeczy and **Ian G. Dobbins**
Washington University in St. Louis

Abstract

The Remember/Know procedure, developed by Tulving (1985) to capture the distinction between the conscious correlates of episodic and semantic retrieval, has spurned considerable research and debate. However, only a handful of reports have examined the recognition content beyond this dichotomous simplification. To address this, we collected participants' written justifications in support of ordinary old/new recognition decisions accompanied by confidence ratings using a 3-point scale (high/medium/low). Unlike prior research, we did not provide the participants with any descriptions of Remembering or Knowing and thus, if the justifications mapped well onto theory, they would do so spontaneously. Word frequency analysis (unigrams, bigrams, and trigrams), independent ratings, and machine learning techniques (Support Vector Machine - SVM) converged in demonstrating that the linguistic content of high and medium confidence recognition differs in a manner consistent with dual process theories of recognition. For example, the use of 'I remember', particularly when combined with temporal or perceptual information (e.g., 'when', 'saw', 'distinctly'), was heavily associated with high confidence recognition. Conversely, participants also used the absence of remembering for personally distinctive materials as support for high confidence new reports ('would have remembered'). Thus, participants afford a special status to the presence or absence of remembering and use this actively as a basis for high confidence during recognition judgments. Additionally, the pattern of classification successes and failures of a SVM was well anticipated by the Dual Process Signal Detection model of recognition and inconsistent with a single process, strictly unidimensional approach.

> "One might think that memory should have something to do with remembering, and remembering is a conscious experience."
>
> (Tulving, 1985, p. 1)

In order to examine the conscious experience of remembering, Tulving (1985) developed the Remember-Know procedure, which requires participants to indicate whether items endorsed as previously studied during a recognition test are Remembered (i.e., bring to mind a specifics of the prior encoding episode) or Known (i.e., known to have been recently experience but without specific recollections). According to Tulving, Remember and Know reports measure autonoetic (self-knowing) and noetic (knowing) consciousness respectively. Autonoetic consciousness, or the awareness of personal time including the past and the future, is characterized by retrieval from episodic memory (i.e., personally experienced

Correspondence concerning this article should be addressed to: Diana Selmeczy, Department of Psychology, Washington University in St. Louis, Campus Box 1125, One Brookings Drive, St Louis, MO, 63130, dselmecz@wustl.edu.
Diana Selmeczy, Department of Psychology, Washington University in St. Louis
Ian G. Dobbins, Department of Psychology, Washington University in St. Louis

events) while noetic consciousness, or learned knowledge that is unaccompanied by personal awareness of its acquisition, is characterized by retrieval from semantic memory (i.e., generalized knowledge). A host of research has examined the effects of various encoding and retrieval manipulations on Remember-Know rates demonstrating that these responses can vary independently and in opposite directions (for reviews see Gardiner & Java, 1993; Gardiner & Richardson-Klavehn, 2000; Rajaram & Roediger, 1997).

Since its inception however, memory theorists have interpreted Remember-Know responses differently, with some favoring a dual process view that posits separable retrieval processes or information dimensions contributing to Remember and Know reports (e.g., Jacoby, 1991; Rajaram, 1996; Wixted & Mickes, 2010; Yonelinas, 2002), while others assert that Remember-Know responses merely reflect differing strength levels along a single, undifferentiated strength of evidence dimension (e.g., Donaldson, 1996; Dunn, 2004; Wixted & Stretch, 2004). We refer to this latter approach as strictly unidimensional. Critically, these two perspectives potentially lead to different predictions regarding the content of justifications that observers might offer in support of their Remember and Know decisions. A strictly unidimensional model assumes that recognition evidence varies only in subjective intensity and thus to the extent it makes content predictions at all, it predicts that descriptions of varying intensity might differentiate confidence distinctions or remember versus know ratings. For example, observers might distinguish Remembering versus Knowing using words such as “definitely”, “certain”, “absolutely” for the former, and “somewhat sure”, “possibly”, or “likely” for the later. We refer to this general class of words as intensity modifiers. Furthermore, from a strictly unidimensional approach these modifiers would be the same for old and new judgments, since both rely upon the same unidimensional scale, with modifiers simply indicating extremeness with respect to the center of the scale.

In contrast, Tulving’s theorizing that Remembering and Knowing are linked to fundamentally different states of conscious experience (and different underlying memory systems) suggests that the content of justifications for these two reports should categorically differ. Additionally, given that highly confident judgments of novelty are not linked to the conscious experience of remembering, these should also differ considerably from Remember reports in terms of the contents of justification. However, very few studies have actually examined the memorial content associated with Remember-Know responses. That is, although the qualities associated with a remembering versus knowing are clearly described in experimental instructions given to participants before they are asked make the distinction (e.g., Rajaram, 1993), very few studies directly ask participants to describe additional thought processes or content beyond the dichotomous Remember/Know judgment. This is somewhat surprising, as the typical Remember/Know instructions rest on a host of assumptions about the kinds of conscious content and experiences participants should have during the two putatively different retrieval experiences; assumptions which have in large part, not been tested.

One of the few studies that examined freely provided memorial content during recognition memory was conducted by Strong in 1913 where participants reported their experiences during a recognition test. Strong noted that some words, generally of low confidence, were

recognized in the absence of any associations, while other reports clearly described emotions, objects, or thoughts associated with the study word. Gardiner, Ramponi, and Richardson-Klavehn (1998) more formally examined verbal descriptions accompanying Remember and Know judgments and had raters classify Remember transcripts into various descriptive categories (see also Dewhurst & Farrand, 2004). In Gardiner et al. (1998) participants completed a recognition test and had to indicate whether words reported as old were Remembered, Known, or simply Guessed. After completing the entire recognition test, participants had to verbally indicate to the experimenter what led them to recognize two randomly selected words as studied from each response category (i.e. Remember, Know, Guess). These verbal responses were originally collected to verify that observers were correctly following Remember-Know-Guess instructions (Gardiner, Richardson-Klavehn, & Ramponi, 1997), but after discovering the potentially informative detail provided in these reports, the authors more closely examined their content. The authors concluded that Remember responses "reflect the use of effortful strategies, associations, and imagery" (p. 5) and Know responses lack "any indication that they involved recollection of any specific contextual details" (Gardiner et al. 1998, p. 7). Additionally, two raters categorized remember responses in order to demonstrate particular characteristics that commonly occurred in Remember reports; these categories included intra-list associations, extra-list associations, item-specific images, description of item's physical features, and self-referential statements. The authors concluded that these subjective reports of Remembering and Knowing, along with other evidence (e.g. Gardiner & Gregg, 1997), reflect distinct states of awareness and support a dual-process view of recognition. Extending this finding, McCabe, Geraci, Boman, Sensenig, and Rhodes (2011) had participants think-aloud and generate thoughts associated with each individual study item. These descriptions were then compared with verbal justifications provided for words reported as 'old' during a recognition test, which was administered a day later using either Remember-Know or confidence judgments (Sure Old vs. Probably Old). Results revealed that recollection of think-aloud descriptions most often occurred during Remember and Sure reports; however, a low proportion of recollection also occurred in Know and Probably reports.

Although the Gardiner paper provides important evidence that the freely reported content of justifications for Remember and Know reports differs, some aspects of the design potentially limit the generality of the conclusions. First, as is standard practice, subjects were given extensive instructions about the presumed characteristics of Remembering and Knowing (and Guessing) prior to testing. This potentially imposed a demand characteristic that encouraged participants to provide justifications that aligned with the experimenters' instructions. Second, the justifications were provided after a delay with respect to encoding and when the initial Remember/Know distinctions were indicated, raising the possibility that forgetting or other delay related changes in content might have occurred. Third, justifications underlying new reports were not collected, which is important since neither Know nor New reports depend upon conscious recollection under the dual process framework, hence one might expect some similarities in the content provided to justify these reports. Finally, the characteristics potentially differentiating Remembering and Knowing were not statistically analyzed and thus their reliability is uncertain.

To address these questions, we examined the content of written justifications accompanying recognition decisions for both correctly identified old items (hits) and correctly identified new items (correct rejections) following two different, fairly typical encoding conditions. These justifications were collected immediately after each recognition response. Critically, the design also differed from Gardiner et al. (1998) in that we did not provide any descriptions of Remembering or Knowing to participants, but instead used conventional old/new response instructions along with simple ratings of report confidence (high/medium/low). This manipulation precludes subjects from providing justifications that might artificially align with current conceptualizations of Remembering and Knowing. Generally, under dual process models of recognition, high confidence old reports should be considerably more likely to contain experiences of remembering than medium confidence reports whereas medium confidence old reports should be more likely to contain experiences of knowing than high confidence old reports. Thus, it should be possible to obtain justifications that demonstrate important content differences between remembering and knowing even without telling the observers about the types of experiences thought to support the distinction, providing them these labels, or employing think aloud techniques. One might even expect that the observers would spontaneously use the actual terms formalized by Tulving when justifying high versus medium confidence recognition experiences; namely, ‘remember’ and ‘know’.

To address these questions, we focused on the written content of recognition justifications associated with high vs. medium recognition confidence, for old and new reports, using three different analyses. First, we analyzed the frequency of single, double, and triple word sequences, which we refer to as n-grams (unigrams, bigrams, and trigrams), across confidence categories to see whether particular sequences were more heavily associated with particular confidence levels; for example, is the bigram ‘I remember’ spontaneously offered in support of high confidence old reports reliably more often than medium confidence old reports? The second analysis used a machine learning algorithm (Support Vector Machine) trained on one data set to learn the differences between high and medium confidence old reports, and then tested on a second independent data set. Using the dual process signal detection (DPSD) model of Yonelinas (1994) we tested a set of predictions regarding how the classifier would both succeed and fail when applied to new content in the various confidence categories. The logic of these predictions is spelled out in the appropriate methods section below. Finally, we also had human raters code the justifications for four semantic characteristics that were potentially theoretically relevant for the distinction between Remembering and Knowing and derived in large part from the discussion of Gardiner et al. (1998) and from our own informal consideration of the justifications. Combined, these different analyses provide a much more complete analysis of recognition memory content than in prior literature, and statistically quantify putative differences between subjects’ self-reported content accompanying different levels of recognition memory confidence.

Experiments 1 and 2

Method

Participants—Experiment 1 included 27 Washington University students (average age =20, 16 female). Experiment 2 included 28 Washington University students (average age =20.3, 13 female). Participants were either paid $15 or received course credit for their participation. All participants provided informed consent in accordance with the University's Institutional Review Board.

Materials and Procedure—Observers entered their responses via keyboard and presentation and timing was controlled via Matlab's Psychophysics Toolbox (version 3.0.8) (Brainard, 1997; Pelli, 1997). For each participant, words were randomly selected from a 1216 item pool with an average of 7.09 letters, 2.34 syllables and Kučera-Francis frequency of 8.85.

Participants completed a total of two study/test cycles. Experiment 1 and 2 were identical in all aspects except for the encoding task performed during study. During the study phase in Experiment 1 participants performed a syllable counting task (1,2, 3 or more syllables?), while in Experiment 2 participants were simply told to memorize the presented words for an upcoming memory test. The use of two moderately different encoding tasks ensured that our content analysis findings would be fairly general to the types of tasks often used in recognition memory research. Participants were given a maximum of 2 seconds to make a syllable judgment (Experiment 1) or view each presented word (Experiment 2). Recognition testing immediately followed each study phase, with subjects indicating whether randomly intermixed old and new items were studied ("old") or novel ("new") (100 old items, 100 new items). After each self-paced old/new recognition decision, participants provided confidence on a 3-point scale (low, medium, high). Participants were told that on small subset of trials they would be asked to justify their confidence rating with the following instructions: "Please describe in as much detail as possible why you chose (low/medium/high) confidence level." Participants provided a typed justification for correctly identified old responses (hit) and correctly identified new responses (correct rejection) for each confidence level (low, medium, high) once throughout each test phase for a maximum of 6 total justifications per study/test and 12 total justifications for the experiment (2 per each combination of confidence and response type). Participants' first justification was prompted after their third response of a particular confidence type (e.g., third high confidence hit) in order space out justifications throughout the test. The current report focuses primarily on high and medium confidence justifications because these are most theoretically relevant to Tulving's original Remember/Know distinction. However, the SVM analysis also examines low confidence reports as the DPSD model makes an interesting prediction with respect to these as well.

Results and Discussion

Preparation of typed justifications for analyses—Spelling errors were corrected for each justification and contractions were completed (e.g., replaced "don't" with "do not"). We omitted four responses in Experiment 1 and four responses in Experiment 2 due to

incomplete responding (e.g. participant accidentally hit enter before typing in a response). Additionally, three responses in Experiment 1 and two responses in Experiment 2 were removed because participants clearly indicated they intended to give a different response (e.g. "I meant to choose medium confidence level."). Table 1 indicates the total number of responses collected for each response type and confidence level totaled across the two study/test cycles and subjects. Table 2 provides a basic summary of overall recognition performance. Table 3 provides the average proportion of each response type by confidence level. The Supplement contains the detailed transcripts of each participant's reported justifications for all confidence levels.

Linguistic, n-gram Analysis

Hits—Although participants provided justifications for high, medium, and low confidence reports, here we compare high vs. medium confidence reports. Tulving's original characterization of autonoetic consciousness suggests that experiences of remembering should be linked to high confidence in the context of simple recognition testing, and Yonelinas' (1994) DPSD model specifically assumes recollection leads to high confidence. Thus, we expect high confidence reports to often contain recollection-linked content not associated with medium confidence reports. Alternatively, if high confidence reports simply reflect greater intensity of a unidimensional memory signal, high confidence reports should instead contain intensity modifiers indicating greater certainty or describing a more vivid memory signal relative to medium confidence reports, but the nature of these modifiers would not suggest a fundamentally different experience across medium and high confidence old reports.

Our first analysis examines the frequency of unigrams, bigrams, and trigrams, used during response justifications for hits. The provided justifications were collapsed across runs and across subjects into eight documents, namely, high confidence old reports, medium confidence old reports, high confidence new reports and medium confidence new reports, for two separate experiments. The key question was whether the frequency of n-gram usage reliably differed from chance expectations (viz. 50 percent) across the confidence distinction captured by the documents. Using MATLAB (2007) we counted the occurrence of each n-gram in high confidence and medium confidence documents separately for Experiment 1 and Experiment 2. A given n-gram does not differentiate between high and medium confidence when its distribution is consistent with a binomial probability of 0.50 given its total frequency of occurrence. If this null hypothesis is rejected, then the n-gram is assumed to reliably characterize a certain level of confidence. For each n-gram, we calculated the z-value associated with a binomial distribution with parameters, p=0.5 and N= total number of occurrences. Positive z-values indicate that the word occurred more often in high confidence reports whereas negative z-values indicate that the word occurred more often in medium confidence reports. To increase power, we combined z-scores from Experiment 1 and Experiment 2 weighting each by Stouffer's method (Stouffer, Suchman, DeVinney, & Star, 1949) using relative frequency. We then used the normal approximation to the binomial to determine the p-values associated with each combined z-value and the results of these analyses are reported in Table 3. In order to reduce the number of comparisons performed we did not include words that occurred infrequently across the two experiments, using the

median total frequency of occurrence as a lower cutoff. Additionally, because this approach nonetheless involves a large number of paired comparisons, we provide a correction for multiple comparisons using the false discovery rate procedure (Benjamini & Hochberg, 1995). Thus Table 4 reports both the uncorrected p-values and the FDR adjusted p-values. Because of the novel nature of this investigation, we provide both the corrected and uncorrected p-values for the reader and used an uncorrected two-tailed p-value of .06, which captured important distinctions supported via the machine learning approach we report later in the results section. The same n-gram analysis was conducted for bigrams and trigrams.

Based on Tulving's notion of remembering one might expect to see the word "remember" and phrases associated with conscious remembering experiences differentiating high from medium confidence old recognition judgments. Scanning of Table 4 confirms this prediction. Fully characterizing the table is difficult, however, the n-grams seem to fall into at least two broad categories, namely those reflecting intensity modifiers (e.g., 'positive', 'distinctly', 'I vaguely') versus those linked to the conscious experience of remembering (e.g., 'I remember', 'I thought about') which generally predicted high recognition confidence. Additionally, justifications of high confidence also sometimes contained temporal information, for example "when I", "when this", "after", and "when I saw". These temporal content words such as "when", "when I", and "first" typically referred to when the word appeared during the prior study phase (e.g., "I visualized cracking open a walnut when this word first appeared on screen"). "Myself" was generally used in the context of "thinking to myself" or "saying to myself" indicating that participants are recalling prior thoughts about the study word in a manner consistent with Tulving's assertion that remembering is associated with autonoetic consciousness. That is, observers are indicating that they are aware that the word constituted an element of a particular past personal episode in which they participated. The phrases "I remember saying", and "I thought about" also occurred more often in high confidence reports and likely indicate specific associations previously linked with the recognition probe.

The n-gram analysis clearly supports Tulving's original use of the word 'Remember' as highly diagnostic of autonoetic recognition experiences. However, its use appears more complicated than simply indicating the conscious experience of recollection. Participants also appear to use the absence of Remembering to guide other confidence assignment. Thus, the word "remember" when associated with medium confidence reports tends to reflect the failure or poverty of remembering, for example, "Think I remember" and "I vaguely remember." Additionally, as we will see below in the section examining correct rejections, observers also use the absence of remembering in a strategic fashion when assigning confidence to judgments of novelty.

Overall, it is clear from Table 4 that the use of the word 'remember' without negative modifiers, and often in conjunction other contextual information, is largely confined to high, not medium confidence old judgments. In contrast, the term 'know' is clearly not used spontaneously by participants to justify medium confidence recognition judgments. This absence is consistent with the tendency of many researchers to eschew this label because the lay usage of 'knowing' connotes high confidence or certainty and thus the label could be potentially confusing to subjects. Instead researchers often use a Remember/Familiar

distinction (e.g., Dobbins, Kroll, & Liu, 1998; Donaldson, 1996; Norman, 2002) and indeed the word “familiar” did occur more often during medium confidence reports than high confidence reports in Table 4 (e.g., ‘looks familiar’, ‘familiar but’).

A variety of words were used significantly more often during medium confidence justifications as opposed to high confidence justifications. For example qualifiers reducing certainty were quite common including words such as “but”, “but I”, “if”, “if I” that when used during medium confidence reports reflected general uncertainty (e.g., “I feel like I already saw this word, but I am not entirely positive”). Negations such as “not”, “do not”, “I do not”, “am not”, “I am not” were also common and may indicate a lack of memory retrieval (e.g., “I am not totally sure if I saw this word”). The single word “sure” also occurs more often in medium confidence reports; however, it tends to be preceded by the word ‘not’ or other qualifying adjectives such as “not completely sure” and “am not sure”. Perhaps the most surprising and robust indication of medium confidence was use of ‘am’ (“am not”, “am”, “am pretty”, “I am not”, “but I am”, etc.). We are not aware of a prior characterization of Knowing or Familiarity that necessarily predicts this, but here we speculate that it reflects that fact that recognition based upon familiarity is necessarily grounded in the perceptual present. Indeed, researchers often talk of ‘feelings of familiarity’ and this reflects the assumption that familiarity reflects a current, perhaps relatively automatic response to a memory probe. This is also conceptually consistent with dual process frameworks that focus on the fluency of processing at the time of recognition report (Jacoby, 1991), or which link familiarity to perceptual processing of the probes at the time of test (Mandler, 1980; Rajaram, 1996). Thus, the current data suggest that endorsing a memory probe as studied, while concentrating on one’s current phenomenological reactions to the probe, signifies medium confidence recognition. In turn, this may serve as a feasible operational definition of familiarity-based endorsements. Also, to presage our findings when looking at correct rejections, we note in advance, that many of the words linked to medium confidence old reports were also indicators of medium confidence new reports. This similarity in content across old and new materials for medium confidence content is something we also examine in the support vector machine analysis and it is consistent with the idea that medium confidence old and new reports lie on the same single dimension.

Correct Rejections—Response justifications associated with correct rejections were analyzed using the same method as described above for hits. Results are shown in Table 5.

Surprisingly, the word “remember” was also associated with the correct rejection of new items. High confidence correct rejections use words such as “remembered”, “have remembered”, and “would have remembered”, suggesting that observers use the clear absence of remembering as indicative that an item is new. Critically, these results suggest that observers use subjective memorability heuristics (Brown, Lewis, & Monk, 1977) during recognition judgments perhaps far more frequently than currently thought, with this heuristic being fairly important to assigning high confidence to correct rejection reports. Although subjective memorability is often presumed to be associated with high confidence correct rejections, this is the first experiment to empirically verify that observers do in fact routinely engage in thought processes consistent with this heuristic.

In contrast, medium confidence correct rejection justifications, were unsurprisingly, more likely to indicate uncertainty. For example, medium confidence responses contained negation words such as “not”, “cannot”, “I cannot”, “am not”, “I am not” more often than high confidence responses. Additionally, the qualifiers “but”, “but I”, “but it”, “but I am”, “the list but”, and “this word but” occur more often during medium confidence reports.

Considering both hits and correct rejections, the n-gram analysis suggests that remembering or its unexpected absence is used to indicate both high confidence hits and high confidence correct rejections during standard recognition testing. This underscores the fact that observers put a premium on this introspective state during recognition memory testing consistent with the idea that it is linked to a fairly distinct and salient form of conscious experience. When it is clearly present it is often accompanied by other contextual details linked to the prior experiences and recognition is highly confident. When it is surprisingly absent given materials that the observer finds personally distinctive or salient, then rejection is often highly confident. In contrast, the results also suggest that feelings of familiarity are used when endorsing items as either recognized or novel with medium confidence, and observers appear to be heavily focused on their current phenomenological reactions to the probes in these situations with n-grams containing ‘am’ being highly prevalent. Indeed, there is a striking similarity of the terms indicative of medium confidence old and medium confidence new reports.

Rater Analysis of Justifications

Although linguistic n-gram analysis is informative, the approach cannot capture subtle semantic aspects of justifications that may span highly variable or complex sequences of words. Given the discussion available in Gardiner et al. (1998), and after reading through all the responses we created an ad-hoc set of four categories of justification content that we were interested in examining. These categories included 1.) Personal experiences outside of experiment, 2.) Imagery, feelings, and thoughts, 3.) Notable absences of memory, and 4.) Strategies to memorize words. Specific instructions regarding these ratings are included in the Appendix. Although these categories were ad-hoc, they are generally consistent with the descriptions used by Gardiner et al. (1998) for Remember-Know reports. For this analysis we only used responses from Experiment 2 in order to decrease the number of responses that needed to be coded and because we thought that the free encoding manipulation in that experiment might have encouraged more complexity or variability in the subsequent justifications. Raters were provided with detailed instructions and examples of each category, and rated each response for the presence or absence of a particular category. The justifications provided to the raters were randomized and raters were blinded to response type and confidence level. Raters were recruited through on campus flyers and were paid $15 for their help. We collected three raters per category and selected the two having the highest inter-rater reliability (inter-rater reliability ranged from 0.62– 0.71). The presence of a particular category of content was scored dichotomously by each rater (1 present, 0 absent) and then the scores were summed across raters so that each response justification had a rating of 0 (category indicated absent by both raters), 1 (category indicated present by one rater), or 2 (category indicated present by both raters). These scores were then averaged across study/test cycles for each confidence (high or medium) and response category (hit or

correct rejection). Unlike the n-gram analysis, this analysis approach allowed us to treat subjects as a random factor and use conventional ANOVA analyses.

To assess whether particular categories of content occur more often for certain types of justifications, we conducted a 2 × 2 repeated measures ANOVA with factors of item type (correct rejections vs. hits) and confidence level (high vs. medium) for each ad hoc content category (See Table 6 for descriptive statistics). For Category 1 (personal experience outside the experiment), results revealed a main effect of confidence level ($F(1,23)=15.49$, $\eta^2 =0.402$, $p<.001$), reflecting higher occurrences during high confidence responses relative to medium confidence responses. The main effect for item type ($F(1,23)=0.22$, $\eta^2 =.009$, $p=.65$) and the interaction between item type and confidence level ($F(1,23)=0.61$, $\eta^2 =.026$, $p=.44$) were not significant. Overall these results suggest that high confidence reports mention personal experiences outside of the experiment more often than medium confidence reports, for both hits and correct rejections. Although it is somewhat surprising that correct rejections also contain significantly more instances of personal experiences during high confidence reports, this may be because high confidence correct rejections often involved subjective memorability heuristics (see Category 3 analysis below). Although the instructions tried to guide the raters to focus on personal experiences outside of the experiment that arose during prior study of the words, new responses containing a notable absence of memory also tended to be linked to personal experiences as well (e.g., “I really like limes and I would have remembered if they were on the list”). Thus the category raters may have generally rated these instances of subjective memorability as a ‘personal experience outside the laboratory’ and ignored the instructions noting that we were looking for cases in which these experiences were reflected upon during the word’s initial study (and then subsequently reported during testing).

For Category 2 (Imagery, feelings, and thoughts), results revealed a main effect of confidence ($F(1,23)=17.74$, $\eta^2 =.435$, $p<.001$), reflecting higher occurrences during high confidence responses relative to medium confidence responses. The main effect of item type ($F(1,23)=3.44$, $\eta^2 =.130$, $p=.08$) and the interaction between item type and confidence level ($F(1,23)=3.59$, $\eta^2 =.135$, $p=.07$) approached significance. Although the interaction did not meet conventional levels of significance, we conducted follow up planned comparisons based on prior literature that suggests remember reports contain extra-list associations and item specific images (Gardiner et al., 1998) and recollected responses are likely to be associated with high confidence reports (Yonelinas, 2002). Thus, we hypothesized that high confidence hits should contain more instances of Category 2 than medium confidence hits. Follow up t-test revealed that Category 2 occurrences in high confidence hits were more common than during medium confidence hits ($p <.001$). In contrast, high confidence correct rejections did not significantly differ from medium confidence correct rejection ($p=.131$). Overall, these results suggest that high confidence reports contain more instances of imagery, feelings, and thoughts than medium confidence reports and this pattern seems to be more robust for hits.

For Category 3 (Notable absence of memory), results revealed a main effect of confidence level ($F(1,23)=4.28$, $\eta^2 =.157$, $p=.05$), reflecting higher occurrences during high confidence responses relative to medium confidence responses. The main effect of item type

($F(1,23)=84.14$, $\eta^2 =.785$, $p<.001$) was also significant, reflecting higher occurrences of Category 3 during correct rejections vs. hits. These main effects were conditioned by a significant interaction between item type and confidence level ($F(1,23)=6.68$, $\eta^2 =.225$, $p=.02$). Follow up tests (Tukey's HSD) revealed that high confidence correct rejections contained significantly more instances of Category 3 than medium confidence correct rejections ($p=.01$). In contrast, the difference between high confidence hits and medium confidence hits was non-significant ($p=0.99$). These results suggest that notable absences of memory occur more often during correct rejections than hits, and critically they occur more often during high confidence correct rejections than medium confidence correct rejections.

For Category 4 (Strategies to memorize words), results revealed a main effect for confidence level ($F(1,23)=14.35$, $\eta^2 =.384$, $p<.001$), reflecting higher occurrence during high confidence vs. medium confidence reports. The main effect of item approached significance ($F(1,23)=3.94$, $\eta^2 =.146$, $p=.06$), reflecting higher occurrence of Category 4 during hits vs. correct rejections. These main effects were conditioned by a significant interaction between confidence level and item type ($F(1,23)=6.36$, $\eta^2 =.217$, $p=.02$). Follow up tests (Tukey's HSD) revealed no significant difference between high vs. medium confidence correct rejections ($p=0.61$). In contrast, instances of Category 4 occurred significantly more often in high confidence hits relative to medium confidence hits ($p<.001$). In summary, subjects report prior strategies to memorize words more often when providing high vs. medium confidence hits, while they do not do so more often for high vs. medium confidence correct rejections.

Overall these category analyses demonstrate that high relative to medium confidence hits contain more instances of personal experiences outside of the experiment; imagery, feelings, and thoughts; and more instances of the remembrance of memorization strategies recruited during study. These results are consistent with the descriptions of Gardiner et al. (1998) of remember reports, however Gardiner and colleagues did not directly statistically compare content categories. Additionally, imagery, feelings, and thoughts as well as strategies to memorize words seem to occur more often during hits relative to correct rejections. In contrast, correct rejections contain significantly more instances of notable absences of memory, and critically this occurs more often during high confidence correct rejections than medium confidence rejections. Thus, it appears as though observers use the subjective memorability heuristic when justifying new reports and this occurs more often during high confidence; a finding that converges with the n-gram analysis.

Support Vector Machine Analysis

Machine learning algorithms are often applied to text classification problems with one of the most successful being Support Vector Machines (SVM). The SVM we initially used attempts to parse high and medium confidence recognition hit justifications using a linear decision boundary. The goal of the classifier is to find a decision surface that isolates the categories and which is maximally distant from the most confusing cases from the two categories (Hamel, 2009). This is easy to illustrate in the case of a perceptual classification involving stimuli with two continuous stimulus features such as height and weight (Figure 1). Here, each dimension is a feature value and the plus and minus signs illustrate the

distribution of two stimulus categories in this 2D space. The cases touching the margins of the decision surface are known as the 'support vectors' and again, the goal of the algorithm is ensure that these are maximally distant from the decision surface, based on a large body of research demonstrating that this constraint leads to classifiers that generalize well (Hamel, 2009). This general approach is known as maximum margin classification.

In the case of text classification, each unique word in a document constitutes a feature and different coding schemes can be used to quantify the feature values of words within the documents. Here we used a simple binomial scheme which simply indicates whether the feature/word is present or absent (1 or 0) in the combined pair of justifications for each subject's confidence category. The algorithm trains and tests on a document term matrix in which each row represents a provided confidence justification and each column a particular word feature. Prior to training several choices have to be made about which terms from the texts should be incorporated into the document term matrix. The current training and testing matrices were constructed by removing all punctuation, converting all words to lowercase, and removing all numbers. Additionally, two procedures normally done in text classification were omitted. First, we did not use the stem word procedure. This procedure truncates or reduces all tenses and uses of a word to a lowest common element and is efficient and important for large-scale text classification. For example, if one wanted to classify a document as reflecting a discussion of memory, then one would not want to treat 'remember', 'remembers', 'remembered', or 'remembering' as different instances/features and they would all be collapsed into the stem 'remember'. In contrast, we assumed that different tenses of words such as remember might be distributed differently across confidence reports and hence avoided the stem word procedure. The other procedure often used is the removal of so called stop words such as 'the', 'I', 'is', 'am', etc., that are often considered useless for document classification and which therefore unnecessarily increase the size of the document term matrices. However, because our initial analyses already demonstrated that such words were likely useful we omitted this procedure as well.

The SVM was implemented using the R statistical language (R Core Team 2012) and the package RTextTools (Jurka et al. 2012). We trained the SVM on justifications that were collected in Experiment 1 and then validated the classifier on independent justifications collected in Experiment 2[1]. Because these are separate experiments using separate subjects, this is a strong validation test. The classifier was trained with a linear kernel and a cost value of .10.

Dual Process Signal Detection Model Predictions—As noted in the introduction, if recognition judgments were based on a strictly unidimensional strength signal then at best one would expect that confidence justifications might be mildly distinguishable based on intensity modifiers. For example high confidence hits might be accompanied by words such as 'definitely', 'certain', 'absolutely', whereas medium confidence hits might be accompanied by content such as 'possibly', 'probably', 'likely', etc. Additionally, these same terms should tend to distinguish high versus medium confidence correct rejections as

[1]The main findings of the SVM analysis also replicate when training on Experiment 2 justifications and validating on Experiment 1 justifications.

well, since intensity extremes in either direction are analogous. The dual process signal detection model (DPSD) of Yonelinas (1994) model begins with such a unidimensional familiarity process in which evidence varies continuously either towards familiarity or novelty and hence confidence distinctions based on that familiarity process in isolation would be predicted to differ only in terms of intensity modifiers (Figure 2). However, the model also assumes some old items can elicit conscious recollection of prior contextual information, and assumptions about this recollection process compared to the familiarity process lead to fairly specific SVM predictions. First, recollection is assumed to reflect a threshold retrieval process. Thus some portion of studied materials will fail to exceed threshold and hence no conscious recollections will be available (the threshold assumption). Critically, recollection should also be completely absent for items judged new. The second assumption is that recollection is very highly valued by the participants in recognition situations, such as standard recognition confidence paradigms, in which observers rate materials as old or new and provide basic confidence judgments (viz., high, medium, or low). Given this, successful recollection, no matter how modest, is assumed to lead the subject to report high confidence. This is the recollection mapping assumption. It is important to note that neither the recollection threshold nor recollection mapping assumptions require one to believe all recollections are equally vivid or complete. The recollection mapping assumption merely assumes subjects subjectively rate the utility of recollection much higher than that of familiarity during recognition, even if the former entails only modest recollections of prior context. The recollection threshold assumption instead assumes that conscious recollection can completely fail for some subset of studied materials and is absent for novel materials. However, when an old item exceeds threshold, recollections can vary in extent and specificity, although the DPSD mapping assumption again assumes that even modest recollections will map to high report confidence in this type of paradigm.

Given these simple assumptions, the DPSD model and prior dual process theory make a series of novel and specific predictions about the expected performance of a SVM classifier when applied to the content of confidence justifications. Before discussing these we wish to draw the distinction between classifier sensitivity and classifier specificity. Sensitivity refers to whether or not the classifier can successfully distinguish the categories of a validation data set of the same kind upon which it was trained. So for example, does a classifier trained to distinguish Hemlock from Pine do well when it sees a new set of Hemlock and Pine data? Specificity however, refers to whether the learned distinction is relatively unique among a broader pool of possible categories. For example, if the above classifier trained on Hemlock and Pine completely failed in every attempted dichotomous classification of other conifers, one would contend that the learning was highly specific. This implies that some of the features of Hemlock and/or Pine that were critical for classification are also unique among conifers as a whole. As we note below, the DPSD makes predictions regarding both sensitivity (categorical distinctions within hits) and specificity (the uniqueness of learned distinctions to hits versus correct rejections).

1. High Sensitivity to High vs. Medium Confidence Hit Content: This prediction arises from the recollection mapping assumption that recollection is restricted to high confidence

hits and absent from medium confidence hits, which instead are assumed entirely reliant upon familiarity. If this assumption is true (or largely true), then the classifier will perform well because the content of the two confidence reports will strongly differ.

2. High Specificity to High vs. Medium Confidence Hits Content (Relative to Correct Rejections): In contrast to a strictly unidimensional model, the DPSD model predicts that a classifier trained on the distinction between high and medium confidence hits will catastrophically fail when applied to high and medium confidence correct rejections, because of the recollection threshold assumption. High confidence correct rejections cannot contain recollective content under the model, and this content is the key information supporting high classifier performance when applied to hits. Interestingly, the DPSD model not only predicts classifier failure, but the nature of the failure. Because the classifier depends upon episodic recollection when trained on high confidence hits, and such content is absent in high confidence correct rejections, it should classify most correct rejections as medium confidence regardless of whether the actual confidence was medium or high. That is, it will be biased towards medium confidence classifications because the recollection features it uses for high confidence hits are absent from the correct rejection data.

3. Modest Sensitivity to Medium versus Low Confidence Hit Content: Under the DPSD model these two response categories are distinguished only by continuous familiarity. Thus they represent only an approximate or fuzzy categorical distinction and so success is expected to be modest, dependent primarily on differences in the described intensity of the familiarity feelings. Thus there should be a clear decline in classifier performance when one compares high and medium confidence hit success rates to medium and low confidence success rates. In other words, under the DPSP model, there is less information available to distinguish medium and low confidence reports than high versus medium confidence reports and so classifier performance should decline when going from the latter to the former.

4. No Specificity for the Medium versus Low Confidence Hit Content: Unlike the case when considering high versus medium confidence hits (where specificity is assumed high), the DPSD model predicts that medium versus low confidence judgments rely on subjective intensity differences that are analogous for both hits and correct rejections. Thus although the classifier trained on medium versus low confidence hits is expected to achieve only modest sensitivity, it nonetheless should transfer that modest ability well when applied to correct rejections as these also rely on moderate differences in perceived familiarity.

These four novel predictions about the performance of the SVM classifier represent a fairly strong test of dual process theory ideas and they are clearly not consistent with a strictly unidimensional view of recognition evidence. To preview the results, all four predictions were confirmed.

When applied to high versus medium confidence hit justifications in Experiment 2, the classifier trained on Experiment 1 performed extremely well, correctly labeling 93% of the text justifications. Table 7 shows the confusion matrix resulting from the validation test. In contrast, when this same classifier was applied to high versus medium confidence correct rejection justifications in Experiment 2, performance plummeted to 58%, a value not

different from chance (p = .27). Additionally the decline in performance was significant (.93 vs. .58; χ^2=15.49, p < .001) with the proportions demonstrating both the sensitivity and the specificity of the classifier. This in turn means that there are features of the reports that yield clear differences between high and medium confidence hits, and that these features are not present in high versus medium confidence correct rejections, a pattern consistent with the idea that recollection clusters in high confidence hits and is absence in correct rejections. Additionally, as anticipated above, the classifier failed in a particular manner when applied to correct rejections. Namely, it 'thought' that most (77%) correct rejections were of medium confidence, regardless of actual confidence (Table 7). Again, this should occur under the DPSD framework because correct rejections should lack the content uniquely linked to high confidence hits, namely episodic recollection.

Turning to medium versus low confidence hits, the classifier achieved only modest sensitivity at 68%. This is consistent with the idea that this distinction rests only on graded familiarity differences, and this success rate is reliably lower than that achieved by the classifier used for high versus medium confidence hits (.93 vs. .68; χ^2=8.84, p = .003). This drop off in performance is anticipated by the DPSD because it assumes an actual categorical distinction between high and medium confidence hits, but only a graded familiarity difference between medium and low confidence hits. Finally, when the classifier trained on medium versus low confidence hits was applied to medium versus low confidence correct rejections, there was minimal decline in its modest performance (63%). Indeed its success rate for hits and correct rejections did not reliably differ (.68 vs. .63; χ^2=0.11, p = .739) which demonstrates that the classifier has no material specificity whatsoever; a finding also consistent with the DPSD framework and the notion that medium versus low confidence judgments rest on analogous familiarity intensity differences for the two classes of materials.

In order to examine the similarity of findings across the n-gram analysis and the SVM approach, we extracted the feature weights used in the SVM algorithm when trained on high versus medium confidence hits. These weights represent relative feature importance during classification with positive values indicating features predicting high confidence and negative values indicating features predicting medium confidence. Table 8 shows the 40 most influential tokens (20 in each direction) out of the 427 words actually in the classification algorithm. The table converges with the n-gram analyses above in many respects. For example, words such as 'remember, 'being', 'I', and 'thought' were particularly discriminant towards high confidence hit classification. In contrast, words such as 'but', 'not', 'am' and 'vaguely' were particularly discriminant towards medium confidence hit classification. Unlike the n-gram analysis however, the SVM enables the classification of individual subject justifications and it also was able to illustrate the sensitivity and specificity of the distinction between high and medium confidence hit justifications.

Finally, for the sake of thoroughness, we also applied the classifier, trained on the high versus medium confidence hits in Experiment 1 here, to the verbal justification data of Gardiner et al. (1998). It performed quite well, classifying 88% of the justifications correctly. Given the differences in paradigms, this outcome is remarkable and it supports the idea that recollection and familiarity processes, as defined by Tulving and captured in the

Remember/Know paradigm are reliably (if not exclusively) mapped to high and medium confidence reports in our basic recognition paradigm.

In summary, the SVM analysis demonstrated a highly reliable content difference between high and medium confidence hits. This distinction was specific to hits as the same classifier failed when applied to correct rejections indicating that there was content in hits that was unavailable in correct rejections. Additionally, the manner of the failure, in which most correct rejections were labeled as medium confidence, indicates that the content that was missing, was that present in high confidence hits, namely reports of recollection. In contrast, when moving to medium versus low confidence hits, the classifier demonstrated a limited but reliable ability to parse the report content (modest sensitivity). However, this limited ability was preserved when we applied the same classifier to medium and low confidence correct rejections, demonstrating that the content distinctions were shared across hits and correct rejections in this case and presumably reflected gradations in perceived familiarity.

General Discussion

The current findings demonstrate that the content of recognition justifications varies reliably across confidence categories (high versus medium confidence) and report types (hits and correct rejections). This was illustrated using three complementary analysis methods. The n-gram analysis identified unigrams, bigrams, and trigrams that differentiated confidence levels and the latter were particularly informative. For example, the trigrams ‘I would have’ and ‘would have remembered’ were highly indicative of high versus medium confidence correct rejections and illustrated the use of a subjective memorability heuristic, whereby the observer used the absence of remembrance for words judged distinctive (presumably combined with some sensation of low familiarity or high novelty) as a strong indicator that the word is unstudied (Brown et al., 1977; Dobbins & Kroll, 2005; Ghetti, 2003). Although prior work has suggested the use of such a heuristic, it has not been validated through content analysis such as done here, and it has not been thought to routinely play a role in standard recognition paradigms because the materials are not manipulated to be particularly personally distinctive to the observers. However, the current report challenges that notion. Despite this, the n-gram approach has drawbacks. It requires the collapsing of individual reports into single documents and combining experiments to achieve sufficient power. It also faces a large multiple comparison problem during statistical inference.

We also used human raters potentially capable of identifying certain themes that span larger sequences than triplets and/or which may be expressed in a highly variable fashion across individuals. For example, observers might use vastly different strategies during encoding with a majority of the words used to describe those strategies differing across reports. Raters can presumably overcome this type of data variability and identify these high level constructs in the justifications. Of course, doing so requires understanding on the part of the raters and our use of raters illustrated both this benefit and drawback. Although we gained converging evidence for the remembrance of prior strategy use, and for the use of a subjective memorability heuristic, it appeared that the raters might not have been as selective as we wished when rating personal experiences outside of the experiment (Category 1). In this light it is probably important to note that the investigators themselves

may be more optimal raters (provided they are blinded during rating) as they are likely to have a fuller understanding of the characteristics of the particular heuristics and strategies that may be present in the texts.

In contrast to n-gram approach, the use of machine learning algorithms such as Support Vector Machines is highly powered, as demonstrated here by the successful ability to highly reliably classify individual observer justifications (combined for the two reports of each confidence category) that where wholly independent of the training of the classifier. The method also allowed us to jointly examine the sensitivity and specificity of the classifier demonstrating high sensitivity and specificity for the distinction between high and medium confidence hits, and modest sensitivity with no specificity for the distinction between medium versus low confidence hits. Despite these strengths, the classifier treats single words as features (so called Bag-Of-Words approach), and thus it may miss subtle conceptual information that critically depends not just upon the presence of particular words in a text, but also upon their ordering. Nonetheless, the approach has an important strength that is well suited to the current investigation; namely, it does not reduce a category to a specific word or several words, but weighs a large collection of verbal features when assigning categories and hence is not ‘fooled’ by single words that are seemingly out of context. For example, the word ‘remember’ spans high and medium confidence hit categories. However, it is used in fundamentally different ways in each (i.e., occurs in conjunction with different sets of words), and the classifier was sensitive to these different uses.

Relevance for Decision Models of Recognition

Much of the debate regarding the basis or bases of recognition memory judgment has centered on the comparison of statistical decision models of recognition memory. This area received a renewed interest in part because of an influential paper by Donaldson (1996) that demonstrated that many Remember and Know dissociations could be easily accommodated within a strictly unidimensional strength framework that merely assumed that Remembering and Knowing simply reflected different extremities along a single strength dimension (see also Dunn, 2004) leading to claims that the distinction between remembering and knowing was epiphenomenal or artifactual. While this strictly unidimensional framework could accommodate many of the patterns post hoc it notably did not predict them in advance. Regardless, the current data demonstrate the limitation of the strictly unidimensional approach which does not anticipate current data.

Instead, the pattern of content findings is supportive of dual process models of recognition (for review see Yonelinas, 2002) and in particular, the SVM performance patterns were predicted by the DPSD decision model illustrated in Figure 2. However there are alternative dual process models and we next consider the Continuous Dual Process (CDP) model of Wixted & Mickes (2010) in light of the SVM and n-gram findings. The CDP model is a more complex decision model that eschews a threshold assumption for recollection and instead posits separate, statistically independent underlying recollection and familiarity channels that feed summed signals into a final recognition evidence value, which then determines recognition confidence (Figure 3). If this summed value exceeds the old/new

recognition criterion, observers will then go on to assess probes for evidence of recollection and then familiarity during paradigms that combine simple recognition confidence judgments with additional Remember/Know requirements. In these situations participants first assess the recollection channel to see if evidence along this dimension surpasses their criterion for remembering, and if so they report the item as remembered. If not, then the familiarity dimension is consulted and the participant reports the item as Known if its evidence along this dimension exceeds a second familiarity decision criterion. If the item exceeds neither criterion it is then deemed a guess (Figure 3). Under the model, all recognition probes, whether new or old, elicit both a continuous familiarity signal and a recollection signal. We next consider the CDP model in light of the SVM and n-gram analysis results.

In the case of the SVM, the DPSD model predicted high classifier performance when applied to high and medium confidence hits because recollective experiences were assumed to be confined to high recognition confidence, whereas medium confidence solely depends upon familiarity. Although classification was not perfect, the extremely high success rate of 93% suggested a near categorical distinction, which is quite a feat given that only two reports were sampled for each confidence category per subject. Furthermore, as noted above, the model also predicted how the classifier should fail when applied to potentially analogous correct rejection confidence justifications; namely, it should mistakenly rate these as generally medium confidence. These types of predictions do not naturally arise from the CDP model because it assumes that both recollection and familiarity are continuous signals spread liberally across recognition confidence options. To better illustrate this we used parameter values from a simulation in Wixted and Mickes (2010) that were deemed typical of the CDP, and generated 200 triplets of target evidence values (overall strength, familiarity strength, and recollection strength). In Figure 4 the recollection and familiarity strengths are denoted by the grey and black squares respectively and they are plotted in relation to the total summed recognition strength on the X-axis, which determines the confidence of the observer. Three evenly spaced, illustrative recognition criteria are used to demark recognition confidence.

It is clear from the figure that recollection and familiarity signals are each positively, moderately, and similarly correlated to final recognition confidence. This is captured by the 95% confidence ellipses for the 200 values drawn from each channel, which are almost completely overlapping. This of course stems from the fact that the overall recognition evidence is simply the sum of the two signals and the difference in slopes of the ellipses is a function of the recollection channel targets having higher variance than the familiarity channel targets. Critically, it is highly unlikely that one would start with this type of model representation and then consequently predict the patterns of SVM classifier sensitivity and specificity demonstrated in the results because neither recollection nor familiarity signals are uniquely indicative of any location along the X axis and hence any particular confidence level. The fact that the CDP does not predict the pattern of SVM performance in the current study is in fact not surprising since it was designed to advance the hypothesis that recollection is continuous and spread across all confidence options. Given that stance, to predict that a classification algorithm designed to make categorical content distinctions would succeed when applied to recognition hit confidence justifications would be strange.

Turning to the n-gram analysis, the CDP model instead fares conceptually better than the DPSD model because there are a minority of medium confidence reports in which modest recollection occurs, as indicated by the bigram 'vaguely remember' or the trigram 'think I remember' (Table 4). As noted earlier, the recollection mapping assumption of the DPSD model instead holds that all occurrences of recollection, even modest, should map onto the high confidence report option during simple recognition tasks. This is not because all recollections are assumed to contain the same amount of content, but because subjects are assumed to view any contextual recollection in this task as considerably more useful than familiarity evidence. Below are five instances of this phenomenon:

1. Circus is a somewhat strange word and I vaguely remember seeing it on the word list before because the word itself looks pretty.
2. Table is 2 syllables, but the word table spoon is a compound noun, so one might initially think that it was only 2 syllables, but because the first word has 2 syllables, it actually has 3 syllables in the word. I vaguely remember some process like this going on in my head, but cannot be sure.
3. I do not really remember sounding out this word in the first task, but it is possible that I did because I think I remember there was a word that had to do with birds or flight.
4. The word was familiar, and I think I remember seeing it, but I am not sure.
5. I have no idea why I think I remember this word. Maybe I was thinking about water towers or the clock tower or towering something or other, or maybe I was thinking of the towers test (neuropsychology). But, I am pretty sure I was thinking about something like that, so I am pretty sure it was on the list.

Some of these reports clearly would not be characterized as remembrances under the Tulving (1985) framework because the remaining content of the justification actually suggests a failure to retrieve episodic content reliably linked to the particular test probe considered. Nonetheless, it is clear from perusal of these and the remaining medium confidence reports that sometimes subjects will recover information about the prior study episode when encountering a given recognition probe and yet map the recognition response onto medium confidence. Thus these reports suggest that the recollection mapping assumption of the DPSD model can only approximately hold with subjects (or some subset of subjects) choosing to use the medium confidence option for modest recollective experiences. It is important to note that this does not invalidate the recollection threshold assumption, which merely holds that for some studied probes no conscious prior recollective content can be recovered.

Although the presence of modest recollection in the medium confidence hit category favors the CDP model over the DPSD model, the data do not provide evidence for the CDP assumption that recollection is completely continuous, which would require demonstrating that recollection is present in not only medium confidence hits, but also low confidence hits and misses as well. Further, as shown in Figure 4, one might expect recollective experiences should occasionally be quite vivid and strong even for recognition accompanied only by medium confidence. This can easily be seen by looking at the range of recollection signals

captured between the low and high recognition criteria in the figure, which are often higher than those in the adjacent high recognition confidence category. Unfortunately, because we only sampled two reports from each confidence bin, and did not sample errors, there is likely insufficient coverage to test these more nuanced predictions of the CDP model.

Overall, the SVM and n-gram data weigh heavily against a strictly unidimensional account of recognition and favor dual process interpretations. When comparing the CDP and DPSD models the data do not uniquely favor either. The DPSD model well anticipated the pattern of SVM classifier performance and was the motivating force behind actually applying the classifier and n-gram analyses to recognition content. In contrast, the CDP model does not anticipate the SVM performance and in fact the continuous and noisy nature of the recollection and familiarity evidence depicted in Figure 4 instead suggests the classifier should have generally struggled even if recollection and familiarity content differ. However, the CDP model does anticipate that recollective experiences should be present in medium confidence recognition reports and this was confirmed. This in turn means that the recollection mapping assumption of the DPSD, which was critical for making the SVM predictions, can only hold approximately.

The key differences between the two models revolve around the valuation of recollection on the part of the observer and the relation of recollection to overall recognition performance. Under the DPSD model, and other dual process characterizations (e.g., Jacoby, 1991; Mandler, 1980; Tulving, 1985), observers are assumed to place considerably more value on recollection than familiarity when judging memoranda as recognized. In contrast, the CDP model assumes that observers weight the two equally, which is reflected by the assumption that observers combine recollection and familiarity signals by simply summing them during basic recognition decisions. This means that psychologically, they do not favor one type of evidence over the other during the assignment of recognition confidence. One could perhaps relax this assumption by somehow weighting recollection more than familiarity during the summing step, but the implications of this are unclear and it would add another free parameter to the model.

Additionally, the two models differ considerably in the directness with which claims of remembering are linked to overall recognition performance. In the CDP model there is actually no direct connection between an observer's Remember rate and his or her overall recognition accuracy because the Remember rate is determined by a variable criterion within the recollection channel (Figure 3). In contrast, under the DPSD model the Remember rate directly tracks the recollection process and should be an extremely reliable indicator of overall recognition accuracy. These kinds of differences may provide fruitful avenues for future model comparison but are not testable via content analysis.

The Potential for Demand Characteristics

The current data demonstrate an association between old/new recognition confidence and the qualitative content of justifications used to support these decisions. This said, it is important to note that we cannot directly establish that the content provided by participants drives report confidence because we have no way of directly manipulating that content; a problem inherent in all psychological investigations that use subjective or self report data.

However, it is important to emphasize that our claims do not rest on intuitive examination of these reports, but on objective statistical analyses and classification algorithms. Regardless, to address the potential for demand characteristics to confound our current data it is important to be fairly specific about how they might operate.

First we consider a non-episodic heuristic in which observers merely give answers they believe are consistent with their provided confidence based on intuitive theories about memory functioning. In other words they provide justifications that are not drawn from episodic memory, but from general beliefs and knowledge about how they think memory content should map to recognition confidence. We view this interpretation as unlikely because of the complexity of the required heuristics and the specificity of the provided reports. For example, we doubt that many individuals entering the experiment are aware of the subjective memorability construct of Brown et al. (1977) and hence reject new items with high confidence first, and then go onto construct a story about how the personal distinctiveness of each particular item would have led to a remembrance had it been studied. For example, ‘I am a Christian so I was paying attention to all the religiously loaded words, and I do not recall seeing this one.’ Additionally, under a non-episodic demand characteristic explanation of our data, one would have to conclude that many of the provided justifications, including the one above, were potentially fabrications on the part of the subjects; a conclusion we find incongruous with their basic willingness to participate for minimal returns. Finally, many of the high confidence old justifications contained context information that was in fact objectively verifiable. For example, ‘The first word shown was monopoly. When I saw monotony I was reminded of monopoly, so I definitely remember both.’ This participant did in fact study Monopoly and Monotony in that order and appears to be illustrating an important memory principle that has garnered renewed interest, namely, the role of study list reminding in the facilitation of final recognition and recall (Jacoby & Wahlheim, 2013). When we informally examined all high confidence reports we found that such objectively verifiable content occurs most often in Experiment 2, and focusing on this experiment we observed that 10 of 11 responses that contained objectively verifiable content were in fact correct. The one incorrect response was only a minor error where the participant claimed a particular word appeared as the first item on the study list, when in fact it had appeared as the second item. Additionally, the study by McCabe et al. (2011) demonstrated that the majority of the justifications collected during recognition testing accurately matched the think aloud protocols collected during prior study. That is, very few responses were classified as incorrect recollections where the participant provided a completely different response justification during test than what was reported during study. Thus, at least in the case of high confidence recognition or subjective reports of remembering, participants often appear to recover verifiable contextual information.

None-the-less, asking participants to justify a memory report may be somewhat incongruous with many of our everyday experiences, since observers may often act reflexively during memory-based decisions without explicitly reflecting on various experiences of consciousness. Thus, intentionally asking participants to justify their recognition decisions may be somewhat difficult, especially in the case of familiarity based judgments where there is no specific episodic content available to report. In fact, in McCabe et al. 2011 participants were specifically instructed to report “the word seemed familiar” or it “rang a bell” for

responses that they could not recall any specific information. When episodic content is not available, reporting justifications may be particularly difficult and observers, although not intentionally, may rationalize in order to appear consistent with their current confidence judgment. Although we demonstrate that this is unlikely to be the case with high confidence reports it seems less easily ruled out for medium confidence reports. However, for these reports the most reliable content appeared to be intensity modifiers that simply reflected the perceived intensity of familiarity or novelty feelings.

Putting a non-episodic heuristic aside as a potential demand characteristic, we now turn to one with an episodic basis, termed differential search. Under the differential search account the differences in episodic content are genuine, and subjects do in fact recover recollections mostly for high confidence recognition. However, this results because subjects search more vigorously or for longer durations following high confidence old reports than medium or low confidence old reports because they feel the need to recover such content to justify the initial high confidence rating. This explanation does not help account for correct rejection content, but it could in principle result in recollections being heavily associated with high confidence recognition. Although this is an interesting idea, it seems unlikely because it assumes that observers could in fact provide similar levels of episodic detail for items recognized with medium or low confidence if they were just induced to search memory longer. However, there are prior recognition studies that demonstrate that low confidence recognition often accompanies chance recovery of source memory information (e.g., Ingram, Wixted, & Mickes, 2012; Slotnick & Dodson 2005) which strains the notion that participants could recover contextual information for these materials if they just searched longer. Additionally, in Gardiner et al. (1998) recognition testing occurred 24 hours after the initial encoding period. Critically, justifications were collected for a small random sample of test items after the entire recognition test had been completed and participants were simply asked to justify why they reported an item as ‘old’ at the testing phase. In this case, under a differential search account, one would have to posit that subjects remembered providing a prior ‘Remember’ earlier, and then that they engaged in a vigorous search of memory for the prior day’s encounter to try to substantiate having given a ‘Remember’ during the recognition test. Thus, our current paradigm rules out a demand characteristic that might arise from giving subjects detailed Remember/Know instructions, and the Gardiner et al. (1998) report delayed justifications until after the entire recognition test making differential search strategies less plausible. Given that the SVM developed in our study also reliably distinguished the Remember/Know justifications of Gardiner et al. (1998) at 88%, the total range and nature of the data lead us to view the demand characteristics outlined above as highly unlikely.

Instead, the current findings suggest reliable and important differences in the content that accompanies simple recognition confidence judgments. This content is consistent with dual process theories of recognition and therefore suggests that rigorous content analyses applied to recognition and other item-based memory tasks may provide fruitful data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported in part by a grant from the National Institute of Mental Health R01MH073982.

References

Bekkerman, R.; Allan, J. Technical report. University of Massachusetts; 2003. Using bigrams in text categorization. www.cs.umass.edu/~ronb/papers/bigrams.pdf

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological). 1995:289–300.

Brainard D. The psychophysics toolbox. Spatial Vision. 1997; 10(4):433–436.10.1163/156856897X00357 [PubMed: 9176952]

Brown J, Lewis V, Monk A. Memorability, Word-Frequency and Negative Recognition. Quarterly Journal of Experimental Psychology. 1977; 29:461–473.10.1080/14640747708400622

Dewhurst S, Farrand P. Investigating the phenomenological characteristics of false recognition for categorised words. European Journal of Cognitive Psychology. 2004; 16:403–416.10.1080/09541440340000088

Dobbins IG, Kroll NEA. Distinctiveness and the Recognition Mirror Effect: Evidence for an Item-Based Criterion Placement Heuristic. Journal of Experimental Psychology Learning, Memory, and Cognition. 2005; 31:1186–1198.10.1037/0278-7393.31.6.1186

Dobbins IG, Kroll NE, Liu Q. Confidence-accuracy inversions in scene recognition: a remember-know analysis. Journal of Experimental Psychology Learning, Memory, and Cognition. 1998; 24:1306–1315.10.1037//0278-7393.24.5.1306

Donaldson W. The role of decision processes in remembering and knowing. Memory & Cognition. 1996; 24:523–533.10.3758/BF03200940 [PubMed: 8757500]

Dunn JC. Remember-Know: A Matter of Confidence. Psychological Review. 2004; 111:524–542.10.1037/0033-295X.111.2.524 [PubMed: 15065921]

Gardiner JM, Gregg VH. Recognition memory with little or no remembering: Implications for a detection model. Psychonomic Bulletin & Review. 1997; 4:474–479.10.3758/BF03214336

Gardiner JM, Java RI. Recognition memory and awareness: An experiential approach. European Journal of Cognitive Psychology. 1993; 5:337–346.

Gardiner JM, Ramponi C, Richardson-Klavehn A. Experiences of remembering, knowing, and guessing. Consciousness and Cognition. 1998; 7:1–26.10.1006/ccog.1997.0321 [PubMed: 9521829]

Gardiner JM, Richardson-Klavehn A, Ramponi C. On Reporting Recollective Experiences and "Direct Access to Memory Systems. Psychological Science. 1997; 8:391–394.10.1111/j.1467-9280.1997.tb00431.x

Gardiner, JM.; Richardson-Klavehn, A. Remembering and Knowing. In: Tulving, E.; Craik, FIM., editors. The Oxford Handbook of Memory. New York: Oxford University Press; 2000. p. 229-244.

Ghetti S. Memory for nonoccurrences: The role of metacognition. Journal of Memory and Language. 2003; 48:722–739.10.1016/S0749-596X(03)00005-6

Hamel, LH. Knowledge Discovery with Support Vector Machines. Hoboken, NJ: John Wiley & Sons; 2009.

Ingram KM, Mickes L, Wixted JT. Recollection can be weak and familiarity can be strong. Journal of Experimental Psychology Learning, Memory, and Cognition. 2012; 38:325–339.10.1037/a0025483

Jacoby LL. A process dissociation framework: Separating automatic from intentional uses of memory. Journal of Memory and Language. 1991; 30:513–541.10.1016/0749-596X(91)90025-F

Jacoby LL, Wahlheim CN. On the importance of looking back: The role of recursive remindings in recency judgments and cued recall. Memory & cognition. 201310.3758/s13421-013-0298-5

Jaeger A, Cox J, Dobbins IG. Recognition confidence under violated and confirmed memory expectations. Journal of Experimental Psychology: General. 2012; 141:282–301.10.1037/a0025687 [PubMed: 21967231]

Jurka, TP.; Collingwood, L.; Boydstun, AE.; Grossman, E.; van Atteveldt, W. RTextTools: Automatic Text Classification via Supervised Learning. R package version 1.3.9. 2012. http://CRAN.R-project.org/package=RTextTools

Mandler G. Recognizing: The judgment of previous occurrence. Psychological Review. 1980; 87(3): 252.10.1037/0033-295X.87.3.252

MATLAB version 7.5.0. Natick, Massachusetts: The MathWorks Inc; 2007.

McCabe DP, Geraci L, Boman JK, Sensenig AE, Rhodes MG. On the validity of remember-know judgments: Evidence from think aloud protocols. Consciousness and Cognition. 2011; 20:1625–1633.10.1016/j.concog.2011.08.012 [PubMed: 21963257]

Norman KA. Differential effects of list strength on recollection and familiarity. Journal of Experimental Psychology Learning, Memory, and Cognition. 2002; 28(6):1083–1094.

Pelli D. The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision. 1997; 10:437–442.10.1163/156856897X00366\ [PubMed: 9176953]

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2012. URL http://www.R-project.org/

Rajaram S. Remembering and knowing: Two means of access to the personal past. Memory & Cognition. 1993; 21:89–102.10.3758/BF03211168 [PubMed: 8433652]

Rajaram S. Perceptual effects on remembering: recollective processes in picture recognition memory. Journal of Experimental Psychology Learning, Memory, and Cognition. 1996; 22:365–377.10.1037/0278-7393.22.2.365

Rajaram, S.; Roediger, HL. Remembering and knowing as states of consciousness during retrieval. In: Cohen, J.; Schooler, JW., editors. Scientific Approaches to Consciousness. Mahwah, New Jersey: Lawrence Erlbaum Associates Inc Hillsdale, NJ; 1997. p. 213-240.

Schacter DL, Kaszniak AW, Kihlstrom JF, Valdiserri M. The relation between source memory and aging. Psychology and Aging. 1991; 6:559–568.10.1037/0882-7974.6.4.559 [PubMed: 1777144]

Slotnick SD, Dodson CS. Support for a continuous (single-process) model of recognition memory and source memory. Memory & cognition. 2005; 33:151–170.10.3758/BF03195305 [PubMed: 15915801]

Stouffer, SA.; Suchman, EA.; DeVinney, LC.; Star, SA. The American soldier Vol 1: Adjustment during Army Life. Princeton: Princeton University Press; 1949.

Strong E. The effect of time-interval upon recognition memory. Psychological Review. 1913; 20:339–372.

Tulving E. Memory and consciousness. Canadian Psychology/Psychologie Canadienne. 1985; 26:1–12.

Wixted JT, Mickes L. A continuous dual-process model of remember/know judgments. Psychological Review. 2010; 117:1025–1054.10.1037/a0020874 [PubMed: 20836613]

Wixted JT, Stretch V. In defense of the signal detection interpretation of remember/know judgments. Psychonomic Bulletin & Review. 2004; 11:616–641.10.3758/BF03196616 [PubMed: 15581116]

Yonelinas AP. Receiver-operating characteristics in recognition memory: evidence for a dual-process model. Journal of Experimental Psychology Learning, Memory, and Cognition. 1994; 20:1341–1354.10.1037/0278-7393.20.6.1341

Yonelinas AP. The Nature of Recollection and Familiarity: A Review of 30 Years of Research. Journal of Memory and Language. 2002; 46:441–517.10.1006/jmla.2002.2864

Appendix. Rating Instructions

Instructions

You will be shown the responses of participants who completed a recognition memory test where they had to decide whether presented words were “old” (from the study list) or “new” (first appearance of the word in the experiment). For example, if you were asked whether the word “participants” was in the first sentence above (without looking back), you should

indicate the word is “old”. In contrast if you were shown “pickle” you would respond “new” because it was not mentioned earlier. Participants were also asked to indicate their confidence in this decision and to justify their decision with one or a few brief sentences.

Your task is to rate each of these justifications by indicating whether or not a certain characteristic is present. You will simply indicate whether the characteristic is present (yes) or absent (no) and indicate your confidence (high, medium, low) in the presence or absence of the particular characteristic.

Description of Characteristic

Personal experience outside of experiment

Response mentions a personal experience outside the experiment that they thought of when they encountered the word during the prior study list.

Examples

This word reminded me of our family vacations to the beach when I was a child.

I just went to the library and checked out one of my favorite books (Great Expectations), so I thought it was funny when this word popped up.

I have weird personal memories associated with tweed, like how it was really popular in junior high and how I owned this really hideous pink tweed blazer and wore it all the time when I was twelve or thirteen.

Description of Characteristic

Imagery, feelings, and thoughts

Response mentions specific imagery, feelings, or thoughts associated with the word from when they encountered it during the prior study list.

Examples

I remember thinking of a sunny day and feeling very happy when I saw the word “sun” before.

I pictured a gun when seeing this word, and since that is fairly easy to remember, I am positive it was in the previous list.

I remember seeing this word and thinking that sometimes it is also spelled with an “o” instead of an “a”.

Description of Characteristic

Notable absence of memory

Response mentions thoughts, feelings or images that should have been recalled if the word had been previously encountered, but which are currently clearly absent. Thus a key

characteristic of the response is that the participant focuses on the fact they did not remember seeing the word **and** they are confident they would have remembered thoughts, feelings, or images from a prior encounter if it had occurred.

Examples

I really like cars and I would have thought about the car I am currently working on if I saw “car” earlier.

I definitely would have pictured scaffolding had I saw the word. I am sure I did not, so it is new.

I really like limes and would have remembered if they were on the list.

Description of Characteristic

Strategies to memorize words

Response mentions specific strategies that the participant used in an attempt to memorize words.

Examples

I tried to create images for each word and I remember imagining a water bottle when I saw the word “water”.

I created a story to help me memorize the list of words and I know I saw “cat” before because I remember tying to together with the word “hat”.

I tried to group all the food words together and I remember adding “pineapple” to this category.

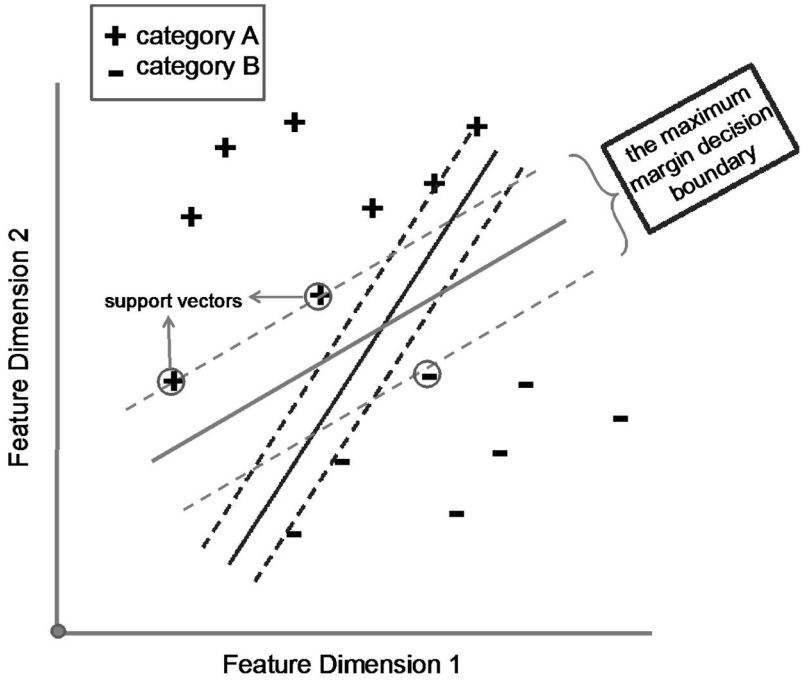

Figure 1.

Example of category classification using support vector machines. The figure depicts a simple 2D example of classification where two categories (A: plus symbols and B: minus symbols) are classified based on two continuous feature dimensions. The goal of the classifier is to determine a decision boundary that maximally separates the most confusing cases from the decision surface (maximum margin classification). The cases that touch the margin are termed support vectors. Note that the dark grey decision boundary results in a greater margin separating the support vectors than the light grey decision boundary.

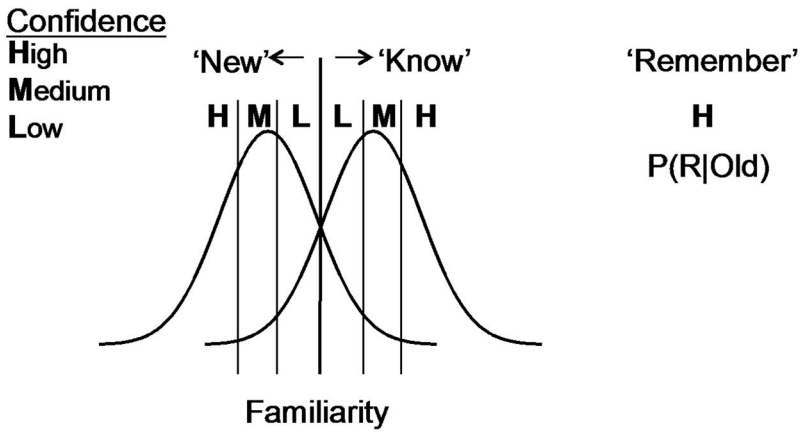

Figure 2.

Dual process signal detection model (DPSD). This model assumes that familiarity follows an equal variance signal detection model, where confidence judgments are based on a distance to criterion account (i.e., items further from the old/new decision criterion are reported with greater confidence). Recollection is a qualitatively distinct process that occurs when an old item's evidence surpasses an independent threshold and specific episodic context is recovered. Recollection is modeled by the probability of recollection, given an old item (p(R|Old), and this is generally assumed to lead to reports of high confidence.

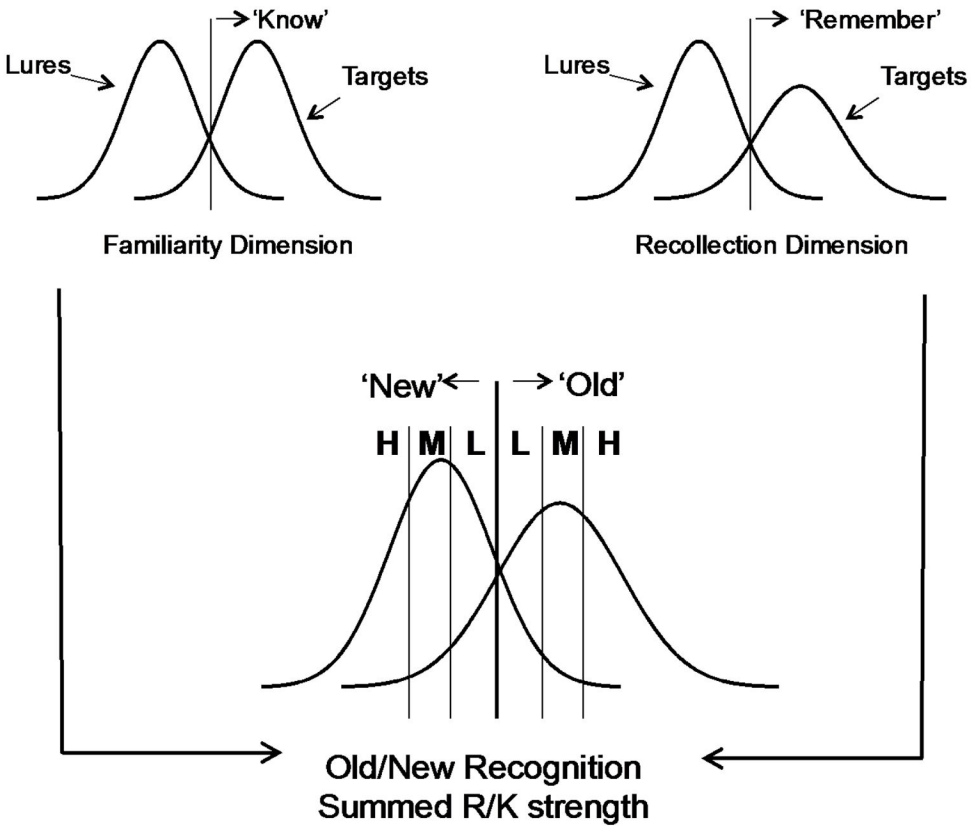

Figure 3.

Continuous dual process model (CDP). Under this model recollection and familiarity are separate, orthogonal dimensions where both processes follow a continuous signal detection model. Familiarity assumes an equal variance model whereas recollection assumes an unequal variance model. Observers make old/new recognition assessments by evaluating a hybrid signal, which is the sum of both independent processes, and determining whether this summed signal surpasses and old/new decision criterion. Confidence ratings follow a distance to criterion account using the summed recollection/familiarity signal. For items reported as old, observers assess for evidence of recollection and then for familiarity in order to determine if an item is Remembered vs. Known.

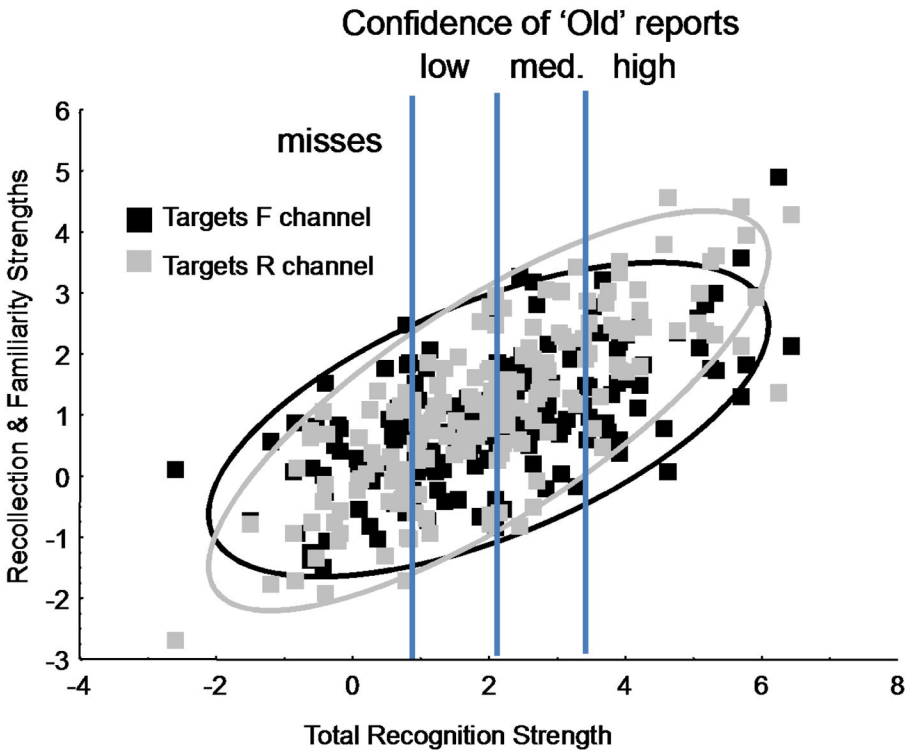

Figure 4.

Continuous Dual Process Model Mapping of Recollection and Familiarity to Confidence. A sample of 200 total target strengths were generated by summing random samples drawn from normal distributions representing a familiarity channel ($\mu = 0.8$ and $\sigma = 1$) and recollection channel ($\mu = 1.0$ and $\sigma = 1.4$). The separate Recollection (gray) and Familiarity (black) values are plotted against the y-axis and are then summed to represent the total recognition strength (x-axis). The vertical lines are illustrative confidence criterion and the ellipses are 95% confidence regions around the Recollection and Familiarity channel target values. Due to the highly overlapping Recollection and Familiarity values across low, medium, and high confidence old reports, this model is unlikely to give rise to our obtained Support Vector Machine results of high sensitivity and specificity for high vs. medium confidence hit classification.

Table 1

Number of Justifications Collected in Each Confidence Category

Experiment 1					
Hits			Correct Rejections		
High	Medium	Low	High	Medium	Low
52	52	42	51	52	45

Experiment 2					
Hits			Correct Rejections		
High	Medium	Low	High	Medium	Low
55	50	40	49	51	51

Table 2

Average Recognition Performance (Standard Deviations in Parenthesis)

Experiment 1		
Hit Rate	**False Alarm Rate**	**d′**
0.71 (0.09)	0.22 (0.08)	1.39 (0.38)

Experiment 2		
Hit Rate	**False Alarm Rate**	**d′**
0.72 (0.16)	0.23 (0.12)	1.49 (0.69)

Table 3

Average Proportion of Responses by Confidence Level

Experiment 1											
Hits			Misses			False Alarms			Correct Rejections		
High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low
0.50	0.13	0.08	0.05	0.12	0.11	0.04	0.09	0.10	0.25	0.30	0.23

Experiment 2											
Hits			Misses			False Alarms			Correct Rejections		
High	Medium	Low	High	Medium	Low	High	Medium	Low	High	Medium	Low
0.42	0.21	0.09	0.06	0.12	0.11	0.05	0.11	0.07	0.29	0.30	0.19

Table 4

Unigram, Bigram, Trigram Results for Hits

Unigram									
	Experiment 1			Experiment 2			Combined		
	High Confidence	Medium Confidence	z	High Confidence	Medium Confidence	z	Weighted z	p-value	Adj. p value
but	2	36	−5.52	3	31	−4.80	−7.31	0.000	0.000
not	3	43	−5.90	7	34	−4.22	−7.21	0.000	0.000
sure	2	28	−4.75	7	20	−2.50	−5.20	0.000	0.000
am	8	35	−4.12	12	31	−2.90	−4.96	0.000	0.000
medium	0	11	−3.32	0	7	−2.65	−4.22	0.000	0.000
pretty	2	10	−2.31	0	13	−3.61	−4.22	0.000	0.000
familiar	3	6	−1.00	0	15	−3.87	−3.84	0.000	0.002
is	8	28	−3.33	8	15	−1.46	−3.59	0.000	0.004
cannot	0	6	−2.45	1	9	−2.53	−3.43	0.001	0.007
looks	0	5	−2.24	0	6	−2.45	−3.31	0.001	0.010
vaguely	0	7	−2.65	0	4	−2.00	−3.29	0.001	0.010
if	1	13	−3.21	2	4	−0.82	−3.27	0.001	0.010
have	8	18	−1.96	4	16	−2.68	−3.19	0.001	0.012
be	0	8	−2.83	3	8	−1.51	−2.88	0.004	0.031
think	4	17	−2.84	6	10	−1.00	−2.86	0.004	0.031
a	16	35	−2.66	23	33	−1.34	−2.78	0.005	0.038
there	2	9	−2.11	2	7	−1.67	−2.69	0.007	0.047
been	0	2	−1.41	0	5	−2.24	−2.60	0.009	0.054
or	4	9	−1.39	2	10	−2.31	−2.59	0.010	0.054
completely	0	3	−1.73	0	3	−1.73	−2.45	0.014	0.070
seems	0	2	−1.41	0	4	−2.00	−2.42	0.015	0.070
did	0	3	−1.73	1	5	−1.63	−2.24	0.025	0.110
I	106	145	−2.46	120	129	−0.57	−2.15	0.032	0.132
that	15	26	−1.72	22	31	−1.24	−2.03	0.042	0.160
like	4	11	−1.81	3	5	−0.71	−1.93	0.054	0.173
do	2	4	−0.82	3	9	−1.73	−1.91	0.056	0.173

Unigram

	Experiment 1			Experiment 2			Combined		
	High Confidence	Medium Confidence	z	High Confidence	Medium Confidence	z	Weighted z	p-value	Adj. p value
positive	3	6	−1.00	2	7	−1.67	−1.89	0.059	0.178
head	8	3	1.51	3	0	1.73	1.91	0.056	0.173
being	6	2	1.41	4	1	1.34	1.91	0.056	0.173
came	4	0	2.00	2	1	0.58	1.95	0.052	0.173
saying	2	1	0.58	4	0	2.00	1.95	0.052	0.173
remember	44	37	0.78	49	31	2.01	1.97	0.049	0.173
distinctly	5	0	2.24	1	1	0.00	2.08	0.038	0.148
myself	4	1	1.34	7	2	1.67	2.11	0.035	0.141
after	2	0	1.41	4	0	2.00	2.42	0.015	0.070
first	5	4	0.33	11	1	2.89	2.51	0.012	0.062
in	26	15	1.72	22	11	1.91	2.54	0.011	0.059
stuck	3	0	1.73	4	0	2.00	2.64	0.008	0.051
when	9	2	2.11	15	2	3.15	3.79	0.000	0.002

Bigrams

	Experiment 1			Experiment 2			Combined		
	High Confidence	Medium Confidence	z	High Confidence	Medium Confidence	z	Weighted z	p-value	Adj. p value
but I	1	21	−4.26	1	16	−3.64	−5.60	0.000	0.000
am not	0	22	−4.69	2	11	−2.50	−5.31	0.000	0.000
I am	8	32	−3.79	12	29	−2.65	−4.55	0.000	0.001
word but	1	13	−3.21	0	5	−2.24	−3.77	0.000	0.006
pretty sure	1	8	−2.33	0	9	−3.00	−3.77	0.000	0.006
think I	2	13	−2.84	0	6	−2.45	−3.55	0.000	0.012
am pretty	2	6	−1.41	0	10	−3.16	−3.35	0.001	0.020
medium confidence	0	7	−2.65	0	4	−2.00	−3.29	0.001	0.021
I think	3	15	−2.83	2	7	−1.67	−3.28	0.001	0.021
I cannot	0	4	−2.00	1	9	−2.53	−3.09	0.002	0.036
not positive	0	4	−2.00	0	5	−2.24	−3.00	0.003	0.045
it is	1	12	−3.05	3	4	−0.38	−2.87	0.004	0.053
a medium	0	6	−2.45	0	2	−1.41	−2.77	0.006	0.063

Bigrams									
	Experiment 1			Experiment 2			Combined		
	High Confidence	Medium Confidence	z	High Confidence	Medium Confidence	z	Weighted z	p-value	Adj. p value
vaguely remember	0	6	−2.45	0	2	−1.41	−2.77	0.006	0.063
looks familiar	0	4	−2.00	0	3	−1.73	−2.64	0.008	0.087
if I	1	8	−2.33	0	2	−1.41	−2.58	0.010	0.092
cannot be	0	3	−1.73	0	3	−1.73	−2.45	0.014	0.106
not completely	0	3	−1.73	0	3	−1.73	−2.45	0.014	0.106
have been	0	2	−1.41	0	4	−2.00	−2.42	0.015	0.106
it looks	0	2	−1.41	0	4	−2.00	−2.42	0.015	0.106
I vaguely	0	4	−2.00	0	2	−1.41	−2.42	0.015	0.106
sure I	0	6	−2.45	3	7	−1.26	−2.34	0.019	0.126
not sure	0	7	−2.65	2	2	0.00	−2.30	0.022	0.138
familiar but	0	3	−1.73	0	2	−1.41	−2.23	0.026	0.149
this one	0	3	−1.73	0	2	−1.41	−2.23	0.026	0.149
but it	1	2	−0.58	0	5	−2.24	−2.21	0.027	0.149
I might	0	1	−1.00	0	4	−2.00	−2.18	0.029	0.149
sure this	0	1	−1.00	0	4	−2.00	−2.18	0.029	0.149
word is	0	4	−2.00	0	1	−1.00	−2.18	0.029	0.149
because it	1	6	−1.89	1	3	−1.00	−2.14	0.033	0.159
do not	1	3	−1.00	2	8	−1.90	−2.13	0.033	0.159
that I	4	9	−1.39	6	13	−1.61	−2.11	0.035	0.165
did not	0	3	−1.73	1	4	−1.34	−2.04	0.041	0.179
I have	6	13	−1.61	2	6	−1.41	−2.03	0.042	0.179
completely sure	0	2	−1.41	0	2	−1.41	−2.00	0.046	0.179
similar to	0	2	−1.41	0	2	−1.41	−2.00	0.046	0.179
familiar it	0	1	−1.00	0	3	−1.73	−1.96	0.050	0.179
totally sure	0	3	−1.73	0	1	−1.00	−1.96	0.050	0.179
I do	2	3	−0.45	2	8	−1.90	−1.90	0.058	0.195
my head	8	3	1.51	3	0	1.73	1.91	0.056	0.193
one of	4	0	2.00	7	3	1.26	1.92	0.055	0.193
high confidence	3	0	1.73	1	0	1.00	1.96	0.050	0.179
thought about	3	0	1.73	1	0	1.00	1.96	0.050	0.179

Bigrams

	Experiment 1			Experiment 2			Combined		
	High Confidence	Medium Confidence	z	High Confidence	Medium Confidence	z	Weighted z	p-value	Adj. p value
to me	3	0	1.73	1	0	1.00	1.96	0.050	0.179
it stuck	2	0	1.41	2	0	1.41	2.00	0.046	0.179
when this	2	0	1.41	2	0	1.41	2.00	0.046	0.179
list I	1	1	0.00	9	2	2.11	2.08	0.038	0.174
in my	9	4	1.39	6	0	2.45	2.29	0.022	0.138
remember this	6	0	2.45	2	1	0.58	2.45	0.014	0.106
I thought	5	0	2.24	4	1	1.34	2.53	0.011	0.102
to myself	3	1	1.00	6	0	2.45	2.59	0.010	0.092
when I	6	2	1.41	11	2	2.50	2.87	0.004	0.053
I remember	31	21	1.39	34	15	2.71	2.87	0.004	0.053

Trigrams

	Experiment 1			Experiment 2			Combined		
	High Confidence	Medium Confidence	z	High Confidence	Medium Confidence	z	Weighted z	p-value	Adj. p value
I am not	0	20	−4.47	2	11	−2.50	−5.11	0.000	0.000
but I am	0	15	−3.87	1	7	−2.12	−4.42	0.000	0.001
think I remember	0	9	−3.00	0	5	−2.24	−3.71	0.000	0.010
word but I	0	11	−3.32	0	3	−1.73	−3.66	0.000	0.010
I think I	2	13	−2.84	0	6	−2.45	−3.55	0.000	0.011
am pretty sure	1	6	−1.89	0	9	−3.00	−3.53	0.000	0.011
I am pretty	2	6	−1.41	0	10	−3.16	−3.35	0.001	0.017
this word but	0	6	−2.45	0	4	−2.00	−3.15	0.002	0.031
a medium confidence	0	6	−2.45	0	2	−1.41	−2.77	0.006	0.084
medium confidence level	0	6	−2.45	0	2	−1.41	−2.77	0.006	0.084
am not positive	0	3	−1.73	0	3	−1.73	−2.45	0.014	0.149
I vaguely remember	0	4	−2.00	0	2	−1.41	−2.42	0.015	0.149
because it is	0	5	−2.24	0	1	−1.00	−2.39	0.017	0.149
chose a medium	0	5	−2.24	0	1	−1.00	−2.39	0.017	0.149
am not completely	0	2	−1.41	0	3	−1.73	−2.23	0.026	0.217
am not sure	0	6	−2.45	2	2	0.00	−2.04	0.042	0.222
not completely sure	0	2	−1.41	0	2	−1.41	−2.00	0.046	0.222

Trigrams									
	Experiment 1			Experiment 2			Combined		
	High Confidence	Medium Confidence	z	High Confidence	Medium Confidence	z	Weighted z	p-value	Adj. p value
I do not	1	2	−0.58	2	8	−1.90	−1.98	0.047	0.222
but I do	0	1	−1.00	0	3	−1.73	−1.96	0.050	0.222
I cannot be	0	1	−1.00	0	3	−1.73	−1.96	0.050	0.222
pretty sure that	0	3	−1.73	0	1	−1.00	−1.96	0.050	0.222
part of the	1	2	−0.58	0	4	−2.00	−1.95	0.052	0.222
one of the	2	0	1.41	7	2	1.67	1.93	0.053	0.222
remember this word	4	0	2.00	2	1	0.58	1.95	0.052	0.222
high confidence level	3	0	1.73	1	0	1.00	1.96	0.050	0.222
I thought about	3	0	1.73	1	0	1.00	1.96	0.050	0.222
I remember saying	2	0	1.41	2	0	1.41	2.00	0.046	0.222
so it stuck	2	0	1.41	2	0	1.41	2.00	0.046	0.222
when this word	2	0	1.41	2	0	1.41	2.00	0.046	0.222
when I saw	3	2	0.45	6	0	2.45	2.17	0.030	0.222
I remember this	6	0	2.45	2	1	0.58	2.45	0.014	0.149
word I remember	5	1	1.63	4	0	2.00	2.47	0.014	0.149
I remember it	3	0	1.73	4	0	2.00	2.64	0.008	0.113

Table 5

Unigram, Bigram, Trigram Results for Correct Rejections

Unigram									
	Experiment 1			Experiment 2			Combined		
	High Confidence	Medium Confidence	z	High Confidence	Medium Confidence	z	Weighted z	p-value	Adj. p value
but	7	28	−3.55	7	29	−3.67	−5.10	0.000	0.000
medium	0	5	−2.24	0	8	−2.83	−3.58	0.000	0.021
so	6	10	−1.00	10	27	−2.79	−2.96	0.003	0.079
not	65	80	−1.25	45	79	−3.05	−2.93	0.003	0.079
the	35	68	−3.25	46	54	−0.80	−2.89	0.004	0.079
sure	10	19	−1.67	7	18	−2.20	−2.70	0.007	0.121
cannot	2	11	−2.50	1	2	−0.58	−2.56	0.010	0.150
familiar	2	7	−1.67	0	5	−2.24	−2.54	0.011	0.150
I	126	155	−1.73	131	160	−1.70	−2.42	0.015	0.188
possible	0	5	−2.24	1	3	−1.00	−2.37	0.018	0.198
it	44	47	−0.31	38	63	−2.49	−2.06	0.040	0.324
many	0	1	−1.00	1	6	−1.89	−2.01	0.044	0.340
just	6	7	−0.28	2	11	−2.50	−1.96	0.050	0.342
think	11	11	0.00	6	19	−2.60	−1.95	0.051	0.342
some	1	5	−1.63	2	5	−1.13	−1.92	0.054	0.342
am	19	28	−1.31	17	26	−1.37	−1.90	0.058	0.342
all	11	5	1.50	7	2	1.67	2.12	0.034	0.295
definitely	6	0	2.45	3	2	0.45	2.17	0.030	0.285
have	28	22	0.85	43	25	2.18	2.26	0.024	0.243
remembered	5	1	1.63	11	2	2.50	2.95	0.003	0.079

Bigrams									
	Experiment 1			Experiment 2			Combined		
	High Confidence	Medium Confidence	z	High Confidence	Medium Confidence	z	Weighted z	p-value	Adj. p value
am not	1	12	−3.05	1	10	−2.71	−4.08	0.000	0.009
not sure	0	6	−2.45	1	9	−2.53	−3.43	0.001	0.065
word but	0	9	−3.00	2	6	−1.41	−3.18	0.001	0.070

Bigrams									
	Experiment 1			Experiment 2			Combined		
	High Confidence	Medium Confidence	z	High Confidence	Medium Confidence	z	Weighted z	p-value	Adj. p value
but I	3	12	−2.32	4	13	−2.18	−3.17	0.002	0.070
it was	5	10	−1.29	3	15	−2.83	−3.00	0.003	0.100
sure if	0	6	−2.45	0	3	−1.73	−2.97	0.003	0.100
but it	0	4	−2.00	0	4	−2.00	−2.83	0.005	0.120
medium confidence	0	4	−2.00	0	4	−2.00	−2.83	0.005	0.120
or not	0	4	−2.00	0	3	−1.73	−2.64	0.008	0.175
there is	0	3	−1.73	0	3	−1.73	−2.45	0.014	0.224
but there	0	2	−1.41	0	4	−2.00	−2.42	0.015	0.224
just do	0	2	−1.41	0	4	−2.00	−2.42	0.015	0.224
possible that	0	4	−2.00	0	2	−1.41	−2.42	0.015	0.224
so I	4	8	−1.15	8	18	−1.96	−2.26	0.024	0.275
is why	0	2	−1.41	0	3	−1.73	−2.23	0.026	0.275
as a	0	3	−1.73	0	2	−1.41	−2.23	0.026	0.275
before but	0	3	−1.73	0	2	−1.41	−2.23	0.026	0.275
the meaning	0	3	−1.73	0	2	−1.41	−2.23	0.026	0.275
is possible	0	4	−2.00	0	1	−1.00	−2.18	0.029	0.290
if I	2	9	−2.11	5	9	−1.07	−2.14	0.032	0.290
I just	1	3	−1.00	1	6	−1.89	−2.14	0.033	0.290
however I	1	4	−1.34	0	3	−1.73	−2.04	0.041	0.290
chance I	0	2	−1.41	0	2	−1.41	−2.00	0.046	0.290
not know	0	2	−1.41	0	2	−1.41	−2.00	0.046	0.290
not positive	0	2	−1.41	0	2	−1.41	−2.00	0.046	0.290
could have	0	2	−1.41	1	5	−1.63	−2.00	0.046	0.290
I cannot	2	8	−1.90	1	2	−0.58	−1.98	0.047	0.290
feel that	0	1	−1.00	0	3	−1.73	−1.96	0.050	0.290
it could	0	1	−1.00	0	3	−1.73	−1.96	0.050	0.290
number of	0	1	−1.00	0	3	−1.73	−1.96	0.050	0.290
it or	0	3	−1.73	0	1	−1.00	−1.96	0.050	0.290
am confident	1	0	1.00	3	0	1.73	1.96	0.050	0.290
and it	1	0	1.00	3	0	1.73	1.96	0.050	0.290

Bigrams

	Experiment 1			Experiment 2			Combined		
	High Confidence	Medium Confidence	z	High Confidence	Medium Confidence	z	Weighted z	p-value	Adj. p value
have no	0	1	−1.00	9	2	2.11	2.01	0.044	0.290
at all	8	3	1.51	5	1	1.63	2.11	0.035	0.290
and I	9	5	1.07	13	5	1.89	2.14	0.032	0.290
I definitely	6	0	2.45	0	1	−1.00	2.25	0.024	0.275
all I	4	0	2.00	2	0	1.41	2.42	0.015	0.224
have remembered	5	1	1.63	10	2	2.31	2.80	0.005	0.120
would have	14	3	2.67	24	10	2.40	3.34	0.001	0.065

Trigrams

	Experiment 1			Experiment 2			Combined		
	High Confidence	Medium Confidence	z	High Confidence	Medium Confidence	z	Weighted z	p-value	Adj. p value
I am not	1	11	−2.89	1	10	−2.71	−3.96	0.000	0.016
this word but	0	7	−2.65	0	2	−1.41	−2.93	0.003	0.234
am not sure	0	4	−2.00	1	7	−2.12	−2.79	0.005	0.234
not sure if	0	4	−2.00	0	3	−1.73	−2.64	0.008	0.265
but I am	0	7	−2.65	1	1	0.00	−2.54	0.011	0.287
there is a	0	3	−1.73	0	3	−1.73	−2.45	0.014	0.287
just do not	0	2	−1.41	0	4	−2.00	−2.42	0.015	0.287
medium confidence level	0	4	−2.00	0	2	−1.41	−2.42	0.015	0.287
not think I	1	0	1.00	0	6	−2.45	−2.25	0.024	0.357
that it was	0	3	−1.73	1	5	−1.63	−2.24	0.025	0.357
word but it	0	2	−1.41	0	3	−1.73	−2.23	0.026	0.357
it is possible	0	4	−2.00	0	1	−1.00	−2.18	0.029	0.357
sure if I	0	4	−2.00	0	1	−1.00	−2.18	0.029	0.357
think I saw	1	1	0.00	0	5	−2.24	−2.08	0.038	0.357
is why I	0	2	−1.41	0	2	−1.41	−2.00	0.046	0.357
I just do	0	1	−1.00	0	3	−1.73	−1.96	0.050	0.357
the list but	0	1	−1.00	0	3	−1.73	−1.96	0.050	0.357
the word but	0	1	−1.00	0	3	−1.73	−1.96	0.050	0.357
remember this word	2	6	−1.41	0	3	−1.73	−1.93	0.053	0.357
and I think	3	0	1.73	1	0	1.00	1.96	0.050	0.357

Trigrams									
	Experiment 1			Experiment 2			Combined		
	High Confidence	Medium Confidence	z	High Confidence	Medium Confidence	z	Weighted z	p-value	Adj. p value
I am confident	1	0	1.00	3	0	1.73	1.96	0.050	0.357
I have no	0	1	−1.00	9	2	2.11	2.01	0.044	0.357
it would have	3	0	1.73	2	0	1.41	2.23	0.026	0.357
at all I	4	0	2.00	2	0	1.41	2.42	0.015	0.287
word I am	4	0	2.00	3	0	1.73	2.64	0.008	0.265
would have remembered	5	1	1.63	10	2	2.31	2.80	0.005	0.234
I would have	7	2	1.67	19	6	2.60	3.01	0.003	0.234

Table 6

Average Category Ratings (Standard Deviations in Parenthesis)

	Hits		Correct Rejections	
Category	**High Confidence**	**Medium Confidence**	**High Confidence**	**Medium Confidence**
1.) Personal Experience Outside of Experiment	0.38 (0.44)	0.06 (0.21)	0.27 (0.55)	0.09 (0.28)
2.) Imagery, Feelings, and Thoughts	1.32 (0.70)	0.76 (0.79)	0.90 (0.77)	0.72 (0.66)
3.) Notable Absence of Memory	0.11 (0.25)	0.15 (0.30)	1.35 (0.67)	1.04 (0.62)
4.) Strategies to Memorize Words	1.38 (0.70)	0.76 (0.66)	0.90 (0.81)	0.65 (0.73)

Table 7

Classifier Performance Applied to Experiment 2 (Trained on Hits - Experiment 1)

	Hits Classifier Judgment			
Origin	**Medium**	**High**	**Count**	**% Corr**
Medium	26	1		
High	3	25	55	93
	Correct Rejections Classifier Judgment			
Origin	Medium	High	Count	% Corr
Medium	23	4		
High	18	8	53	59
	Hits Classifier Judgment			
Origin	Medium	Low	Count	% Corr
Medium	22	5	50	68
Low	11	12		
	Correct Rejections Classifier Judgment			
Origin	Medium	Low	Count	% Corr
Medium	17	10	54	63
Low	10	17		

Table 8

Classifier Feature Weights, High vs. Medium Conf. Hits (Trained on Exp 1)

High Confidence Weight	Word	Medium Confidence Weight	Word
0.22	being	−0.38	but
0.21	remember	−0.36	not
0.16	previously	−0.23	sure
0.16	so	−0.22	think
0.14	previous	−0.16	a
0.14	was	−0.16	medium
0.14	two	−0.14	have
0.13	thought	−0.14	vaguely
0.12	with	−0.13	if
0.11	this	−0.12	looks
0.11	in	−0.11	me
0.1	seeing	−0.11	just
0.1	phase	−0.1	am
0.1	I	−0.1	time
0.1	pronounce	−0.1	line
0.1	hammock	−0.1	movie
0.1	here	−0.1	reminded
0.1	number	−0.1	least
0.1	an	−0.1	long
0.09	counting	−0.1	must