Journal of Experimental Psychology: Learning, Memory, and Cognition

Mnemonic Discrimination Language Evinces Recollection Rejection of Similar Lures

Christopher N. Wahlheim, Ian G. Dobbins, and Bayley M. Wellons

Online First Publication, November 13, 2025. https://dx.doi.org/10.1037/xlm0001556

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology:
Learning, Memory, and Cognition

ISSN: 0278-7393

https://doi.org/10.1037/xlm0001556

Mnemonic Discrimination Language Evinces Recollection Rejection of Similar Lures

Christopher N. Wahlheim[1], Ian G. Dobbins[2], and Bayley M. Wellons[1]
[1] Department of Psychology, University of North Carolina at Greensboro
[2] Department of Psychological and Brain Sciences, Washington University in St. Louis

Mnemonic discrimination of visual objects entails differentiating among repetitions of target objects, unstudied foil objects, and, critically, lure objects that are similar, but not identical to, studied objects (e.g., a different coffee mug than what was studied). Correctly rejecting lures may involve hippocampal pattern separation, a process that orthogonalizes representations of similar experiences. However, lures can also be rejected when recollection of studied objects enables detection of changed lure features. The present study examined whether verbal justifications of recognition decisions in an object-based mnemonic discrimination task could reveal recollection rejection as the primary basis for lure rejections. Across multiple study–test trials, participants studied everyday objects and, at test, attempted to classify similar lures, studied targets, and novel foils. Participants sometimes verbally justified their decisions. Machine learning classifiers showed that verbal justifications discriminated among different classifications given to the same item types for both in- and out-of-sample data. Lure rejection language often expressed the use of recollection of studied objects to detect changes in perceived objects (viz., recollection-rejection strategy). Verbal justifications also discriminated correct from incorrect responses better than numeric confidence, which could not be explained by a model assuming a one-dimensional memory strength signal. Finally, verbal justifications best predicted accurate recognition decisions for all item types at the highest level of subjective confidence, which further implicated the use of recollection. The present findings verify that lure rejections in mnemonic discrimination tasks do not only reflect hippocampal pattern separation but also suggest that rejections reflect recollection of studied targets.

Keywords: language, machine learning, mnemonic discrimination, recognition memory, recollection rejection

Supplemental materials: https://doi.org/10.1037/xlm0001556.supp

Mnemonic discrimination of visual objects entails identifying when a currently perceived object is similar but not identical to an earlier-encoded object. For example, a recognizer may conclude that a coffee mug on the counter is not another mug they recently used because the previous mug had a nicer handle. Mnemonic discrimination can be studied by showing participants pictures of objects, then later instructing them to reject lures that are similar to those objects (Stark et al., 2013). Such tasks have been used to assess hippocampal pattern separation, a process that mitigates interference by encoding distinctive representations of similar inputs (Marr, 1971). Pattern separation has been inferred from the ability to reject similar lures and related hippocampal activation (for a review, see Stark et al., 2019). However, lure rejections may also involve hippocampal pattern completion, a process that reinstates representations and enables detection of changed features via comparison (Norman & O'Reilly, 2003). The present study examined whether the language used to justify decisions in a mnemonic discrimination task shows how memory and perception are compared during successful versus failed mnemonic discrimination. This is relevant for understanding the processes underlying lure rejections and the utility of language classifiers for identifying the qualitative bases for recognition decisions.

The Processes Underlying Similar Lure Rejections

Routines lead people to encode similar events and objects. One challenge for the memory system is to encode distinctive representations to prevent memory errors. Mnemonic discrimination tasks are used to characterize the processes supporting this

Jeffrey Starns served as action editor.

Christopher N. Wahlheim https://orcid.org/0000-0002-2381-1493

The materials, data, and analysis scripts are available on the Open Science Framework (https://osf.io/g5t9a/). Christopher N. Wahlheim and Ian G. Dobbins contributed equally to this work, which was supported by the National Science Foundation, Grant 2317124, awarded to Ian G. Dobbins. The authors thank Paige Kemp for assisting with experiment programming and data collection.

Christopher N. Wahlheim played a lead role in project administration, resources, and supervision and an equal role in conceptualization, data curation, formal analysis, investigation, methodology, visualization, writing–original draft, and writing–review and editing. Ian G. Dobbins played an equal role in conceptualization, data curation, formal analysis, investigation, methodology, visualization, writing–original draft, and writing–review and editing. Bayley M. Wellons played a supporting role in data curation, methodology, and writing–review and editing.

Correspondence concerning this article should be addressed to Christopher N. Wahlheim, Department of Psychology, University of North Carolina at Greensboro, 296 Eberhart Building, P.O. Box 26170, Greensboro, NC 27402-6170, United States. Email: cnwahlhe@uncg.edu

aspect of memory, emphasizing how hippocampal subfields and cortical regions projecting to the hippocampus cooperate to resolve interference (for a review, Amer & Davachi, 2023). Studies of hippocampal computations have focused on pattern separation in the dentate gyrus subfield, which assigns distinct neural codes to similar experiences, and pattern completion in the CA3, which retrieves representations from partial inputs (Marr, 1971; Norman & O'Reilly, 2003; O'Reilly & McClelland, 1994; Treves & Rolls, 1994). However, it is challenging to isolate these processes during recognition and identify their neural correlates because lures that are similar to studied items have unique features that may evoke pattern separation while also having shared features that may evoke pattern completion (for reviews, see Hunsaker & Kesner, 2013; Liu et al., 2016).

This complexity in assessing these processes is illustrated by studies using the Mnemonic Similarity Task (MST; for a review, see Stark et al., 2019). In study–test MSTs, participants study everyday objects, then, in the next phase, make recognition decisions about repetitions of studied objects, changed versions of studied objects (referred to as lures), and new objects (referred to as foils). Such studies have assumed that rejections of lures manifest pattern separation, and this is supported by converging evidence. For example, populations with hippocampal deficiencies, such as older adults, show disproportionate deficits in lure rejections (e.g., Stark et al., 2013) and DG/CA3 hyperactivity during lure rejections (e.g., Reagh et al., 2018), suggesting a bias toward pattern completion in the DG (Wilson et al., 2006). However, the assertion that lure rejections reflect pattern separation is complicated because such decisions are process impure. Because lures have the same identities as studied objects, the shared features can cue the retrieval of studied objects that enable the detection of changed features (e.g., Norman & O'Reilly, 2003).

This retrieval monitoring approach, referred to as *recall-to-reject* or *recollection rejection*, is antithetical to the assumption that lure rejections solely reflect distinctive encoding supported by pattern separation. Indeed, studies have long assumed that participants use a recollection-based strategy during mnemonic discrimination (e.g., Hintzman & Curran, 1994). A recollection rejection mechanism has been proposed as a way to avoid false memories of related lures (Brainerd et al., 2003; Gallo, 2010). It has been implicated with computational modeling (Rotello et al., 2000) and by showing that manipulations that impair recollection, such as speeded responding and divided attention, reduce lure rejections (Jones, 2005; Odegard & Lampinen, 2005; Odegard et al., 2008). The use of this strategy has also been measured more directly by requiring participants to indicate if lure rejections were based on recollection of studied objects, which implies a detailed memory (Bowman & Dennis, 2016; Kim & Yassa, 2013; Szőllősi et al., 2020). Those studies suggest that participants used memories of studied objects of various qualities as a basis for lure rejections. In a related study, participants based their lure rejections on self-reported recollection of studied objects more often for objects that they reported attending to during study (Wahlheim et al., 2024). Participants made these reports by indicating whether they were on or off task when intermittent thought probes appeared after each of a subset of studied objects. These findings suggest that lure rejections can be based on a monitoring strategy that requires successful prior encoding of studied objects to detect changes.

Language Classifiers and Recognition Decisions

Although the aforementioned studies provide converging evidence for retrieval-based lure rejections, each approach has limitations. Indirect approaches require researchers to infer strategies using manipulations and modeling without asking participants about the experiences informing their decisions, which leaves room for alternative interpretations. Conversely, in the self-report methods used thus far, the lure rejection strategy is described to participants in the task instructions, and participants are given categorical response options for the strategies that the researcher believes are germane. However, these procedural details may alter participants' decision strategies and create demand characteristics that evoke reports of retrieval monitoring. An approach that partially circumvents these limitations entails participants introspecting on recognition decisions without being told about strategies that they could have used or retrieval experiences that are critical for successful performance (Dobbins & Kantner, 2019). In what follows, we summarize findings showing that this approach can capture qualitatively different bases for recognition decisions, suggesting that strategy use and qualitative bases for decisions in mnemonic discrimination tasks can also be characterized using this method.

The natural language approach that we advance here was initially applied to recognition memory of words to determine if recognizers use language implying that recollection and familiarity were used as distinctive bases for endorsing test items (Selmeczy & Dobbins, 2014). These retrieval processes have been measured by giving participants detailed descriptions about differences between subjective states according to Tulving's (1985) conception of remembering and knowing and then providing participants with categorical response options that map onto those states after old recognition responses (for an overview, see Umanath & Coane, 2020). One study using this procedure with verbal justifications of remember, know, and guess responses showed that the mnemonic content of justifications matched subjective states that dual-process theories assume accompany recollection- and familiarity-based decisions (Gardiner et al., 1998). However, telling participants about those states and providing categorical response options may have created demand characteristics that influenced the reports. To circumvent this issue, Selmeczy and Dobbins (2014) designed a recognition protocol in which verbal justifications and confidence ratings (high, medium, low) were collected, and participants were not told about subjective states. Word sequences extracted from justifications showed that language implying recollection was most associated with highly confident recognition of studied items.

To ensure that the verbal justifications were not distorted by requiring them after making confidence ratings, another study collected justifications in a recognition protocol with only old/new judgments before justifications (Dobbins & Kantner, 2019). A machine learner was still able to distinguish accurate (hits) from inaccurate (false alarms) decisions using language. The features selected by the machine learner suggested it relied on the distinction between recollective and nonrecollective language to make its predictions. This afforded tests of convergent and divergent validity in which this trained classifier scored the justifications from the other studies that either should or should not differ based on successful recollection (Gardiner et al., 1998; Selmeczy & Dobbins, 2014). Those tests showed convergent validity as the trained classifier predicted justifications from out-of-sample data sets that were

associated with metacognitive judgments that differentiated recollective from nonrecollective bases for accurate recognitions (remember vs. know judgments; high vs. medium confidence). Those tests also showed divergent validity as the trained classifier could not predict differences between decisions that did not differ based on recollective retrieval (know vs. guess; medium vs. low confidence). These findings validate the utility of this approach and suggest that it may be generalizable in detecting the degree of recollection-based retrieval in other recognition tasks.

The generalizability of this approach was tested in another study that trained language classifiers to differentiate justifications from the three aforementioned studies (viz., Dobbins & Kantner, 2019; Gardiner et al., 1998; Selmeczy & Dobbins, 2014) to predict differences in accurate and inaccurate eyewitness identifications from justifications in lineup procedures from out-of-sample data (Dobolyi & Dodson, 2018; Grabman et al., 2019). The guiding assertion was that the lineup procedure was a case of far transfer because testing data came from experiments with different stimuli, participants, and protocols. Nonetheless, the trained classifiers discriminated eyewitness recognition accuracy (for details, see Dobbins, 2022). The findings suggested that the language classifier approach generalizes in showing that trained classifiers predicted lineup identification accuracy and detected subjective states of recollection. Following this approach, a related study showed that language classifiers trained on eyewitness confidence statements predicted lineup identification accuracy within the same data set beyond confidence ratings, suggesting that statements include unique diagnostic information (Seale-Carlisle et al., 2022). These findings suggest that confidence ratings may struggle to capture qualitative differences in the bases for recognition. However, confidence and language may converge in revealing such bases.

The Present Study

Prior findings suggest that mnemonic discrimination tasks evoke retrieval monitoring and that language classifiers that are sensitive to recollection generalize to other tasks. Therefore, we hypothesize that language classifiers applied to verbal justifications for lure rejections will be sensitive to a strategy entailing recollection of studied objects for comparison with lures. The present study tested this hypothesis in two experiments using multitrial, study–test variants of the MST with verbal justification prompts after a subset of each combination of item and response types. To determine the bases for recognition decisions, verbal justifications were assessed using language classifiers that are described in more detail in the experiments proper.

Experiment 1 conformed to the canonical study–test MST with three response options (i.e., old, similar, and new). To clarify when to respond "similar," the instructions indicated that those responses should be given for lures that are the same as studied objects but with a featural change. Consistent with the proposal that lure classifications evoke retrieval monitoring, these instructions should inform participants that one valid approach is to compare the features of viewed test objects with the features of retrieved studied objects. Consequently, we expected language classifiers to be sensitive to such verbalizations following lure rejections.

To minimize the extent to which the instructions and response format alluded to retrieval monitoring as a strategy, we modified Experiment 2 protocol to include only two response options (i.e., old and new). The two- and three-response option MSTs have been routinely used because they are comparably sensitive to variables affecting lure rejections (e.g., Stark et al., 2015). By removing the "similar" response option, the instructions emphasized the retrieval monitoring strategy less. However, because lures share features with studied objects, test items should still naturally evoke recollection rejection evident in verbal justifications. Also, following each old/new response, participants rated their numeric confidence. This method allowed us to directly compare the predictive accuracy of confidence and language, which enabled comparisons of recognition memory theories. We discuss these issues more extensively when introducing Experiment 2.

Experiment 1

In Experiment 1, we characterized the bases for recognition decisions in an MST, including four study–test cycles in which participants attempted to classify target, lure, and foil test items as old, similar, and new. Following a subset of responses, participants also provided verbal justifications for their responses. Using a bag of words approach with individual words as predictors, penalized logistic regression models were trained to discriminate the language supporting correct versus incorrect responses for each item type (viz., targets, lures, and foils). The ability of these trained models to detect accurate responding was assessed using the area under the curve (AUC) of receiver operating characteristics (ROCs). If participants used retrieval monitoring most often when rejecting lures, then language classifiers should reveal AUCs indicating above-change model prediction accuracy.

Method

Transparency, Openness, and Data Availability

We report how we determined sample sizes, data exclusions, manipulations, and measures. The present research complied with the institutional review board at the University of North Carolina at Greensboro (Protocol #IRB-FY24-140). The deidentified data, analysis code, and stimulus materials are available on the Open Science Framework (https://osf.io/g5t9a/). We analyzed the data using R software (R Core Team, 2024) along with functions from the following packages: car (Fox & Weisberg, 2019), emmeans (Lenth, 2021), glmnet (Friedman et al., 2010), lme4 (Bates et al., 2015), pRroc (Robin et al., 2011), quanteda (Benoit et al., 2018), sjPlot (Lüdecke, 2024), and tidyverse (Wickham et al., 2019). The study was not preregistered.

Participants

The stopping rule was to test as many undergraduates from the University of North Carolina at Greensboro as resources allowed over one semester. We tested 126 participants in total and excluded two because the program crashed and four more to ensure that each counterbalancing format included an equal number. We excluded one of the latter four participants because they were 40 years old, which was older than we typically include in undergraduate samples. We excluded the other three participants because they were the last to be tested in formats with too many participants. The final sample included 120 participants. All participants received partial course credit as compensation.

Design and Materials

The experiment used a within-subjects design with three test item types. Target items were exact repetitions of studied items, lure items were different versions of studied items, and foil items were new items that did not appear during study. The materials comprised 432 pairs of object images from the Stark lab database (https://github.com/celstark/MST; Stark et al., 2013). Each pair included two versions of the same object (e.g., two coffee mugs). The database provides normative false alarm rates, reflecting how often lures were identified as targets. These rates were used to create bins varying in levels of perceptual similarity between targets and lures. We used items from Bins 2 through 4 to create an intermediate level of lure confusability. To examine the influence of such confusability on the predictive accuracy of language classifiers on lure recognition accuracy, we included the false alarm rate as a predictor in simple regression models, but found no significant moderation (see Supplemental Section S1).

Each study–test cycle included 108 unique objects. For counterbalancing, we created 12 groups of 36 objects with comparable normative false alarm rates (*M*s = 0.32–0.34, *SD*s = 0.10) and an equal number of objects from each bin (12 each). Each cycle included a unique set of three groups that rotated through item types, thus producing three experimental formats.

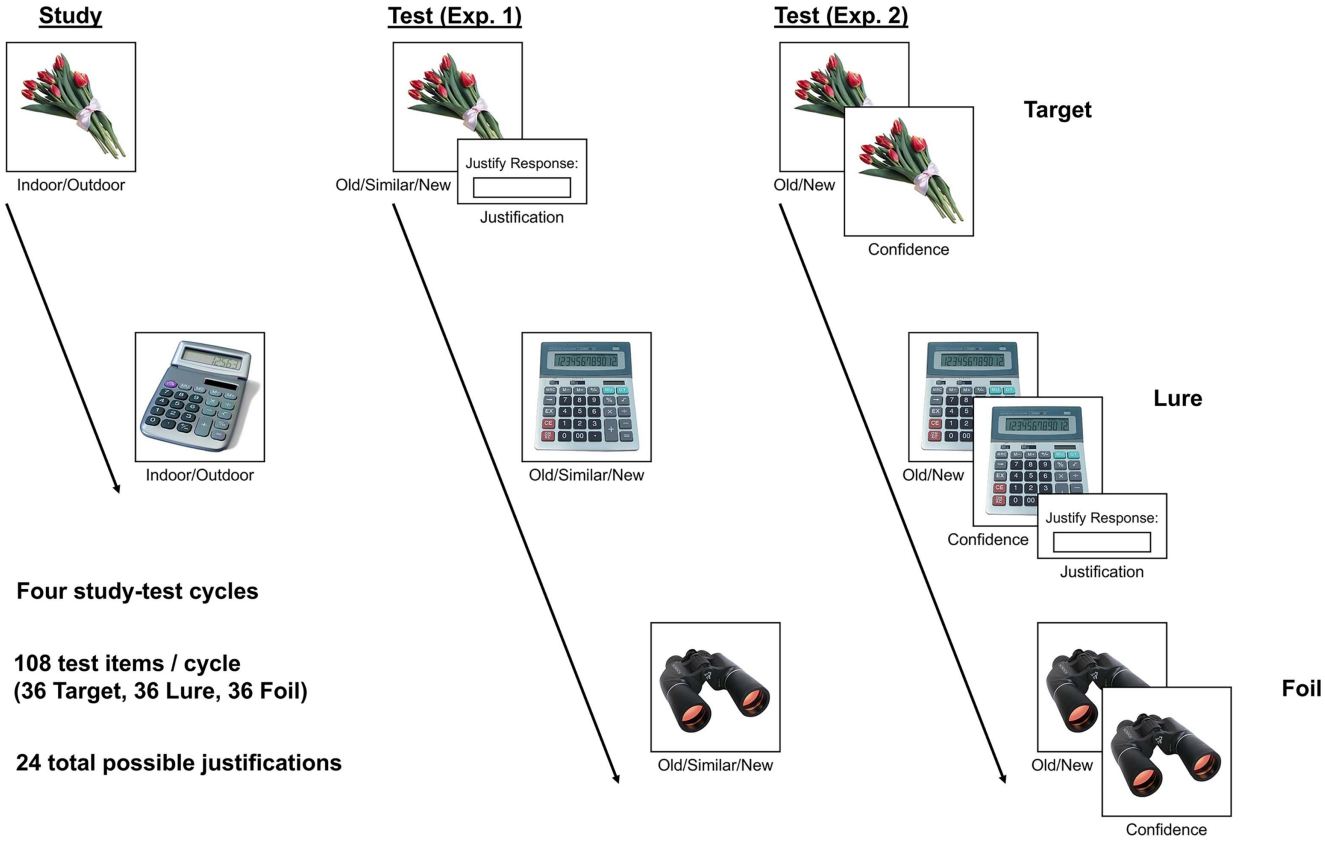
Procedure

Participants were tested in groups of up to four people with an experimenter present. E-prime 3.0 software (Psychology Software Tools, 2016) controlled stimulus presentation. All cycles used the same procedure with different items. Figure 1 displays a procedure schematic.

During study phases, each object appeared for 2 s with a 0.5-s interstimulus interval (ISI). Participants were told that while each object appeared, their task was to press a key to indicate whether the object belonged indoors (V) or outdoors (N). Participants were also told that there were no right or wrong answers. When participants responded before the 2 s deadline, the object remained on the screen for the remainder of the 2 s, and then the program proceeded to the next trial. When participants did not respond before the deadline, the

Figure 1

Schematic of the Experimental Procedures: Experiments 1 and 2



Note. The Mnemonic Similarity Task in the current experiments entailed four study–test cycles in which participants made indoor/outdoor judgments during study and recognition memory decisions at test. The study phases were the same in Experiments 1 and 2, whereas the recognition response options varied from three options in Experiment 1 (old/similar/new) to two options (old/new) followed by numeric confidence ratings (1 = low, 2 = medium, 3 = high) in Experiment 2. Critically, on a minority of the recognition trials (for 24 items), participants were prompted to type a verbal justification for their recognition decisions. Target test items were exact repetitions from the study; Lure test items were repetitions of study items with change features; and Foil test items did not appear during study. See the online article for the color version of this figure.

program advanced to the next trial after the 2 s elapsed. The presentation order was randomly determined by the program.

During test phases, each object appeared until participants made a response (i.e., the presentation duration was self-paced). A 0.5 s interstimulus interval intervened between trials. Participants were told that while each object appeared, their task was to press a key to indicate when the object was old (V), similar (B), or new (N). To ensure that the test matched closely with the procedure from the canonical MST, the instructions about how to respond to test objects were presented in a video from the Stark lab website (https://faculty.sites.uci.edu/starklab/mnemonic-similarity-task-mst/). The video file is also available on the Open Science Framework (https://osf.io/g5t9a/).

After the video, participants read a screen with instructions about the verbal justifications that they would make following recognition responses on some trials. Those instructions stated,

> In addition to deciding whether the objects are OLD (V), SIMILAR (B), or NEW (N), for some trials you will also be asked to provide details for why you made that decision. Please provide as much detail as you can and use the keyboard to make your response. There are no right or wrong answers.

On trials with justifications, objects disappeared, and prompts appeared as a text box in which responses were typed. The order of objects and justification prompts were both randomly determined by the program. The program also constrained the number of justification prompts per response type, such that participants could make a maximum of three justifications for old and similar responses and a maximum of one justification for new responses for each item type. The justification response limit was higher for old and similar responses because those were of primary theoretical interest. Table 1 shows the justification response frequency rates.

Statistical Methods

We compared recognition response accuracy with linear mixed effects models using the *lmer* function from the *lme4* package (Bates et al., 2015). Those models each included a by-subject random-intercept effect and a fixed effect of item type. We then conducted Wald's χ^2 hypothesis tests using the *Anova* function from the *car* package (Fox & Weisberg, 2019) and pairwise comparisons using the *emmeans* function from the *emmeans* package (Lenth, 2021).

We trained one language classifier for each item type to determine whether each classifier could distinguish the one correct response from the two incorrect responses. In both experiments, the correct responses were "old" for targets and "new" for foils. However, because the response options changed from old/similar/new in Experiment 1 to old/new in Experiment 2, the correct response for lures also changed from "similar" in Experiment 1 to "new" in Experiment 2. Thus, "new" responses to lures were incorrect in Experiment 1 and correct in Experiment 2.

We used the *quanteda* R package (Benoit et al., 2018) to tokenize justifications into unigrams (single words), with punctuation, numbers, and symbols removed, and all items set to lowercase. The unigrams were transformed into a document-term matrix, in which rows are justifications and columns are words used within the experiment. Cell entries in each row were the frequency of the words' usage within that justification (i.e., the BoW approach). Each word was used as a predictor without respect to its surrounding words in the justification. Because the number of unique words outnumbers the predicted cases, regularized regression techniques were used to prevent overfitting and yield a unique solution. The sum of the absolute values of the unigram coefficients was constrained during fitting using the least absolute shrinkage and selection operator (Tibshirani, 1996), implemented in the *glmnet* package (Friedman et al., 2010). This removed most predictors (unigrams) from the solution. The degree of constraint placed on the sum of the coefficients was determined by a 10-fold cross-validation procedure, with the final model yielding a minimum cross-validation error. To assess the training performance of each model, the AUC of ROCs was inspected using the *pROC* package (Robin et al., 2011). The cases with the highest probability predictions by the classifiers were inspected to determine the basis of classification.

Results

Test Responses

We first characterized the recognition response probabilities across all items and tests, regardless of whether they included verbal

Table 1
Verbal Justification Response Frequencies in Experiment 1

Response type	Item type		
	Target	Lure	Foil
A.			
Old	360 (100.00%)	360 (100.00%)	241 (66.94%)
Similar	351 (97.50%)	356 (98.89%)	351 (97.50%)
New	116 (96.67%)	118 (98.83%)	120 (100.00%)

Response type	Percentage of justifications completed											
	0%	33%	67%	100%	0%	33%	67%	100%	0%	33%	67%	100%
B.												
Old	0	0	0	120	0	0	0	120	17	23	22	58
Similar	0	2	5	113	0	1	2	117	0	2	5	113
New	4			116	2			118	0			120

Note. (A) Total verbal justification response frequencies summed across participants, and the percentage of participants who received the maximum number of possible justification prompts (in parentheses). (B) Frequencies of participants receiving each possible percentage of prompts.

justifications (Figure 2). Recognition accuracy was compared across conditions using a model with a fixed effect of item type and correct response probability as the outcome variable. A significant effect, $\chi^2(2) = 356.70$, $p < .001$, indicated that recognition accuracy was significantly higher for targets and foils than lures, smallest $t(238) = 16.17$, $p < .001$, and was not significantly different between targets and foils, $t(238) = 0.37$, $p = .93$. This pattern replicates prior findings (Stark et al., 2013), thus confirming that the task was designed appropriately and that participants generally followed the instructions.

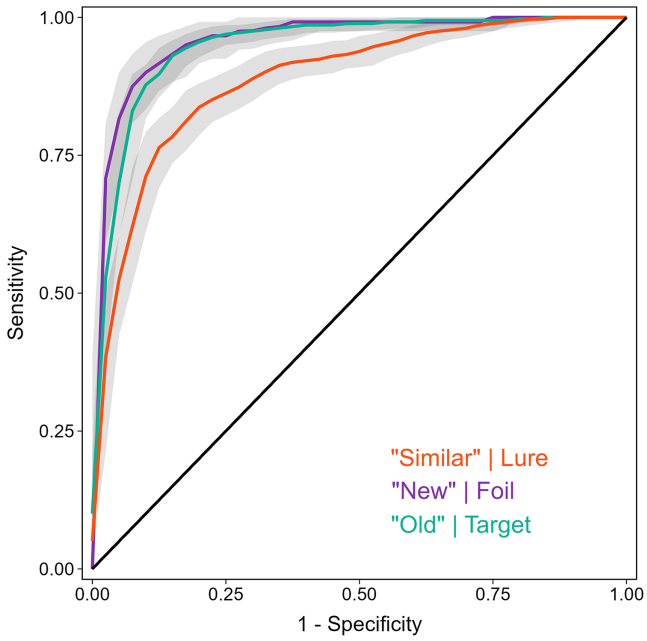
Classifier Performance

We next tasked each classifier with distinguishing the language for the one correct response from the two incorrect responses for each item type. Figure 3 shows high training performance for all three classifiers. The classifier for lures, which compared correct "similar" responses to incorrect "old" and "new" responses, yielded an AUC = .89 (95% CI [.87, .91]); the classifier for foils, which compared correct "new" responses to incorrect "old" and "similar" responses, yielded an AUC = .96 (95% CI [.94, .98]); and the classifier for targets, which compared correct "old" responses to incorrect "similar" and "new" responses, yielded an AUC = .95 (95% CI [.93, .96]). The illustrated 95% confidence intervals across the ROCs indicate that classifier accuracy for targets and foils did not differ and was higher than for lures. Overall, these results suggest that the justification language differed for correct and incorrect responses for every item type.

We examined the face validity of the classifiers by extracting the six justifications with the highest probabilities of reflecting correct responses (Table 2). These justifications clearly indicate that participants used different bases for responses across correct conclusions for the three item types. For example, justifications of correct "similar" responses (Section A) indicated that participants detected changes from remembrances (e.g., "there was a similar wooden stand but the connecting pieces were different"). In contrast, justifications of correct "new" responses (Section B) indicated that neither the item nor anything similar was previously seen (e.g., "its new because i havent seen it befire or anything likeit"). Interestingly,

Figure 3

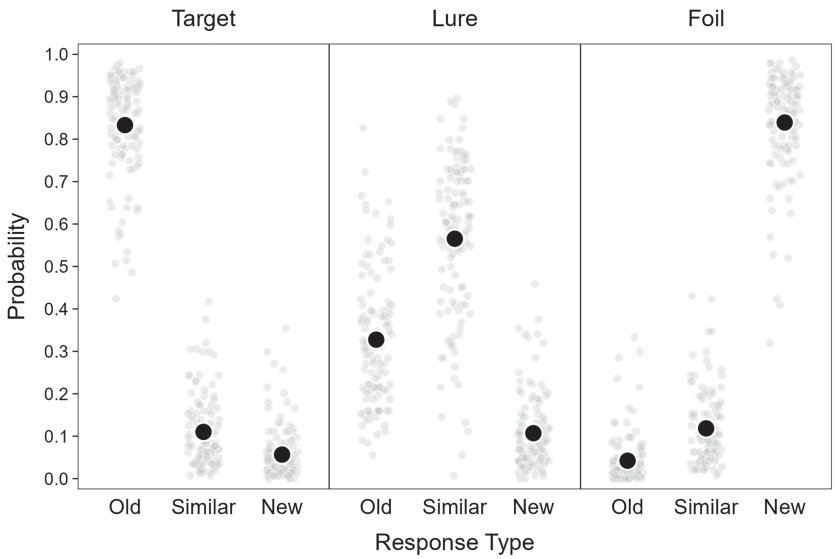
Classifier Training Performance (Receiver Operating Characteristics) in Experiment 1



Note. The quoted labels (left) indicate the recognition responses, and the unquoted labels (right) indicate the item types. Shaded regions are 95% confidence intervals. See the online article for the color version of this figure.

Figure 2

Test Items Response Probabilities in Experiment 1



Note. Larger darker points are aggregate condition probabilities, and smaller lighter points are individual participant probabilities.

Table 2
Extreme Classifier Scores for Item-Matched Analyses: Experiment 1

Test item	*p*lcorrect	Tokenized justification
A. "Similar" > "Old" and "New" classifications of lures		
Saw horse	.996	There was a similar wooden stand but the connecting pieces were different
Perfume bottle	.994	I saw a perfume bottle similar to this one but in a different shape
Bassinet	.994	I saw a crib similar to this one but in a different color
Gas mask	.994	I saw a similar mask before but it was slightly different
Swinging ball toy	.992	I chose similar because it from a different view but it could be old
Bassinet	.991	There was a basinet similar to this one but it was not this exact one
B. "New" > "Old" and "Similar" classifications of foils		
Blow dryer	*.966*	*I have never seen this before. I accidentally named it as old*
Swinging ball toy	.964	It is new because I have not seen it before or anything like it
Pitcher	.957	I have never seen this
Jumper clamps	.950	I did not see the clips like those in my first test
Wind up monkey	.948	I did not see that object in the previous experiment
3 drawer dresser	.942	I do not recall seeing any furniture but specifically not this dresser
C. "Old" > "Similar" and "New" classifications of targets		
Binoculars	.985	I remember putting outdoor for this in part 1
Ballpoint pen	.982	It saw the same pen with the same color in the first exercise
Starfish	.978	I have seen it before in the indoor and outdoor test
Jumper clamps	.976	It was the same clips from the experiment exact same thing
Lariat	.971	I remember seeing that exact object in the first trial because I remember questioning what exactly it was the first time
Starfish	.971	I saw the same exact starfish in the previous items

Note. The italicized outcome in Panel B was incorrect because the model classified the response as "Old" to a Foil item.

the classifier assigned the highest probability in Section B to a response that was technically incorrect. Here, the participant correctly indicated that despite pressing the "old" response key, they had never seen the item (i.e., "i have never seen this before i accidentally named it as old"). Finally, justifications of correct "old" responses (Section C) indicated remembering, often with a belief that the item was identical to the remembrance (e.g., "it was the same clips from the experiment exact same thing"), and sometimes with recollection of contextual details (e.g., "i remember seeing that exact object in the first triall because i remember questioing what exactly it was the first time").

Discussion

Experiment 1 showed that in a canonical MST, verbal justifications of recognition responses distinguished correct from incorrect responses for all test item types. The classifier analyses implicated qualitatively different bases for the three possible correct conclusions and illuminated the cognitive processes underlying similar lure rejections. The findings suggest that lures were often rejected when the features they shared with studied items triggered recollections that enabled change detection (viz. recollection rejection). Conversely, justifications for correct responses to targets and foils suggested that those responses were based primarily on recollection to accept items and the complete absence of remembrance for unstudied items, respectively. Although the classifiers performed very well, their performance must be interpreted cautiously for two reasons. First, participants could have tailored their justifications to match their preceding conclusions. However, the first justification in Section B was strongly predictive of item type even though it conflicted with the initially incorrect conclusion, which suggests participants genuinely reported their experiences instead of justifying their conclusions post hoc. Second, although the least absolute shrinkage and selection operator regression method is designed to limit overfitting, it does not guarantee it. To rule out these alternative explanations, we tested the classifiers for generalization in Experiment 2.

Experiment 2

Experiment 2 generalized the language classifiers from Experiment 1 to responses in a format that should do less to encourage tailoring justifications to match recognition responses. Instead of providing three options at test, participants received two, with instructions to respond "old" to items that matched studied items and "new" to other items, including different versions of studied items. Showing that the classifiers from Experiment 1 predict these responses would mitigate concerns about participants tailoring their justifications and strengthen the conclusion that natural language conveys qualitatively different experiential bases for recognition responses in the MST.

Experiment 2 also examined whether justifications distinguish correct from incorrect responses more accurately than retrospective confidence ratings, a common subjective measure. After each recognition response, participants rated their confidence on a 3-point scale (1 = *low*, 2 = *medium*, 3 = *high*). By comparing the predictive accuracy of verbal justifications and numeric confidence ratings, we were able to contrast the relative amount of information in the language used to support old/new responses with the confidence in those responses. Researchers have used a one-dimensional signal detection account to describe how MST responses are made (Loiotile & Courtney, 2015; Stark et al., 2015). Under this account, targets, lures, and foils yield overlapping evidence distributions reflecting the strength of the match between probes and memories, with the match being strongest for targets, intermediate for lures, and weakest for foils. Conversely, because multiprocess theories assume

qualitatively different bases for responses to the three item types (e.g., Rotello et al., 2000), they imply that statements may be more informative than numeric ratings. If so, then language will better predict item types than confidence, which will be important for determining how often lure rejections are accomplished via recollection rejection instead of the evidence from a one-dimensional evidence signal. Thus, finding that justifications predict MST responses better than confidence, especially for lures, would warrant further consideration about the extent to which inferences about the processes underlying such decisions can be inferred from one-dimensional signal detection models.

Method

Participants

The stopping rule was to end data collection once the sample included at least the same number of usable participants as in Experiment 1, and based on available resources. This rule deviated slightly from Experiment 1 because data collection began late in the semester and had to continue into the following semester. We tested 143 participants in total and excluded seven because they could not complete the experiment in the allotted time, one because the program crashed, and three because they were the last to be tested in the formats that included too many participants, which ensured that each format included an equal number of participants. The final sample included 132 participants. All participants received partial course credit as compensation.

Design and Materials

The design and materials were the same as in Experiment 1.

Procedure

The procedure maintained key aspects from Experiment 1, including the participant testing method, number of study–test cycles, and item presentation rates. However, Experiment 2 test procedure differed in the test phases (see Figure 1). In the test phases, participants read instructions instead of watching a video. The instructions stated,

> Now, I am going to test your memory for those items that you just viewed. For each item, you need to indicate if the item is OLD (V) or NEW (N). OLD refers to an item that you just saw, and you should press the V key on the keyboard. New refers to 1) an item that is similar to one that you just saw, but not exactly the same, or 2) an item that you have not seen within the context of the experiment. For new items, you should press the N key on the keyboard. A prompt will appear below each picture to remind you which buttons to push. Before you begin, I am going to show you a brief example.

After these instructions, participants viewed a slide showing examples of study and test objects for each item type along with instructions about how to respond (available on the Open Science Framework at https://osf.io/g5t9a/; Wahlheim et al., 2025). After viewing the slide, participants read more instructions that stated,

> After deciding whether the objects are OLD (V) or NEW (N), you will also rate your confidence in that decision. You will make your rating using the following scale: (1) *not at all sure*, (2) *somewhat sure*, and (3) *very sure*. You should rate your decision as *not at all sure* when you feel like you were primarily guessing, *somewhat sure* when you think you might be right, but you have some doubt, and *very sure* when you feel absolutely certain.
>
> Lastly, for some trials, you will also be asked to provide details for why you made an OLD (V) or NEW (N) decision and your confidence judgment. Please provide as much detail as you can and use the keyboard to type your response. There are no right or wrong answers.

Justification prompts appeared after confidence ratings. Because Experiment 2 included fewer response options, the program constrained the number of justification prompts to a higher number than in Experiment 1: Participants could make a maximum of four justifications for each response and item type combination. Table 3 shows the total justification response frequencies.

Results

Test Responses

We first characterized the recognition response probabilities across all items and tests, regardless of whether they included verbal justifications (Figure 4A). Recognition accuracy was compared across conditions using a model with a fixed effect of item type and correct response probability as the outcome variable. A significant

Table 3
Verbal Justification Response Frequencies in Experiment 2

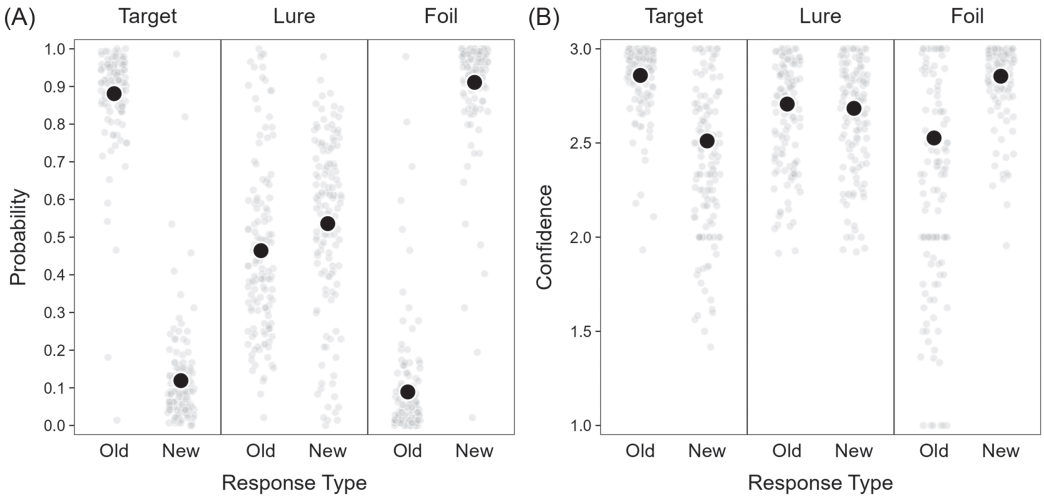
Response type	Item type		
	Target	Lure	Foil
A.			
Old	526 (99.62%)	527 (99.81%)	384 (72.73%)
New	493 (93.37%)	519 (98.30%)	527 (99.81%)

Response type	Percentage of justifications completed														
	0%	25%	50%	75%	100%	0%	25%	50%	75%	100%	0%	25%	50%	75%	100%
B.															
Old	0	0	1	0	131	0	0	0	1	131	10	20	18	8	76
New	3	4	2	7	116	1	0	2	1	128	0	0	0	1	131

Note. (A) Total verbal justification response frequencies summed across participants, and the percentage of participants who received the maximum number of possible justification prompts (in parentheses). (B) Frequencies of participants receiving each possible percentage of prompts.

Figure 4

Test Items Response Probabilities and Confidence Ratings in Experiment 2



Note. Panel A displays recognition response probabilities. Panel B displays confidence ratings that followed recognition responses. Larger darker points are aggregate condition probabilities, and smaller lighter points are individual participant probabilities.

effect, $\chi^2(2) = 394.20$, $p < .001$, indicated that recognition accuracy was significantly higher for both targets and foils than lures, smallest $t(262) = 16.44$, $p < .001$, and was not significantly different between targets and foils, $t(261) = 1.44$, $p = .32$. This pattern replicates prior findings from MSTs with two-response test procedures (Stark et al., 2015), thus confirming that the task was designed appropriately and that participants generally followed the instructions.

Confidence Ratings

We next characterized the confidence ratings across the four study–test cycles for all test items (Figure 4B). A model with fixed effects of item and response type indicated a significant effect of item type, $\chi^2(2) = 14.49$, $p < .001$, no significant effect of response type, $\chi^2(1) = 3.25$, $p = .07$, and a significant interaction, $\chi^2(2) = 532.60$, $p < .001$. Confidence was significantly higher for correct than incorrect responses for targets and foils, smallest $t(642) = 14.48$, $p < .001$, but was not significantly different between response types for lures, $t(642) = 0.15$, $p = .88$.

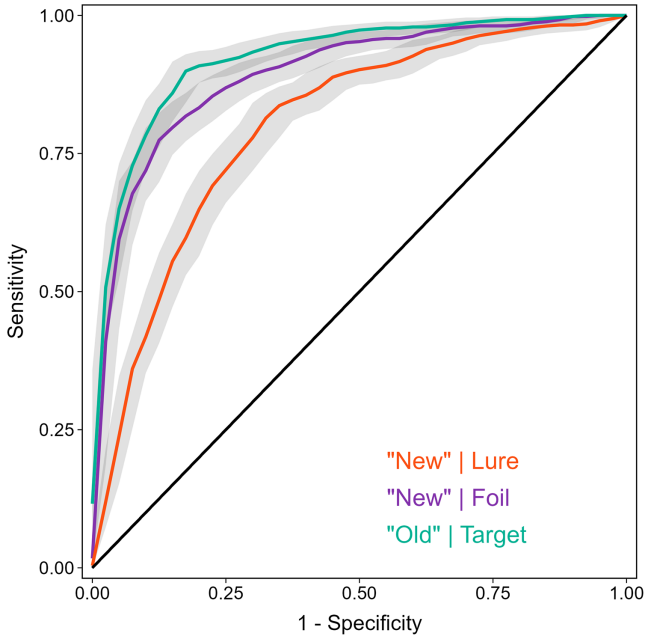
Classifier Performance

After establishing recognition and confidence patterns, we tasked each classifier with distinguishing the language for correct and incorrect responses for each item type. These comparisons were between two responses because there were only "old" and "new" options. Figure 5 shows that training performance was again quite high. The classifier for lures, which compared correct "new" responses to incorrect "old" responses, yielded an AUC = .80 (95% CI [.77, .83]). Additionally, the classifier for foils, which compared correct "new" responses to incorrect "old" responses, yielded an AUC = .89 (95% CI [.87, .91]). Finally, the classifier for targets, which compared correct "old" responses to incorrect "new" responses, yielded an AUC = .92 (95% CI [.91, .94]). The confidence intervals across ROCs indicate that classifier accuracy for targets and foils did not differ and was higher than for lures. Overall, these results suggest that the justification language differed for correct and incorrect responses for every item type.

We examined the face validity of the classifiers by extracting the six justifications with the highest probabilities of reflecting correct responses (Table 4) and found evidence for qualitative differences in

Figure 5

Classifier Training Performance in Experiment 2



Note. The quoted labels (left) indicate the recognition responses, and the unquoted labels (right) indicate the item types. Shaded regions are 95% confidence intervals. See the online article for the color version of this figure.

Table 4
Extreme Classifier Scores for Item-Matched Analyses: Experiment 2

Test item	*p*lcorrect	Tokenized justification
A. "New" > "Old" classifications of lures		
Lock	.967	It is new but i dont there was a similar object to it just at a different angle or slightly different
Matches	*.955*	*The first two pictures i really dont remeber seeing them but the last one i did see but thought it was maybe a bit different*
Computer tower	.937	The brand of the pc was different i did not recognize the rooster compass or the other item from the previous section
Ruler	.935	There was a ruler that i saw but it was not a wood one it was a metal one
Film reel	.933	I remembred that there was an item like this one but it was different i think the color was also different
Graduation cap	.925	The past one was a graduation cap but this one is not exactly that one
B. "New" > "Old" classifications of foils		
Binoculars	.982	This object would have been in the outdoor group because you use this to see something far away that is outdoors but the object is stored
Doorbell	.981	I dont remember seeing anything like this item in the first round i am also not sure what this item is
Axe	.975	I do not remember seeing an axe in the first part i dont recall seeing a tool similar to an axe either
Lock	.973	I do not recall seeing this image in the first section
Megaphone	.970	I do not think i saw this item in the first round there may have been a different picture of this item
Pumpkin	.968	I did not see this item before therefore this item is new
C. "Old" > "New" classifications of targets		
Bin	.994	When first seeing it i told myself it depends how you use it to decide if it was an indoor or outdoor item
Lariat	.993	I saw it and didnt know what it was so i just guessed indoor or outdoor which is why i remember it
Dog bowl	.992	I remember the dog bowl because i thought for a second on if it should be indoor or outdoor
Perfume bottle	.991	I remeber trying to figure out what this object was i dont really know what its used for or if its outdoor or indoor
Pumpkin	.991	I knew this one was old because i pressed n for outdoor since its something you keep outside for halloween
Pumpkin	.989	I remember debating whether it was an indoor thing or outdoor so i know it was the same picture

Note. The italicized outcome in Panel A was incorrect because the model classified the response as "Old" to a lure item.

correct responses across item types. Justifications of correct "new" responses to lures (Panel A) indicated that participants detected changes from remembrances (e.g., "the brand of the pc was different"). In contrast, justifications of correct "new" responses to foils (Panel B) indicated that the item was not previously seen, and often that nothing similar had been seen (e.g., "i do not remember seeing an axe in the first part i dont recall seeing a tool similar to an axe either"). Finally, justifications of correct "old" responses (Panel C) often reported the experience of recollecting contextual details (e.g., "when first seeing it i told myself it depends how you use it to decide if it was an indoor or outdoor item"). Overall, these findings suggest that the item types are correctly recognized based on different kinds of mnemonic evidence, even when the same response is made to different item types (i.e., "new" responses for lures and foils).

Comparing Prediction Accuracy for Classifiers and Confidence Ratings

We tested the generalizability of the classifiers by first using the trained classifiers from Experiment 1 to predict response accuracy in Experiment 2, pitting these predictions against confidence ratings. To obtain each prediction, Experiment 2 justifications were transformed into a document-term matrix, and the weights of each classifier from Experiment 1 were applied to the frequencies of words common to both experiments. The summed values of the weights resulted in the predicted log-odds that the justification is accurate for a given item type. Hence, there are three predicted log-odds for each Experiment 2 justification, one from each Experiment 1 classifier. These were used in three sets of hierarchical regressions that each pitted the relevant classifier from Experiment 1 against confidence ratings in Experiment 2.

Experiment 2 included three pairs of correct and incorrect justifications made to the same items (i.e., "new" compared to "old" responses for lures, "new" compared to "old" responses for foils, and "old" compared to "new" responses to targets). For each pair, we used hierarchical logistic regressions to predict recognition accuracy. Step 1 tested whether the relevant classifier from Experiment 1 predictions predicted recognition accuracy. Step 2 added confidence ratings to see if they contributed over and above classifier predictions. Step 3 added the interaction between classifier predictions and confidence ratings. Importantly, this step allowed us to test the idea that recollective processes inform the language used and subjective confidence, such that confidence increases as recollective language becomes more diagnostic of

recognition accuracy. We consider this possibility more fully in the Discussion section.

For the comparison of "new" and "old" responses to lures (Table 5A), Step 1 shows that Experiment 1 classifier for "similar" responses to lures strongly predicted lure rejections in Experiment 2. However, confidence ratings on Step 2 did not reliably predict lure rejections, and the contribution of the language classifier's predictions remained largely unchanged. Consistent with this, the model fit did not improve, $\chi^2(1) = 0.16$, $p = .69$. Therefore, confidence did not play a role in predicting lure rejections in the presence of language-based predictions. However, the interaction between these predictors (Step 3) was significant, and the model fit improved, $\chi^2(1) = 26.31$, $p < .001$. Figure 6A illustrates the Classifier × Confidence interaction predicting accurate lure responses. On low-confidence trials, language did not predict accuracy, as shown by the substantial uncertainty around the model's predicted and largely flat trend. However, as confidence increased to medium and high, so too did the slope and precision of the language predictions, indicating that language became more determinative of accurate rejection of lures as numeric confidence increased.

Next, for the comparison of "old" and "new" responses to targets (Table 5B), Step 1 shows that the Experiment 1 classifier for "old" responses to targets strongly predicted target recognition in Experiment 2. Step 2 shows that both the classifier and confidence ratings significantly predicted recognition accuracy, and the model fit was improved, $\chi^2(1) = 48.82$, $p < .001$. Finally, the interaction between these predictors (Step 3) was significant, and the model fit improved, $\chi^2(1) = 71.75$, $p < .001$. Figure 6B shows that—as for lures—as confidence increased, the language became more diagnostic of accurate acceptance of targets.

Finally, for the comparison of "new" and "old" responses to foils (Table 5C), Step 1 shows that the Experiment 1 classifier for "new" responses to foils strongly predicted correct rejections in Experiment 2. Step 2 shows both the classifier and confidence ratings significantly predicted recognition accuracy, and the model fit was improved, $\chi^2(1) = 99.09$, $p < .001$. However, unlike the previous two comparisons, Step 3 yielded no significant interaction and did not significantly improve model fit, $\chi^2(1) = 2.61$, $p = .11$. Figure 6C shows that the slopes for language-based predictions increased moderately (additively) with increases in confidence.

Comparing Item Type Predictions From Language Classifiers and Confidence Ratings

The findings thus far suggest that qualitatively different subjective experiences reflected in justification language distinguish accurate responses across the item types. These findings also suggest that language distinguishes these outcomes better than confidence ratings. We tested this conclusion further using the language classifiers from Experiment 1 to classify the item types in Experiment 2 as lures, targets, or foils. This approach when applied to neurophysiological data is sometimes referred to as "decoding," reflecting that the goal is to determine what type of item the participant is evaluating and not necessarily the accuracy of their judgments, although the two will be linked. We performed the same

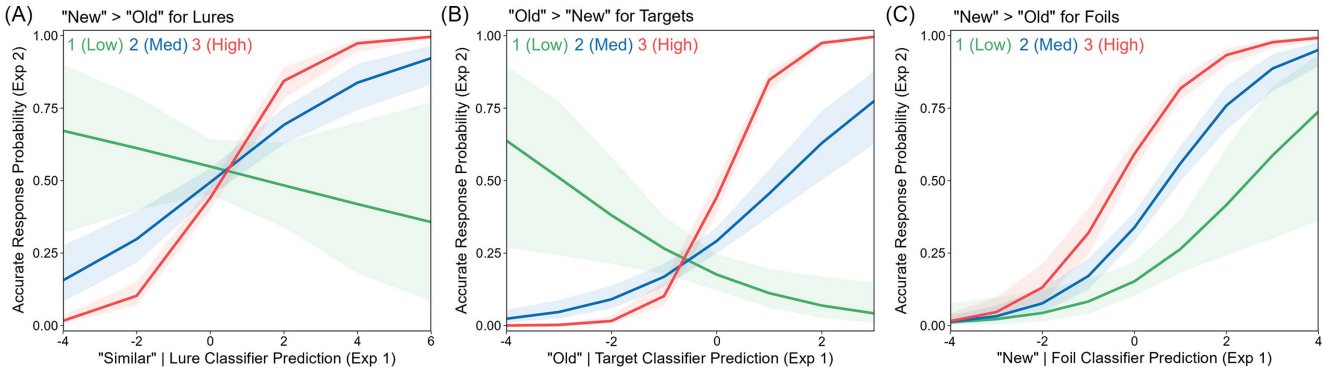
Table 5
Hierarchical Regression Predicting Recognition Accuracy From Language and Confidence

	Model 1		Model 2		Model 3	
Predictor	*OR*	Statistic	*OR*	Statistic	*OR*	Statistic
A. "New" > "Old" for lures (1,046 observations)						
Step 1						
Language	1.95	**10.09**	1.94	**9.90**	2.07	**10.08**
Step 2						
Confidence			0.97	−0.40	0.87	−1.86
Step 3						
Language × Confidence					1.42	**5.09**
R^2 Tjur	.11		.11		.14	
B. "Old" > "New" for targets (1,019 observations)						
Step 1						
Language	4.10	**13.80**	3.49	**11.98**	4.09	**12.31**
Step 2						
Confidence			1.74	**6.73**	1.52	**4.86**
Step 3						
Language × Confidence					2.21	**8.85**
R^2 Tjur	.28		.32		.37	
C. "New" > "Old for foils (911 observations)						
Step 1						
Language	2.96	**12.28**	2.78	**11.00**	2.81	**10.96**
Step 2						
Confidence			2.08	**9.28**	1.95	**7.83**
Step 3						
Language × Confidence					1.15	1.62
R^2 Tjur	.21		.30		.30	

Note. The pseudo-R^2 value Tjur is the absolute value of the difference between the average predicted probability for accurate trials, subtracting that of inaccurate trials. The language predictors above refer to the language classifiers from Experiment 1 for "similar" responses to lures (Panel A), "old" responses to targets (Panel B), and "new" responses to foils (Panel C). Bolded statistics are significant at $p < .001$. *OR* = odds ratio.

Figure 6

Language Classifier by Confidence Interactions When Predicting Recognition Response Accuracy



Note. Interactions predicting accurate lure responses (Panel A), target responses (Panel B), and foil responses (Panel C). Shaded regions are 95% confidence intervals. See the online article for the color version of this figure.

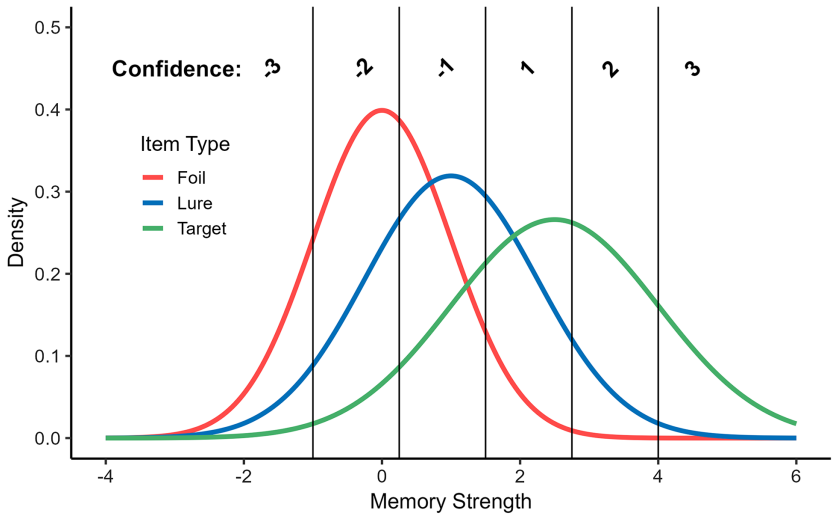
analysis using the combination of old/new and confidence responses—which we report first in what follows—and compared the success of the two approaches to determine if language classifiers continue to prevail.

Predictions From Confidence Ratings. We begin with the analysis of confidence ratings, which allowed us to test the one-dimensional signal detection model. Confidence ratings ranged from 1 to 3 following "old" and "new" recognition responses. These were first combined into one signed confidence predictor by negatively signing "new" confidence ratings, resulting in six values from −3 to +3. This coding is assumed under an extant signal detection model of MST recognition responses (Loiotile & Courtney, 2015; see Figure 7). Under this model, targets yield the greatest memory strength because they exactly match the studied items, lures yield intermediate memory strength because their match is close, but not identical, to the studied items, and foils yield the weakest memory strength because neither they nor anything similar was studied. According to the model, participants make confidence ratings by placing decision criteria along the strength axis, such that confidence increases monotonically with item strength. The model also allows for flexible differences in the relative variance of the evidence distributions of the item types. Given this one-dimensional representation, the model assumes that extreme confidence ratings are more likely to reflect foils or targets, whereas intermediate values are more likely to reflect lures.

We tested this signal detection model using signed confidence ratings in a multinomial logistic regression model predicting the item type. Although signal detection and logistic regression models are not identical, they assume similar representations. Therefore, the pattern of success in one implies the same pattern in the other. Here, if confidence is limited in predicting item types under the logistic approach, it should be similarly limited when

Figure 7

Strength Distributions for Mnemonic Similarity Task Items According to a One-Dimensional Signal Detection Account



Note. See the online article for the color version of this figure.

post hoc fitting a one-dimensional signal detection model (for a discussion, see DeCarlo, 1998). During multinomial regression of three categorical outcomes (i.e., item types), two contrasts are estimated for each predictor. With one confidence predictor, the model estimates the log-odds of the foil–target contrast and of the lure–target contrast. The target class, therefore, serves as the reference level. Table 6 (top rows) displays the results of the confidence model, showing that confidence strongly discriminates foils from targets and reliably but less strongly discriminates lures from targets.

To clarify the model's performance, we translated its predictions into table form. Multinomial logistic regression provides three estimated probabilities for the three possible item types on each trial, and the maximum reflects the model's choice (for details, see Agresti, 2012). Table 7 shows the observed confidence rating counts across the item types (Panel A) and the one-dimensional model's predictions (Panel B). Such tables are referred to as confusion matrices because they show both the degree of correct prediction and the degree to which errors of prediction are distributed to incorrect outcomes. While the overall association between item types and predictions is significant, $\chi^2(4) = 64.11$, $p < .001$, the success rate (.36) is modest and barely higher than chance (∼.33). Moreover, the model identified lures poorly, yielding few "lure" classifications (Panel B, left, middle), of which the most were incorrectly given to targets (69), not to lures (35). The model also struggled with "target" classifications, assigning more of those to lures (527) than to targets (526). Stated differently, when numeric confidence predicted item types, the model massively confused lures with foils and targets, and the model also failed to associate lures with their correct item type. This is consistent with the model's more modest discrimination of lures and targets than foils and lures, shown by multinomial regression (Table 6, top rows).

We next reran the multinomial logistic regression analysis but restricted confidence to the highest absolute levels (i.e., −3 and 3). Typically, for decision models, classification should improve with increased certainty in responses. However, extreme confidence values in the one-dimensional signal detection model are almost entirely linked to foil and target items (see Figure 7). Consequently, the model should perform even more poorly when attempting to produce "lure" classifications. Indeed, the model completely failed to classify any items as lures (Table 7B, right, middle), even though lures comprised a third of all items. This failure is easily explained by the raw counts in Table 7A (middle row) because the extreme confidence ratings for lures were about evenly distributed between ratings of −3 (308) and 3 (342). These modeling results suggest that if the item types only differed on a one-dimensional memory strength signal, captured by confidence ratings, then lure items must be difficult to distinguish from foils and targets. This is not a limitation of multinomial regression; it is a consequence of treating confidence and memory strength as one-dimensional in the MST. Because the lure distribution intervenes between the target and foil distributions (Figure 7), the link between numeric confidence and the item types must be weak (Table 7). The multinomial regression using confidence illustrates this weakness.

Predictions From Language Classifiers. We next conducted analyses of the language classifier predictions analogous to those for confidence ratings. Each recognition accuracy classifier from Experiment 1 was used to render predictions for the justifications of Experiment 2. This produced three log-odds estimates for each trial, reflecting the models' predictions that the current trial reflects an accurate response to a lure, foil, or target. These scores were entered into a multinomial logistic regression to predict the item type on each trial. Table 6 (bottom rows) shows that the lure accuracy classifier reliably discriminated lures from targets, the foil accuracy classifier reliably discriminated foils from targets, and the target accuracy classifier reliably discriminated targets from foils.

Table 8 clarifies the model's performance and predictions by taking the three model-estimated probabilities for the three possible item types on each trial with the maximum indicating model's choice. Table 8 (left) shows a significant association between this model's classifications and the item types, $\chi^2(4) = 429.94$, $p < .001$, and its overall success rate (.50) was well above chance (∼.33). Both the strength of association and success rate were clearly superior to the confidence-based classifier. As for confidence, the pattern of success and confusion in Table 8 is illustrative. Although the language classifiers performed better than the confidence-based classifier, the language approach often confused lures with targets: Table 8 (left) shows that lures were more likely to be labeled as lures (430), as compared to foils (192), but the model also often identified lures as targets (424). This is far better than for confidence predictions, but it still suggests the language of errors to lures is more similar to target recognition than foil rejection, perhaps reflecting the shared features of lures and targets.

Finally, we restricted the analysis to extreme confidence ratings (−3 and 3) and repeated the procedure to see if performance further deteriorated when classifications were restricted to high certainty

Table 6
Multinomial Regression Estimates Predicting Item Types From Confidence Ratings (Top Section) and Language Classifier Predictions (Bottom Section)

Model	Predictor	*OR*	Statistic	*p*
Confidence	Confidence (Foil vs. Target)	0.87	**−8.11**	<.001
	Confidence (Lure vs. Target)	0.96	**−2.18**	=.029
Language	Lure accuracy classifier (Foil vs. Target)	0.87	−1.49	=.138
	Lure accuracy classifier (Lure vs. Target)	1.50	**4.59**	<.001
	Foil accuracy classifier (Foil vs. Target)	1.91	**9.20**	<.001
	Foil accuracy classifier (Lure vs. Target)	1.05	0.77	=.442
	Target accuracy classifier (Foil vs. Target)	0.63	**−6.21**	<.001
	Target accuracy classifier (Lure vs. Target)	0.98	−0.32	=.750

Note. Statistic = *z*-statistic of the Wald for single coefficients. Both models included 2,976 observations. Bolded statistics are significant. *OR* = odds ratio.

Table 7
Performance of Confidence Classifier Predicting Item Types

Item type	Confidence −3	−2	−1	1	2	3
A. Empirically observed confidence counts						
Foil	425	86	16	89	120	175
Lure	308	176	35	21	164	342
Target	238	186	69	12	79	435

	Confidence classifier (all ratings)			Confidence classifier (extreme ratings only)		
Item type	Foil	Lure	Target	Foil	Lure	Target
B. Model predicted confidence counts						
Foil	511	16	384	425	0	175
Lure	484	35	527	308	0	342
Target	424	69	526	238	0	435

responses, as observed using the confidence-based model. The predicted counts in Table 8 (right) show that the model continued to demonstrate a strong association between its classifications and the item types, $\chi^2(4) = 437.53$, $p < .001$, and maintained a high absolute proportion correct (.54). This finding is impossible under the one-dimensional signal detection account (Figure 7). Language classifications, unlike numeric confidence ratings, are therefore not limited to a one-dimensional relationship with the three item types in the current MST variant.

Predicting Experiment 1 Item Types From Experiment 2 Language Classifiers

Overall, the classification approach further validates Experiment 1 language classifiers while also revealing the shortcomings of the one-dimensional, signal detection-based approach using numeric confidence ratings. In the Discussion section, we further explain why this suggests that MST responses are best conceived of as multidimensional and not simply a reflection of a one-dimensional strength of the memory signal evoked by the probes. For completeness, we also performed the reverse analysis by predicting item types in Experiment 1 using the trained language accuracy classifiers of Experiment 2 (see Supplemental Section S2). Although this approach yielded a significant association between predictions and outcomes, the association was less robust than when predicting Experiment 2 item types from Experiment 1 classifiers. This suggests that language classifier training benefits from having unique, correct responses associated with each item type, as in the test procedure from Experiment 1.

Table 8
Performance of Language Classifier Predicting Item Types

Item type	Language classifier (all ratings)			Language classifier (extreme confidence only)		
	Foil	Lure	Target	Foil	Lure	Target
Foil	469	168	274	343	94	163
Lure	192	430	424	101	242	307
Target	182	252	585	94	132	447

Discussion

Experiment 2 showed that the language used to justify MST responses conveyed as much or more information than the numeric certainty in old/new responses. Additionally, the language classifiers trained in Experiment 2 confirmed the training outcomes in Experiment 1 in that language for all item types became more distinctive with accurate responding. For lures, accurate responses were accompanied by language that reported recollection rejection, expressed as detecting changes between test items and remembered study items. For foils, accurate responses were accompanied by language indicating the absence of recollection and, sometimes, claims that neither the test item nor anything like it had been encountered. Finally, for targets, accurate responses were accompanied by claims that study items were recollected and, often, subjective certainty that those items exactly matched test items. Collectively, these differences suggest that correct rejections of lures and foils were primarily made on qualitatively different bases, with both differing from the primary basis for correct target recognitions.

Although these findings replicate the language accompanying analogous outcomes in Experiment 1, proper construct validation of language classifiers requires testing out of sample. Experiment 2 showed that trained classifiers from an independent sample can be used to predict the accuracy of another group performing the MST, even when the tests include different response options. The informativeness of justification language is evident in the coefficients in Steps 1 and 2 of Table 5 and the pseudo-R^2 Tjur statistic, which captures the extremity of the models' probability assignments to targets and lures. Language-based predictions (Step 1) were robust predictors of recognition accuracy in every model. When confidence was added (Step 2), language-based predictions remained robust predictors. Moreover, the coefficients and test statistics were all larger for the language than confidence predictor in the first two steps of each model. It is worth emphasizing that the language models were not fit to the current participants' data. Instead, they reflect the mechanical weighting and combination of individual words learned during training in Experiment 1 to predict recognition accuracy here. Additionally, the multinomial regression approach entailed

fitting *z*-transformed predictions of Experiment 1 language classifiers' scores to predict Experiment 2 item types.

In Experiment 1, we noted that the strong classifier training performance could have reflected participants tailoring their justifications to the recognition responses that immediately preceded the justifications. The clear generalization of the trained Experiment 1 classifiers to Experiment 2 recognition responses eliminates this concern because the response options differed between Experiment 2 (old/new) and Experiment 1 (old/similar/new).

The second illustration of the relative value of language compared to confidence was shown when using both to predict item types: Confidence performed poorly as a predictor, particularly for identifying lures. This could be anticipated from the pattern of confidence counts for lures (Table 7A), which showed higher density around the extreme than intermediate ratings. This pattern deviates from what would be predicted by a model assuming a one-dimensional evidence signal with lures providing intermediate evidence (cf. Loiotile & Courtney, 2015). To further test this model, one could use a more granular confidence scale to isolate where the density is higher for lures than other item types. However, one study found that language uniquely contributed to predictions of eyewitness memory regardless of confidence granularity (Seale-Carlisle et al., 2025). Another way to test this model is to use post hoc fitting procedures to determine if some combination of distribution form, relative distribution variance, distribution position, and criterion locations improves classification. However, the observed counts (Table 7A) cast doubt on the utility of this approach because confidence is ambiguously distributed between correct "New" (negative values) and incorrect "Old" (positive values) responses. The ideal approach is to leverage an information source that disambiguates high certainty that a lure was studied (an error) from high certainty that it is novel (a correct response). Justification language is ideally suited for this purpose.

Why, cognitively, does language improve prediction so dramatically, even in the presence of numeric confidence reports? The strong classifier generalization from Experiment 1 to Experiment 2 suggests that the way participants describe their successful responses to the different item types must be highly consistent regardless of whether there are three or two response options. Critically, language classifiers were sensitive to recognition accuracy during standard recognition tasks because they capitalized on the consistency of words and phrases used to describe successful recollection experiences (Dobbins & Kantner, 2019). This recollection sensitivity hypothesis has gained further support from findings showing that justification language is diagnostic of eyewitness face recognition accuracy (Grabman et al., 2024) and that this language becomes increasingly diagnostic for people with higher recognition abilities.

The present study provides further support to the recollection sensitivity hypothesis because the justifications for targets and lures most reflective of accuracy (Tables 2 and 4) clearly reflect claims of recollection. The present study also extends this interpretation by implicating a second basis of judgment that is unique to the MST and that is also reliably conveyed via language, namely, the detection of changes in test items enabled by the recollection of criterial studied item features. This element of change detection is conveyed by the justifications for correct lure responses, with participants verbalizing features that deviated between study and test items. Conversely, the language supporting correct foil responses was characterized by the absence of recollective experience and, sometimes, remembrance of any item of the same kind. Critically, these distinctions regarding subjective experience are qualitative or multidimensional in that detailed remembrance, the detection of change with respect to remembrance and comparison, and the absence of remembrance altogether are linguistically distinct. Nonetheless, because they can all lead to numerically confident decisions, it means that confidence per se is an inherently impoverished indicator of the basis of judgment if it is assumed to reflect a one-dimensional strength of evidence value. We elaborate on this point in the General Discussion section.

The final aspect of the language analysis worth discussing is the interaction between confidence- and language-based predictions for correct responses to lures and targets (Table 5 and Figure 6). This interaction has been noted in eyewitness research (Seale-Carlisle et al., 2022), but this is the first time it has been shown in the MST. Such an interaction is not surprising if one assumes that recollection, and/or recollection accompanied by the detection of change, influences confidence ratings. Under dual-process models of recognition, this type of assumption is standard, namely, that successful recollection yields high confidence because it situates the item within a specific context, affording more behavioral control than feelings of familiarity or fluency (Dobbins et al., 1998; Jacoby, 1991; Yonelinas, 2002). Hence, from this perspective, language classifiers improved when confidence increased because confidence increased due to conscious recollective experiences and, when relevant, the conscious detection of change with respect to those experiences. Because these are categorically distinct, multidimensional experiences, a one-dimensional confidence scale will poorly discriminate them. However, confidence ratings may capture experiences that are difficult to verbalize such as familiarity and fluency. If those information sources contribute to recognition responses, they may be captured better by numeric confidence than justification language. This idea will need to be explored.

General Discussion

In two experiments, we collected verbal justifications of recognition responses in a popular mnemonic discrimination task—the MST—to assess the qualitative bases for such responses. Correct rejections of similar lures have been portrayed as behavioral manifestations of a hippocampal pattern separation process that encodes events distinctively (Stark et al., 2019). However, lure rejections are not process pure because participants can also make those decisions by detecting how features changed from studied items (Norman & O'Reilly, 2003). Here, the language that participants used to describe lure rejections indicated the use of a recollection rejection strategy in which lures were identified as such by comparing their features to retrieved memories of studied items. This finding was generalizable in that verbal justifications strongly predicted item classification accuracy using three- and two-response option MSTs with the latter also including numeric confidence ratings. Comparisons of the predictive utility of verbal justifications and confidence ratings showed that the former were superior in capturing qualitative differences in bases for responses across lures, targets, and foils. Notably, such differences could not be captured by a one-dimensional signal detection model of MST performance (cf. Loiotile & Courtney, 2015). However, verbal justifications and confidence ratings interacted in showing that language evincing recollection was associated with higher confidence. These results

suggest that language and confidence together may serve to more fully characterize the roles of conscious recollection and automatic familiarity in responses to similar lures. We elaborate on the theoretical implications of these findings in what follows.

Mechanisms Underlying Decisions in Mnemonic Discrimination Tasks

Mnemonic discrimination tasks have enjoyed popularity in human memory research because of their suitability for examining the cognitive and neural mechanisms underlying the recognition of current perceptions that are similar but not identical to the existing memories of earlier events. These tasks have been used to test dual-process models of memory, proposing a distinction between conscious recollection and automatic familiarity (Hintzman & Curran, 1994), the mechanisms of perceptual false memories and group differences therein (Koutstaal et al., 1999), and hippocampal processes that support distinctive encoding (Kirwan & Stark, 2007). The flexibility of these tasks for addressing several memory problems has surely contributed to their broad application. However, such flexibility can undermine a coherent theoretical understanding when the contributions of particular constructs to lure responses are overemphasized. This issue arises in the cognitive neuroscience literature when lure rejections are used to examine hippocampal pattern separation that is supported by inputs from extrahippocampal cortical regions (Amer & Davachi, 2023), while the contribution of comparisons of lures with pattern-completed items from the study phase is less rigorously considered (Stark et al., 2019). Our goal here was to emphasize the contribution of recollection-based pattern completion to lure rejections. We did this by showing that verbalizations indicating that decision strategy predict correct lure rejections.

These findings join a collection of mnemonic discrimination studies that used another subjective report method to identify the qualitative bases for lure rejections. Three studies incorporated remember/know judgments (Tulving, 1985) into MSTs by asking participants to report whether they classified lures by comparing them with memories of studied items based on recollection using "remember" judgments or familiarity using "know" (or "familiar") judgments (Kim & Yassa, 2013; Szőllősi et al., 2020; Wahlheim et al., 2024). Two of those studies showed that lure rejections were more strongly associated with recollection- than familiarity-based retrieval of studied items (Kim & Yassa, 2013; Wahlheim et al., 2024) and that self-reported attention during the encoding of studied objects was associated with recollection-based rejections (Wahlheim et al., 2024). However, the other study, which used a longer study–test delay (10 min) and provided a "guess" response option (Szőllősi et al., 2020), showed that lure rejections were more strongly associated with familiarity-based retrievals of studied items. The shift from recollection- to familiarity-based lure rejections likely reflected the loss of recollection over time. Taken with these prior findings, the present results suggest that recollection rejection is used for lure responses and that recollection rejection depends on the accessibility of studied items. The language describing familiarity-based rejections is unclear because comparing fuzzy memories to perceptions should offer little information about whether features changed. Familiarity-based judgments may instead reflect an intermediate strength signal evoked by test items, as assumed by one-dimensional signal detection accounts (cf. Loiotile & Courtney, 2015). Future studies could illuminate the subjective experiences of familiarity states associated with lure rejections by collecting verbal justifications after those judgments in an MST that does less to promote the use of recollection-based lure rejection, perhaps by reducing study time and increasing study-test retention intervals.

In this vein, studies could also assess the contributions of varying kinds of recollection and familiarity to lure responses using a dual-process signal detection model suited for mnemonic discrimination tasks with similar lures (Parks et al., 2025). Like its predecessors, the lure-optimized dual-process signal detection model estimates process contributions from ROCs derived from old/new recognition decisions and response confidence. The advantage of the model is that, instead of only estimating traditional recollection and familiarity contributions to target classifications, it estimates the contributions of recollection rejection, false recollection, and false familiarity to lure responses. One could apply this model to data from experiments akin to our Experiment 2 while including more verbal justifications to determine whether there is correspondence between recollection rejection language and parameter estimates. Ideal experiments would provide enough observations to power correlations between these measures while maintaining participant engagement and include manipulations that promote the use of recollection rejection, such as instructions to use a retrieval-monitoring strategy (e.g., Rotello et al., 2000).

The present study also showed that the justifications that predicted lure rejections were qualitatively different from those that predicted target recognition. Specifically, whereas justifications for lure rejections implicated recollection of studied items to reject lures, justifications for target endorsements implicated recollection to accept targets. These findings are related to the idea that a common recollection process underlies both judgments (Brainerd et al., 2003). This has been supported by behavioral findings showing similar effects of experimental manipulations that are assumed to selectively affect recollection (e.g., Jones, 2005; Matzen et al., 2011; Odegard & Lampinen, 2005; Odegard et al., 2008). However, contrary to this behavioral evidence, fMRI work has shown activation in distinct brain regions underlying each type of recollection, with target recollection activating the bilateral hippocampus, medial prefrontal regions, and the posterior cingulate, and recollection rejection uniquely activating left-lateralized prefrontal regions but also activating medial prefrontal regions to a lesser extent (Bowman & Dennis, 2016). These results suggest that recollection in lure and target recognition reflects different mechanisms, with rejections engaging monitoring-associated regions (Gallo et al., 2006, 2010).

However, such a conclusion is risky because the slow temporal resolution of the fMRI blood-oxygenation-level-dependent response may mask commonality in early neural activity in tasks that differ in later neural activity. For example, if recollection provides an impetus for affirmative recognition responding generally, then the detection of change enabled by recollection would require an inhibitory response to reject the item as having been studied, even though it triggered a veridical recollection of the study experience. This appears consistent with the behavioral data in Bowman and Dennis (2016) because recollection-based endorsements were faster than rejections based on change detection (1,354 ms vs. 1,744 ms). Under this interpretation, the detection of regions linked to recollection during the rejection contrast may be quite difficult because the blood-oxygenation-level-dependent signal linked to recollection is necessarily temporally conflated with that linked to both change detection and the processes guiding behavioral inhibition. Given this, methods with greater

temporal precision may be required to characterize the temporal sequelae of neural responses underlying successful recollection rejection responses. One could begin to address this problem using MST variants in combination with electroencephalography-based analyses to temporally isolate the retrieval processes that evoke representations of studied items from the comparison process that leads to lure rejections.

Implications for Signal Detection Theory

Experiment 2 in the present study showed that justification language could be more effectively used than the combination of test item judgment (old or new) and numeric confidence to identify which type of stimulus participants were evaluating on each trial. However, it should be noted that the signal detection model was one-dimensional (Figure 7), and more complex multidimensional signal detection models are possible. For example, there are detection theoretic models that place recollection or source memory on a different dimension of strength than perceived familiarity or novelty (Banks, 2000; Rotello et al., 2004), and a similar approach is possible with the MST. Thus, what the current findings demonstrate is not that detection theory in general is insufficient, but that the conceptualization of targets, lures, and foils in the MST as falling along a single dimension of memory strength that reflects how well they match the encoding experience, is insufficient. Indeed, others have argued that more complex approaches using, for example, multinomial processing trees (Lee & Stark, 2023; Villarreal et al., 2022) or evidence accumulation models (Banavar et al., 2024) are more reasonable than a one-dimensional memory strength model for describing the mechanisms underlying lure responses in the MST.

Language to Process Mappings

Previous work using language classifiers has shown that recognition success is predicted by the presence of language conveying recollective experience (e.g., Dobbins & Kantner, 2019). The present findings show that another recollection-linked process is detectable in language, namely, the detection of changed lure features enabled by recollection of studied items (viz., recollection rejection). Here, the reported experience of remembering is augmented by a claim of change, and the latter is clearly reflected in the language. This raises the question of how many distinct memory-linked processes one might capture using language classification approaches, which at this stage is speculative. At a minimum, the language classification approach requires that the process in question be (a) a conscious basis for judgment, (b) verbalizable, and (c) verbalized consistently across people.

To be concrete, consider a process not yet examined via this approach referred to as subjective memorability or the distinctiveness heuristic (Brown et al., 1977; Dobbins & Kroll, 2005; Dodson & Schacter, 2001). This process is relevant here because, like recollection rejection, it can be a basis for confidently rejecting an item as studied. However, it is quite different. Under subjective memorability rejections, participants reject items that do not yield recollections, but given their personal distinctiveness, the participant believes the items would have yielded recollections if studied. For example, if one encounters a family member's name in a test list, the item could be rejected because it would have been recollected if it had actually been studied, resulting in confident rejection because of the absence of recollection. Justification of this response should contain markers of this process, including claims that the item would have been remembered along with why this is believed (i.e., why the item is personally distinctive). Such language is distinct from accepting an item based on recollection or rejecting an item based on a detected change enabled by a recollection. Interestingly, all three of these putative processes refer to recollection but in different ways, namely, its sufficiency (recollection acceptance), its conflict with current perception (recollection rejection), or its unlikely absence (subjective memorability rejection). Future research using the language classification approach may serve as a fruitful way of testing these theoretical distinctions that have often relied on more indirect means of assessment.

Constraints on Generality

The present verbal justification results suggest that recognition responses in the MST reflect qualitatively distinct bases. However, some constraints on generality are worth noting. First, we required justifications on a subset of trials, which may have altered response strategies compared with standard MSTs. However, such reactivity was likely minimal because the response patterns (Figures 2 and 4A) closely replicate prior MST outcomes without verbal justifications (e.g., Stark et al., 2013, 2015). Second, the feature overlap between targets and lures determined recognition accuracy, but response rates might vary with other stimuli. We chose stimuli with moderate perceptual overlap between targets and lures to distribute justification prompts across key response types for targets and lures (Tables 1 and 3). However, more confusable stimuli, such as target and lure words with high semantic similarity or other-race perpetrator and filler faces with similar features, might cause higher error rates. This could undermine the diagnosticity of verbal strategies for predicting recognition accuracy. However, we did not find any preliminary evidence for this because lure confusability did not significantly moderate the predictive accuracy of language classifiers (see Supplemental Section S1).

The language classification approach to characterizing the qualitative bases of recognition responses is effective, but its boundary conditions are unclear. The clear predictive advantage of language over confidence likely requires that the task evoke more than two qualitatively distinct subjective memory experiences. In such cases, a one-dimensional confidence scale may fail to capture categorical differences, as was evident here, where participants invoked at least two distinct forms of recollection. The advantage of language classification is therefore expected to diminish when subjective memory strength is more homogeneous or not easily verbalized, perhaps when familiarity-based responses are prominent. Finally, we did not examine individual differences in justification language or classifier performance. However, prior work using recognition memory tasks without similar lures has shown that language classifiers are more accurate for more skilled recognizers (Dobbins & Kantner, 2019; Seale-Carlisle et al., 2022). This pattern may hold even as lures become less discriminable from targets, as in the MST. Moreover, it remains an open question how cognitive abilities, traits, or other variables such as age will affect verbal justifications. Future research should explore such moderators to characterize the generality of this language-based approach.

Conclusion

In conclusion, the present study examined the extent to which verbal justifications of similar lure responses in the MST could reveal the bases for such responses. Consistent with findings using other measurement approaches, we showed that correct rejections of similar lures were primarily based on a recollection rejection strategy. This basis for responding was qualitatively distinct from the recollective basis used for accepting studied items, as both included language about recollecting studied items, but the language of rejections referenced the detection of changed features. Language classifiers were better able to predict lure classifications than an approach analogous to one-dimensional signal detection models, thus confirming that the latter approach is insufficient for capturing recollection rejection. Indeed, verbal justifications predicted classifications better than numeric confidence. However, an interaction between the two suggests that some combination of measures could be used to more fully characterize the bases for lure responses. A comprehensive understanding of the processes evoked by similar lures will require a combination of verbal justifications, multidimensional modeling, and temporally precise behavioral and neuroimaging measures. This converging methods approach promises to precisely distinguish between recollection in the service of acceptance versus rejection and between rejections based on the presence versus absence of recollection.

References

Agresti, A. (2012). *Categorical data analysis* (2nd ed., [Nachdr.]). Wiley.

Amer, T., & Davachi, L. (2023). Extra-hippocampal contributions to pattern separation. *eLife*, *12*, Article e82250. https://doi.org/10.7554/eLife.82250

Banavar, N. V., Noh, S. M., Wahlheim, C. N., Cassidy, B. S., Kirwan, C. B., Stark, C. E. L., & Bornstein, A. M. (2024). A response time model of the three-choice Mnemonic Similarity Task provides stable, mechanistically interpretable individual-difference measures. *Frontiers in Human Neuroscience*, *18*, Article 1379287. https://doi.org/10.3389/fnhum.2024.1379287

Banks, W. P. (2000). Recognition and source memory as multivariate decision processes. *Psychological Science*, *11*(4), 267–273. https://doi.org/10.1111/1467-9280.00254

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), Article 774. https://doi.org/10.21105/joss.00774

Bowman, C. R., & Dennis, N. A. (2016). The neural basis of recollection rejection: Increases in hippocampal–prefrontal connectivity in the absence of a shared recall-to-reject and target recollection network. *Journal of Cognitive Neuroscience*, *28*(8), 1194–1209. https://doi.org/10.1162/jocn_a_00961

Brainerd, C. J., Reyna, V. F., Wright, R., & Mojardin, A. H. (2003). Recollection rejection: False-memory editing in children and adults. *Psychological Review*, *110*(4), 762–784. https://doi.org/10.1037/0033-295X.110.4.762

Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, word frequency and negative recognition. *The Quarterly Journal of Experimental Psychology*, *29*(3), 461–473. https://doi.org/10.1080/14640747708400622

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*(2), 186–205. https://doi.org/10.1037/1082-989X.3.2.186

Dobbins, I. G. (2022). Recognition language classifiers demonstrate far transfer of learning. *Psychonomic Bulletin & Review*, *29*(4), 1414–1425. https://doi.org/10.3758/s13423-022-02085-1

Dobbins, I. G., & Kantner, J. (2019). The language of accurate recognition memory. *Cognition*, *192*, Article 103988. https://doi.org/10.1016/j.cognition.2019.05.025

Dobbins, I. G., & Kroll, N. E. A. (2005). Distinctiveness and the recognition mirror effect: Evidence for an item-based criterion placement heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1186–1198. https://doi.org/10.1037/0278-7393.31.6.1186

Dobbins, I. G., Kroll, N. E. A., Yonelinas, A. P., & Liu, Q. (1998). Distinctiveness in recognition and free recall: The role of recollection in the rejection of the familiar. *Journal of Memory and Language*, *38*(4), 381–400. https://doi.org/10.1006/jmla.1997.2554

Dobolyi, D. G., & Dodson, C. S. (2018). Actual vs. perceived eyewitness accuracy and confidence and the featural justification effect. *Journal of Experimental Psychology: Applied*, *24*(4), 543–563. https://doi.org/10.1037/xap0000182

Dodson, C. S., & Schacter, D. L. (2001). "If I had said it I would have remembered it": Reducing false memories with a distinctiveness heuristic. *Psychonomic Bulletin & Review*, *8*(1), 155–161. https://doi.org/10.3758/BF03196152

Fox, J., & Weisberg, S. (2019). *An {R} companion to applied regression* (3rd ed.). Sage Publications.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition*, *38*(7), 833–848. https://doi.org/10.3758/MC.38.7.833

Gallo, D. A., Kensinger, E. A., & Schacter, D. L. (2006). Prefrontal activity and diagnostic monitoring of memory retrieval: fMRI of the criterial recollection task. *Journal of Cognitive Neuroscience*, *18*(1), 135–148. https://doi.org/10.1162/089892906775250049

Gallo, D. A., McDonough, I. M., & Scimeca, J. (2010). Dissociating source memory decisions in the prefrontal cortex: fMRI of diagnostic and disqualifying monitoring. *Journal of Cognitive Neuroscience*, *22*(5), 955–969. https://doi.org/10.1162/jocn.2009.21263

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

Grabman, J. H., Dobbins, I. G., & Dodson, C. S. (2024). Comparing human evaluations of eyewitness statements to a machine learning classifier under pristine and suboptimal lineup administration procedures. *Cognition*, *251*, Article 105876. https://doi.org/10.1016/j.cognition.2024.105876

Grabman, J. H., Dobolyi, D. G., Berelovich, N. L., & Dodson, C. S. (2019). Predicting high confidence errors in eyewitness memory: The role of face recognition ability, decision-time, and justifications. *Journal of Applied Research in Memory and Cognition*, *8*(2), 233–243. https://doi.org/10.1037/h0101835

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language*, *33*(1), 1–18. https://doi.org/10.1006/jmla.1994.1001

Hunsaker, M. R., & Kesner, R. P. (2013). The operation of pattern separation and pattern completion processes associated with different attributes

or domains of memory. *Neuroscience and Biobehavioral Reviews*, *37*(1), 36–58. https://doi.org/10.1016/j.neubiorev.2012.09.014

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541. https://doi.org/10.1016/0749-596X(91)90025-F

Jones, T. C. (2005). Study repetition and the rejection of conjunction lures. *Memory*, *13*(5), 499–515. https://doi.org/10.1080/09658210444000197

Kim, J., & Yassa, M. A. (2013). Assessing recollection and familiarity of similar lures in a behavioral pattern separation task. *Hippocampus*, *23*(4), 287–294. https://doi.org/10.1002/hipo.22087

Kirwan, C. B., & Stark, C. E. L. (2007). Overcoming interference: An fMRI investigation of pattern separation in the medial temporal lobe. *Learning & Memory*, *14*(9), 625–633. https://doi.org/10.1101/lm.663507

Koutstaal, W., Schacter, D. L., Galluccio, L., & Stofer, K. A. (1999). Reducing gist-based false recognition in older adults: Encoding and retrieval manipulations. *Psychology and Aging*, *14*(2), 220–237. https://doi.org/10.1037/0882-7974.14.2.220

Lee, M. D., & Stark, C. E. L. (2023). Bayesian modeling of the Mnemonic Similarity Task using multinomial processing trees. *Behaviormetrika*, *50*(2), 517–539. https://doi.org/10.1007/s41237-023-00193-3

Lenth, R. (2021). *emmeans: Estimated marginal means, aka least-squares means* (Version 1.6.3) [Computer software]. https://cran.r-project.org/web/packages/emmeans/index.html

Liu, K. Y., Gould, R. L., Coulson, M. C., Ward, E. V., & Howard, R. J. (2016). Tests of pattern separation and pattern completion in humans—A systematic review. *Hippocampus*, *26*(6), 705–717. https://doi.org/10.1002/hipo.22561

Loiotile, R. E., & Courtney, S. M. (2015). A signal detection theory analysis of behavioral pattern separation paradigms. *Learning & Memory*, *22*(8), 364–369. https://doi.org/10.1101/lm.038141.115

Lüdecke, D. (2024). *sjPlot: Data visualization for statistics in social science* (Version R package Version 2.8.17) [Computer software]. https://CRAN.R-project.org/package=sjPlot

Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences*, *262*(841), 23–81. https://doi.org/10.1098/rstb.1971.0078

Matzen, L. E., Taylor, E. G., & Benjamin, A. S. (2011). Contributions of familiarity and recollection rejection to recognition: Evidence from the time course of false recognition for semantic and conjunction lures. *Memory*, *19*(1), 1–16. https://doi.org/10.1080/09658211.2010.530271

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, *110*(4), 611–646. https://doi.org/10.1037/0033-295X.110.4.611

Odegard, T. N., Koen, J. D., & Gama, J. M. (2008). Process demands of rejection mechanisms of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1296–1304. https://doi.org/10.1037/a0013034

Odegard, T. N., & Lampinen, J. M. (2005). Recollection rejection: Gist cuing of verbatim memory. *Memory & Cognition*, *33*(8), 1422–1430. https://doi.org/10.3758/BF03193375

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus*, *4*(6), 661–682. https://doi.org/10.1002/hipo.450040605

Parks, C. M., Yonelinas, A. P., & Wahlheim, C. N. (2025). *Towards a unified theory of memory for similar episodes*. https://doi.org/10.31234/osf.io/mupqy_v1

Psychology Software Tools. (2016). *E-Prime 3.0* [Computer software]. https://www.pstnet.com

R Core Team. (2024). *R software*. R Foundation for Statistical Computing. https://www.R-project.org/

Reagh, Z. M., Noche, J. A., Tustison, N. J., Delisle, D., Murray, E. A., & Yassa, M. A. (2018). Functional imbalance of anterolateral entorhinal cortex and hippocampal dentate/CA3 underlies age-related object pattern separation deficits. *Neuron*, *97*(5), 1187–1198.e4. https://doi.org/10.1016/j.neuron.2018.01.039

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), Article 77. https://doi.org/10.1186/1471-2105-12-77

Rotello, C. M., Macmillan, N. A., & Reeder, J. A. (2004). Sum-difference theory of remembering and knowing: A two-dimensional signal-detection model. *Psychological Review*, *111*(3), 588–616. https://doi.org/10.1037/0033-295X.111.3.588

Rotello, C. M., Macmillan, N. A., & Van Tassel, G. (2000). Recall-to-reject in recognition: Evidence from ROC curves. *Journal of Memory and Language*, *43*(1), 67–88. https://doi.org/10.1006/jmla.1999.2701

Seale-Carlisle, T. M., Grabman, J. H., Dobolyi, D. G., & Dodson, C. S. (2025). A comparison between numeric confidence ratings and verbal confidence statements. *Journal of Experimental Psychology: Applied*, *31*(1), 12–39. https://doi.org/10.1037/xap0000525

Seale-Carlisle, T. M., Grabman, J. H., & Dodson, C. S. (2022). The language of accurate and inaccurate eyewitnesses. *Journal of Experimental Psychology: General*, *151*(6), 1283–1305. https://doi.org/10.1037/xge0001152

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

Stark, S. M., Kirwan, C. B., & Stark, C. E. L. (2019). Mnemonic smilarity task: A tool for assessing hippocampal integrity. *Trends in Cognitive Sciences*, *23*(11), 938–951. https://doi.org/10.1016/j.tics.2019.08.003

Stark, S. M., Stevenson, R., Wu, C., Rutledge, S., & Stark, C. E. L. (2015). Stability of age-related deficits in the mnemonic similarity task across task variations. *Behavioral Neuroscience*, *129*(3), 257–268. https://doi.org/10.1037/bne0000055

Stark, S. M., Yassa, M. A., Lacy, J. W., & Stark, C. E. L. (2013). A task to assess behavioral pattern separation (BPS) in humans: Data from healthy aging and mild cognitive impairment. *Neuropsychologia*, *51*(12), 2442–2449. https://doi.org/10.1016/j.neuropsychologia.2012.12.014

Szőllősi, Á., Bencze, D., & Racsmány, M. (2020). Behavioural pattern separation is strongly associated with familiarity-based decisions. *Memory*, *28*(3), 337–347. https://doi.org/10.1080/09658211.2020.1714055

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Treves, A., & Rolls, E. T. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus*, *4*(3), 374–391. https://doi.org/10.1002/hipo.450040319

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie canadienne*, *26*(1), 1–12. https://doi.org/10.1037/h0080017

Umanath, S., & Coane, J. H. (2020). Face validity of remembering and knowing: Empirical consensus and disagreement between participants and researchers. *Perspectives on Psychological Science*, *15*(6), 1400–1422. https://doi.org/10.1177/1745691620917672

Villarreal, M., Stark, C. E. L., & Lee, M. D. (2022). Adaptive design optimization for a Mnemonic Similarity Task. *Journal of Mathematical Psychology*, *108*, Article 102665. https://doi.org/10.1016/j.jmp.2022.102665

Wahlheim, C. N., Dobbins, I. G., & Wellons, B. M. (2025, August 1). *The natural language of mnemonic discrimination in visual object recognition*. https://osf.io/g5t9a

Wahlheim, C. N., Garlitch, S. M., Mohamed, R. M., & Weidler, B. J. (2024). Associations among attentional state, retrieval quality, and mnemonic discrimination. *Journal of Memory and Language*, *139*, Article 104554. https://doi.org/10.1016/j.jml.2024.104554

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), Article 1686. https://doi.org/10.21105/joss.01686

Wilson, I. A., Gallagher, M., Eichenbaum, H., & Tanila, H. (2006). Neurocognitive aging: Prior memories hinder new hippocampal encoding. *Trends in Neurosciences*, *29*(12), 662–670. https://doi.org/10.1016/j.tins.2006.10.002

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*(3), 441–517. https://doi.org/10.1006/jmla.2002.2864

Received April 8, 2025
Revision received August 1, 2025
Accepted September 29, 2025 ■