Recognition Receiver Operating Characteristic Curves: The Complex Influence of Input Statistics, Memory, and Decision-making

Olya Hakobyan and Sen Cheng

Abstract

Receiver operating characteristic (ROC) analysis is the standard tool for studying recognition memory. In particular, the curvilinearity and the \( y \)-offset of recognition ROC curves have been interpreted as indicative of either memory strength (single-process models) or different memory processes (dual-process model). The distinction between familiarity and recollection has been widely studied in cognitive neuroscience in a variety of conditions, including lesions of different brain regions. We develop a computational model that explicitly shows how performance in recognition memory is affected by a complex and, as yet, underappreciated interplay of various factors, such as stimulus statistics, memory processing, and decision-making. We demonstrate that (1) the factors in the model affect recognition ROC curves in unexpected ways, (2) fitting R and F parameters according to the dual-process model is not particularly useful for understanding the underlying processes, and (3) the variability of recognition ROC curves and the controversies they have caused might be due to the uncontrolled variability in the contributing factors. Although our model is abstract, its functional components can be mapped onto brain regions, which are involved in corresponding functions. This enables us to reproduce and interpret in a coherent framework the diverse effects on recognition memory that have been reported in patients with frontal and hippocampal lesions. To conclude, our work highlights the importance of the rich interplay of a variety of factors in driving recognition memory performance, which has to be taken into account when interpreting recognition ROC curves.

INTRODUCTION

Recognition memory is the process that allows a participant to indicate whether they have experienced a stimulus before in a preceding study phase. Over the years, a number of different experimental paradigms have been established for studying recognition memory, including the forced-choice task (Bayley, Wixted, Hopkins, & Squire, 2008), remember/know responses (Wixted, 2009), yes/no task (Yonelinas & Parks, 2007), and spontaneous recognition in rodents (Warburton & Brown, 2015). A common practice in the study of recognition memory is to plot the hit rate against the false alarm rate for different response biases (from conservative to liberal), resulting in the receiver operating characteristic (ROC) curve (Green & Swets, 1966). Signal detection theory applied to recognition memory predicts a curvilinear ROC curve that intersects the \( y \) axis at the origin, that is, both low hit and low false alarm rates when the criterion is very conservative. However, experimental recognition ROC curves exhibit a nonzero \( y \) intercept, that is, a significant hit rate at a vanishing false alarm rate (Yonelinas & Parks, 2007). The shape of the recognition ROC curve and its changes in response to experimental manipulations have been discussed extensively using different assumptions about the memory parameters involved. Here, we study how different components of cognition, including perception, decision-making, and memory, can affect recognition judgments, showing that the ROC features are unlikely to reflect process-pure memory-specific parameters.

The dominant account of recognition ROC curves in cognitive neuroscience is arguably the dual-process model. It proposes that the curvilinearity and the \( y \) intercept of recognition ROC curves reflect two qualitatively different memory processes, called familiarity and recollection (Yonelinas, Aly, Wang, & Koen, 2010; Eichenbaum, Yonelinas, & Ranganath, 2007). By contrast, single-process models reject the existence of different memory processes and propose that both curvilinearity and \( y \)-offset are influenced by memory strength (Squire, Wixted, & Clark, 2007). Despite much criticism (Rotello, 2017; Dunn, 2008; Wixted, 2007), many studies in human cognitive neuroscience adopt the dual-process model when interpreting the changes of ROC features caused by development, lesions, and disorders, such as schizophrenia, autism, and Alzheimer disease.

However, both single- and dual-process interpretations of the ROC curves suffer from the same problem: They strongly focus on the properties of the underlying memory system(s) and do not explicitly include other cognitive processes involved in the recognition task and their interaction with other involved components. For instance, consider an experiment where a certain parameter, such as the encoding condition, is manipulated. The resulting outcomes are typically interpreted either as changes in the memory strength (single-process models) or changes in
mnemonic processes, such as familiarity and recollection (dual-process models) evoked by the experimental condition. However, it may well be that the manipulations affect other cognitive processes, such as the way the stimuli are processed or how the decision is made in addition to, or instead of, changes in the mnemonic component. Therefore, a neat mapping of the ROC features to familiarity and recollection components or to memory strength may not be sufficient to understand the more complex interaction between these underlying processes. This proposition might appear strange when discussing recognition “memory”; however, the integrative aspect of memory has been highlighted a long time ago (Atkinson & Shiffrin, 1968), and mounting evidence shows that non-mnemonic factors such as the stimulus material (Taylor, Henson, & Graham, 2007; Cipolotti et al., 2006; Gehring, Toglia, & Kimble, 1976), sensory representation (Cowell, Bussey, & Saksida, 2010), and decision-making (Frank & Hicks, 2016; Aminoff et al., 2012) significantly influence recognition memory. In fact, fitting experimental data with memory-centered models results in systematic residual errors (Kellen & Singmann, 2016), highlighting the need for more careful assumptions about underlying processes. Factors such as sensory input or decision-making are not thought to affect the shape of the ROC curves directly, without changing the mnemonic parameters. However, their role may be more substantial when the assumptions about memory processes are relaxed.

The Role of Stimulus Statistics in Recognition Memory Tests

The relevance of the stimulus representations on the subsequent memory performance has been recognized a long time ago (Nosofsky, 1992; Gehring et al., 1976), and recent developments aim to pin down the specific features and neural correlates of image and word memorability (Xie, Bainbridge, Inati, Baker, & Zaghloul, 2020; Mohsenzadeh, Mullin, Oliva, & Pantazis, 2019; Bainbridge, 2017; Bainbridge, Isola, & Oliva, 2013). In recognition memory studies, the researchers typically try to control for stimulus saliency by making them as homogeneous as possible, for example, by removing verbal cues or outstanding features in pictures. Nevertheless, in some conditions the target–lure similarity is deliberately increased, leading to impaired memory performance. In false memory paradigms, healthy participants report semantically related lures more often than unrelated lures (Chadwick et al., 2016; Roediger & McDermott, 1995; Deese, 1959). Similarly, semantic or orthographic similarity between targets and lures leads to compromised performance in item recognition tests (Heathcote, 2003). Amnesic patients perform poorly on recognition tests when the targets and lures are similar (Bayley et al., 2008; Westerberg et al., 2006; Holdstock et al., 2002), but do about as well as controls when targets are distinct (Turriandiziano, Fadda, Caltagirone, & Carlesimo, 2004; Mayes, Holdstock, Isaac, Hunkin, & Roberts, 2002; Vargha-Khadem et al., 1997). Furthermore, higher inter-stimulus similarity seems to affect recognition performance, giving rise to small list length and list strength effects (Osth, Dennis, & Kinnell, 2014; Kinnell & Dennis, 2012).

Some stimuli seem to result in better performance when amnesic patients are tested. For example, in the study of Cipolotti et al. (2006), the ROC curve of the amnesic patients was comparable to the controls when face stimuli were used. This outcome was interpreted as a selective sparing of the familiarity and recollection processes for facial stimuli. “A receiver operating characteristic analysis revealed that VC’s recollective and familiarity-based recognition processes were well preserved for unknown human faces. In contrast, recollective-based-recognition for verbal and topographical material was at floor.” However, they (the results) also suggest the remarkable and selective preservation of a particular type of non-verbal memory dedicated to faces”. By contrast, some patients with lesioned medial-temporal lobe (MTL) but spared hippocampus showed unimpaired recognition for scenes, but not for other materials (Lacot et al., 2017). Another patient with preserved hippocampus and impaired perirhinal and entorhinal cortices showed no impairment with visual stimuli but compromised performance when tested with aurally presented pseudowords (Martin, Bowles, Mirsattari, & Köhler, 2011). These results, too, are interpreted along the lines of recollection and familiarity processes.

In a similar vein, several studies report more linear ROC curves when switched plurality lures are used (Buratto & Lamberts, 2008; Arndt & Reder, 2002; Norman, 2002; Rotello, Macmillan, & Van Tassel, 2000). It is typically assumed that this task renders familiarity useless, and thus, the linearity of the ROC curves is interpreted as an absence of familiarity and exclusive reliance on recollection. In other words, different inputs are assumed to recruit different memory processes and thereby influence recognition. However, target–lure similarity and its interaction with the decision process may also influence the shape of the ROC curves. We next discuss evidence of such criterion effects in recognition studies.

Criterion Effects in Recognition Experiments

There is a growing interest in dissociating the decision-making systems in recognition memory from the function of memory modules (Bastin et al., 2019; Cox & Shiffrin, 2017; Turner, Van Zandt, & Brown, 2011) and examining the role of different experimental conditions on criterion placement. There are multiple reports of between-list criterion shifts, for example, when strengthened lists are compared with weak lists (Hirshman, 1995), when the lists are tested at different delay durations (Stretch & Wixted, 1998), or when the responses for the lists are biased toward a certain criterion based on experimenter feedback (Han & Dobbins, 2008). Within-list criterion shifts—for
example, when items from different encoding conditions are tested together—are less well established. A large number of studies reported no such effect (Verde & Rotello, 2007), whereas some others have observed within-list criterion shifts by strong explicit hints, that is, marking the items encoded in strong and weak conditions with different colors and screen locations at test (Franks & Hicks, 2016). Although it is clear that memory strength can influence criterion placement (at least in between-list comparisons), changes of response bias are assumed to have no effect on the ROC curves apart from influencing the placement of the data points along the curve. Similarly, the familiarity and recollection estimates from the dual-process model are supposed to be unaffected by criterion changes. Consequently, studies showing how criterion changes might affect the shape of the ROC curves (Benjamin, Diaz, & Wee, 2009; Mueller & Weidemann, 2008; Van Zandt, 2000) have gone relatively unnoticed, particularly in cognitive neuroscience.

### Anatomy of Recognition Memory

Selective hippocampal lesions have been reported to lead to a decrease in the intercept of the recognition ROC curve in human patients (Cipolotti et al., 2006; Wais, Wixted, Hopkins, & Squire, 2006; Yonelinas et al., 2002; Yonelinas, Kroll, Dobbins, Lazzara, & Knight, 1998) and rats (Sauvage, Fortin, Owens, Yonelinas, & Eichenbaum, 2008; Fortin, Wright, & Eichenbaum, 2004). When lesions extend beyond the hippocampus, both the intercept and curvilinearity of the ROC curves seem to be affected (Yonelinas et al., 2002). Even when the overall performance was matched between the lesion and control groups, the ROC curves of the lesion group exhibited no significant intercept in contrast to that of the control group (Fortin et al., 2004; Yonelinas et al., 1998). So, changes of the intercept in the lesion group were not simply attributable to the impairments of overall performance. Furthermore, patients with severe atrophy of the mammillary bodies exhibited a decreased intercept, whereas patients with mild atrophy did not (Vann et al., 2009).

The simple decrease of the intercept can be interpreted as the result of either impaired recollection in the dual-process model or weaker memory of the target items, that is, smaller variance ($\sigma$), in the single-process model. Some findings are, however, more difficult to reconcile with one model than the other. For instance, recognition ROC curves of amnesic patients look remarkably similar to those of controls if the study list was shortened from 50 to 10 items and a shorter retention interval was used (Wais et al., 2006). These findings appear to be more consistent with different memory strengths in lesion patients, rather than the lack of one of two qualitatively different processes. However, Yonelinas et al. (2010) objected: "[t]he results do not challenge the notion that familiarity is an equal-variance signal detection process, but they suggest that with very short lists, amnesic people can maintain and recollect two to three words". Amnesic patients could, in principle, recollect some items under more favorable conditions if the hippocampus is not entirely damaged.

On the other hand, it is unclear whether and how the single-process model could account for the result of unilateral lesions to the perirhinal cortex sparing the hippocampus. In one patient, the ROC curve has lower curvilinearity and a higher $y$ intercept (Bowles et al., 2007). This finding has been interpreted, within the dual-process model, as an impairment of the familiarity process, while recollection remains intact. However, lesions mostly constrained to the perirhinal cortex are rare, and to the best of our knowledge, only one recognition ROC curve has been reported for such patient. Caution should be exercised when interpreting single-case studies. The effect of perirhinal lesions on the shape of the ROC curve has not been reported for an animal model, even though perirhinal lesions have been extensively utilized in other paradigms, such as spontaneous recognition (Warburton & Brown, 2015; Winters, Forwood, Cowell, Saksida, & Bussey, 2004).

Studies in humans have shown that lesions of the lateral pFPC reduce the curvilinearity of the ROC curves and leave the intercept intact (Aly, Yonelinas, Kishiyama, & Knight, 2011; MacPherson et al., 2008). A rat study, however, reports the opposite result when the medial pFPC is lesioned (Farovik, Dupont, Arce, & Eichenbaum, 2008). In all cases, there is a larger difference between controls and frontal patients in the false alarm rates than in the hit rates. So, deficits in recognition memory manifest in frontal patients primarily as increased false alarm rates, whereas hit rates are comparable to those in healthy controls (Swick & Knight, 1999). By contrast, hippocampal patients exhibit relatively normal false alarm rates but lower hit rates. Finally, there are reports of more liberal response bias in frontal patients (Bastin, van der Linden, Lekeu, Andrés, & Salmon, 2006; Swick & Knight, 1999).

### Our Approach

Here, we propose a novel multicomponent model that views performance on recognition memory tasks as the result of multiple processes: First, stimuli are processed and represented by a perceptual system. Second, memories of these perceptual representations are stored in and retrieved from a generic memory module. Third, when a test cue is presented, the closest matching stored item is retrieved from memory, akin to the function of an attractor memory network. Fourth, a decision is made whether the retrieved item is similar enough to the cue based on a decision criterion. Fifth, decision criteria are derived from the distribution of target items. This model allows us to account for a wide range of experimental phenomena, propose alternative explanations, and make testable predictions. The empirical findings simulated in our work include the acquisition of typical recognition ROC curves with nonzero $y$ intercept and curvilinearity, the effects of stimulus statistics, meta-knowledge, list length, and...
list strength on recognition performance. In addition, we discuss parsimonious explanations of how the changes seen after hippocampal and frontal lesions might arise. Our model complements the recent, more integrative approaches to recognition memory (Bastin et al., 2019; Cox & Shiffrin, 2017) and ROC curves (Turner et al., 2011; Benjamin et al., 2009; Mueller & Weidemann, 2008; Van Zandt, 2000).

**METHODS**

**Modeling Framework**

We posit that there is no specialized memory system in the brain that implements recognition memory. Instead, we hypothesize that the task in recognition memory experiments engages a combination of several main components: sensory input, a visual representations system, a general-purpose memory module, and a decision-making module (Figure 1). Here, we suggest an abstract model of recognition memory to study the influence of factors such as stimulus statistics, memory fidelity, and criterion placement on recognition memory, as well as their mutual interactions (Figure 1). We discuss each of these components in more detail below and how features of our model can be mapped onto experimental parameters.

**Sensory Input**

The images used in the current study were drawn from the Chicago face database (Ma, Correll, & Wittenbrink, 2015). Half of the 160 available images were randomly selected as targets, and the remaining half as lures. Additionally, we created an additional set of related lures by morphing the target and lure images into each other to varying degrees (Figure 2A). In each simulation session, a specified number of items \( n \in \{30, 70\} \) was randomly drawn from target and lure images.

Structures in the MTL, including perirhinal cortex and hippocampus, receive highly processed sensory input from cortical areas. To incorporate a visual preprocessing stage in our model, we used the biologically inspired, feedforward HMAX network (Figure 2A), which was developed for object recognition. This network consists of alternating layers (denoted by colors) of simple cells and complex cells, which represent increasingly abstract visual features extracted from the input images. The simple cells calculate the weighted sum of the incoming input, whereas complex cells perform a softmax operation. The bypass route works the same way as the main route except it skips some of the intermediate layers for faster processing. The model has been used for classification and identification tasks; however, for our recognition task, we used the output of the S3 layer, because this layer yielded reasonable features for our stimuli.

**Figure 1.** Study hypothesis. Recognition ROC curves are shaped by multiple interacting factors.

**Figure 2.** Sample of input images and visual preprocessing network. (A) Related lures were created by pairing a target (left) with an unrelated lure (right) and morphing them into each other (intermediate). The numbers indicate the morph similarity to the target item in percent. (B) An illustration of the HMAX model (Serre & Riesenhuber, 2004; Riesenhuber & Poggio, 1999). The model consists of alternating layers (denoted by colors) of simple cells and complex cells, which represent increasingly abstract visual features extracted from the input images. The simple cells calculate the weighted sum of the incoming input, whereas complex cells perform a softmax operation. The bypass route works the same way as the main route except it skips some of the intermediate layers for faster processing. The model has been used for classification and identification tasks; however, for our recognition task, we used the output of the S3 layer, because this layer yielded reasonable features for our stimuli.
recognition and classification (Serre & Riesenhuber, 2004; Riesenhuber & Poggio, 1999). The pretrained HMAX network consists of alternating layers of simple and complex cells (Figure 2B). The simple cells are tuned to specific input patterns, whereas complex cells perform a pooling operation. This design results in increasingly abstract representations along the hierarchical layers. We used the responses from the S3 layer, which were highly processed and yet yielded sufficient detail to distinguish between the stimuli from the same category. Because the network generates high-dimensional representations of the images, we performed a principal component analysis to obtain reasonably sized input patterns to store in/retrieve from our memory model. After sorting the components according to the amount of variance that they account for, we selected the first six components that together accounted for at least 80% of the variance. These components formed the perceptual representation of the visual stimulus $x_i$.

## Memory Storage and Retrieval

The perceptual representations $x_i$ are then passed to a memory system with a scaling parameter $p$, which increases the interitem distance (Figure 3A):

$$y_i = p \times x_i$$

(1)

The memory representation of the target items $y_1, y_2, \ldots$ are encoded and stored in the memory module as follows. The patterns are modified by noise, resulting in corrupted memory patterns:

$$\tilde{y}_i = y_i + \epsilon_i$$

(2)

The components of the $n$-dimensional noise vector $\epsilon$ are drawn independently from a normal distribution with zero mean and variance $\omega \in [0.15, 0.8]$.

During retrieval, test items $y'$ are presented, which are drawn from either the set of targets (represented by $y_1, y_2, \ldots$) or lures (represented by $z_1, z_2, \ldots$). The distance $D$ between the retrieval cue and all stored patterns in the memory module is calculated, and the pattern $\tilde{y}$ with the smallest distance to the retrieval cue is returned (Figure 3B):

$$i = \arg \min_y D(\tilde{y}, y')$$

(3)

$$\tilde{y}_i = \tilde{y}_i$$

(4)

This retrieval process is akin to retrieval from an attractor model of auto-associative memory (Fang, Rüther, Bellebaum, Wiskott, & Cheng, 2018; Treves & Rolls, 1994; Hopfield, 1982). The distance metric $D$ is not only relevant for retrieving the correct item but also for making recognition judgments later. Hence, before describing the decision-making process in the model, we briefly discuss our choice of the distance metric.

### Selection of Distance Measure

The most important feature of the memory module in our model is the scaling parameter, which increases the robustness against noise. Namely, when scaling is applied and the noise level is reasonable, the magnitude of the noise vector is small compared with the scales of the memory patterns, leading to low item interference (Figure 4A). This, in turn, increases memory accuracy. As expected, the proportion of correctly retrieved targets decreases with increased noise and increases with the scaling parameter for all three distance measures considered (Figure 4B): Euclidean distance, correlation, and cosine distance.

However, because memory judgments depend on the distance between the cue and the retrieved item, that is, memory precision, we next examined the distance between the targets and their noisy, stored version, $D(y_i, \tilde{y}_i)$, for the different distance measures. For the Euclidean distance measure, the distance is identical to the length of the noise vector, $\epsilon_i$, which is not scaled (Figure 4C). Hence, memory precision does not increase through scaling. By contrast, the correlation and cosine metrics are sensitive to the angle between the vectors instead of their magnitude. Therefore, for these metrics scaling leads to higher precision, that is, a smaller distance between the original and noisy, stored items (Figure 4C). This in turn increases the robustness.
against noise in memory storage and retrieval. For this reason, we consider the scaling parameter as a proxy for memory robustness. In our simulations, correlation and cosine metrics yielded qualitatively similar results. We adopted the correlation metric for all simulations reported here.

**Decision-making**

After an item is retrieved from memory, a recognition judgment is made. The model gives a “yes” response if the distance between the cue and the retrieved item $d_i$ is smaller than the decision threshold $\theta_j$; otherwise, a “no” response is given:

$$d_i = D(\hat{y}_i, y')$$

$$r_i = \begin{cases} 1 & \text{if } d_i \leq \theta_j + \lambda \\ 0 & \text{otherwise} \end{cases}$$

By varying the decision criterion $\theta_j$, we model different confidence levels in recognition memory to compute ROC curves. Because the judgments are made based on the $d_i$s and the distance distributions change depending...
on the conditions, such as noise and the scaling parameter, it is not reasonable to have a fixed set of criteria across conditions. Instead, we used the range model of criterion placement, where the decision criterion depends on the memory strength distribution (Hirshman, 1995). In our implementation, response thresholds were placed in the same manner for all conditions. Namely, for each test, the lower and higher boundaries are identified by calculating the minimum and maximum values of the target distance distribution after the removal of outliers (three standard deviations away from the mean). Next, the specified number of thresholds \( n = 5 \) are evenly spaced across this interval:

\[
\theta_j = \frac{\max(d_{\text{targets}}) - \min(d_{\text{targets}})}{n - 1} j + \min(d_{\text{targets}}) \quad (7)
\]

where \( j = 0, \ldots, n - 1 \). Finally, the most liberal criterion is added as another threshold \( \theta_n = \infty \) to ensure that the point \((1; 1)\) is part of the recognition ROC curve.

Because the lower boundary of the first interval, that is, \( \theta_0 \), is typically very small, only one or very few items could possibly pass the first threshold, making it unreliable to calculate hit rates. We therefore shift all criteria \( \theta_j \) by a small constant \( \lambda \in [0, 0.01] \), so that a reasonable number of items passed the first threshold. Unless indicated otherwise, the default value is \( \lambda = 0.006 \) in our simulations.

ROC Analysis

The curvilinearity and y intercept of the ROC curves are obtained by fitting our data to the dual-process signal detection (DPSD) model using the ROC toolbox (Koen, Barrett, Harlow, & Yonelinas, 2017). This choice does not imply that we adopt the dual-process model in our study in any way. We chose this model to obtain parameters that can be easily compared with many experimental results in recognition memory. In other words, even though we report the parameters R and F for our ROC curves, we do not necessarily consider nonzero values as evidence of recollection and familiarity processes at work. We view these parameters merely as useful estimates of the y intercept and curvilinearity of the ROC curves. In addition to the above-mentioned parameters, the ROC toolbox also provides estimations of general recognition performance, for example, area under the curve (AUC). Similar to experimental studies (Aggleton et al., 2005; Yonelinas et al., 2002), the F parameter was converted to the probability of a hit given a false alarm rate of 0.10 so that its magnitude is comparable to that of R. Maximum likelihood estimation is used for fitting, which minimizes the negative log likelihood.

RESULTS

Emergence of Typical Recognition ROC Curves in a Multicomponent Model

Our multicomponent model simulates the yes/no test of item recognition that is used frequently in studies with human participants. In the simulated study phase, the target items from the study list are encoded in the model’s memory, more specifically, noisy representations of the target items are stored in the memory module. In the simulated testing phase, the test items (targets or lures) are used to trigger memory retrieval from the memory module, which includes the comparison of the cue to all stored items. In analogy to attractor models of auto-associative memory, the stored item with the minimal distance to the cue is retrieved. Finally, the distance between the retrieved item and the cue is compared with a threshold for the recognition judgment (see the Methods section). The distribution of distances between the cue and the retrieved item depends on whether target items (target distribution) or lure items (lure distribution) are used as cue (Figure 5). The target distribution is shifted toward lower values as compared with the lure distribution, because noisy versions of the target items are stored in memory, whereas for the lure items, there are no corresponding stored items.

As noise increases, the target and lure distributions become more similar to each other (Figure 5B), especially when scaling is low or absent. Specifically, the mean and the variance of the target distribution increase (to a lesser extent such changes occur in the lure distribution as well). We conceptualize a decrease in the parameter \( \omega \) as an increase in memory strength, which might be caused by an experimental manipulation, such as decreasing the retention time (see Table 1). One way to counter the effect of memory noise in our model is to increase the scaling factor \( p \). For larger scaling parameters, the target distribution is narrower and closer to zero. We therefore propose that scaling improves the robustness of memory retrieval (see Table 1), a function that has been linked to memory-related structures in MTL. The criteria for deciding on the recognition response were adjusted to uniformly cover the target distance distribution according to Equation 6 (Figure 5, Columns 1–3, dashed lines). This set of decision criteria can be shifted to more liberal values by increasing the parameter \( \lambda \) in Equation 6 (Figure 5, gray dashed lines).

The combination of memory storage, memory retrieval, and decision-making produces a set of yes/no responses, which can be analyzed using standard ROC curves (Figure 5, Column 4). The model’s ROC curves are qualitatively similar to those obtained from experimental participants, showing pronounced curvilinearity and nonzero y-offset for certain parameter settings. The nonzero y-offset arises when the target distribution has a peak near zero, so that the number of target items passing the most conservative threshold is nonzero, while no lure item does. Our computational model allows us to study the effect of its various components on recognition memory, as well as their interplay in an integrated system.

The Effect of the Stimulus Statistics on Recognition Performance

We study the effect of the stimulus statistics on recognition memory by first examining the effects of target–lure
similarity. Because we use images as stimuli in our model, similar lures are represented by morphs that are similar to the targets (Figure 2A).

As the target–lure similarity increases, recognition memory performance in our model declines, that is, the area under the ROC curve decreases (Figure 6A). This decline can be readily understood by examining the target and lure distance distributions. When the test set includes lures that are similar to the targets, the overlap between the two distribution is higher (Figure 6B); hence, false alarm rates will be higher for any given decision criterion.

The decline of recognition performance with target–lure similarity is observed for all values of the robustness parameter $p$. There are two things that we want to highlight here. First, when $p$ is large, there is a larger shift of the distance distribution for the related lures (colored bars in Figure 6B). Second, the recognition performance is still higher compared with the low values of $p$ because the target items are retrieved with high precision (empty bars in Figure 6B). Thus, robust memory retrieval allows to distinguish between the fine details (difference between colored and empty bars), while preserving the correlations between the related items (higher peak close to zero).

It has been hypothesized that longer list length would impair recognition performance because of interference due to the increased number of stored items (Clark & Gronlund, 1996). Such a list length effect has been observed frequently in recall tasks; however, its prevalence is debated in the recognition literature (Kinnell & Dennis, 2011; Cary & Reder, 2003). We therefore tested whether and how this phenomenon might arise in our model. For the list length condition, we simulated recognition tests with varying

![Figure 5. Multicomponent model generates typical recognition ROC curves. (A) Distances between the cue and the retrieved item differ between when target items are cued (target distribution) and when lure items are (lure distribution). Scaling leads to higher correlation between the target items and noisy memory traces, leading to the peak near zero in the target distributions (first to third column). Black dashed lines represent a set of equally spaced decision criteria. Gray dashed lines represent a set of decision criteria that are more liberal than the first set, that is, a larger value of $\lambda$. Recognition ROC curves generated by the model (fourth column) are highly similar to those typically reported by experimental studies. (B) If memory noise is increased, the correlations suffer, and the effect of noise is larger if scaling is small or absent.](image-url)

**Table 1.** Mapping of Model Parameters to Experimental Findings and Manipulations

<table>
<thead>
<tr>
<th>Effect</th>
<th>Manipulation in Model</th>
<th>Manipulation in Experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td>target–lure similarity ↑</td>
<td>morph distance ↓</td>
<td>comparing unrelated to related lures</td>
</tr>
<tr>
<td>robustness to memory retrieval ↑</td>
<td>scaling $p$↑</td>
<td>MTL integrity ↑</td>
</tr>
<tr>
<td>memory strength ↑</td>
<td>memory noise ↓, storing multiple traces</td>
<td>retention time, encoding depth ↑, repeated stimulus presentation</td>
</tr>
<tr>
<td>liberal recognition decisions ↑</td>
<td>threshold offset ↑</td>
<td>pFC damage ↑, meta-knowledge about task structure, personal bias</td>
</tr>
</tbody>
</table>
numbers of study items. All other parameters were identical to isolate the effect. The number of items in the test list was identical \((n = 30)\), irrespective of study list length. Consistent with experimental findings, the ROC curves of the longer lists are shifted to the right in our model (Figure 7A).

This shift affects the ROC parameters differently. First, there are no systematic changes in the \(R\) parameter, although some small fluctuations were observed (Figure 7B, left). By contrast, the \(F\) parameter shows more pronounced changes (Figure 7B, middle). The AUC changes only slightly with increased list length (Figure 7B, right). This pattern was consistent for different values of noise (data not shown). Our results suggest that parameters specific to certain models, such as \(F\) in DPSD, may under certain circumstances be more sensitive to small changes than generic parameters, such as the AUC.

Changes in ROC curve are mainly due to increased false alarm rates. The lure distance distribution shifts toward smaller values with increasing list length, because the more items are stored, the higher the chance of finding a stored item that is similar to a lure (Figure 7C, right). This leads to higher false alarm rates, when the number of studied items increases (Figure 7A, right). For targets, increased list length does not typically affect the correct retrieval, however, when the interference is already high because of larger noise, the increase in list length seems to compound this problem (Figure 7C, left). Given a high interstimulus similarity, as is the case with face stimuli, longer lists provide more target items that could be a better match to the cue than the originally stored item, which has been corrupted by noise. Therefore, the interference increases (Figure 7C, left), and the mean distance between the targets and retrieved items slightly decreases with list length when the noise is high (Figure 7C, middle). This effect is rather small and should not have a large effect on the hit rate, because the recognition of targets depends on memory precision \((d_i\) in Equation 6), not on memory accuracy. In fact, the hit rates for different list lengths were almost identical in our simulations (Figure 7A, middle). Furthermore, because the performance differences are attributable to false alarm rates, our model predicts that the list length effect would be more readily observable if the interstimulus distance is small, consistent with the results of Kinnell and Dennis (2012).

The Effect of Memory Properties on Recognition ROC Curves

**Familiarity versus Recollection?**

Most findings on recognition memory are interpreted in terms of changes in the underlying memory process(es). In particular, the dual-process model is routinely used to
derive estimates for the proposed recollection (R) and familiarity (F) processes, arguably corresponding to the y intercept and curvilinearity of the ROC curve, respectively. We therefore examined how these parameters are affected in a model, which does not implement separate memory processes.

We first examined the effect of the two parameters of our memory module (noise and scaling) on recognition performance for different list lengths, with worse performance for longer lists (left). Whereas the hit rate remains constant for different list lengths (middle), the false alarm rates increased with list lengths (right). The mean of target distance distribution is largely unchanged across conditions (middle), whereas the mean of lure distance distributions decrease irrespective of the noise level because of higher likelihood of matching a stored item more closely (right).

**Figure 7.** The effect of list length on item retrieval. (A) There are small differences in recognition performance for different list lengths, with worse performance for longer lists (left). Whereas the hit rate remains constant for different list lengths (middle), the false alarm rates increased with list lengths (right). (B) The ROC parameters for different list lengths. (C) The list length does not have a large effect on the correct retrieval of targets (left). The mean of target distance distribution is largely unchanged across conditions (middle), whereas the mean of lure distance distributions decrease irrespective of the noise level because of higher likelihood of matching a stored item more closely (right).

**Figure 8.** The effect of memory parameters on the DPSD estimates. All parameters are affected by noise and scaling; however, the R parameter has the highest sensitivity. To facilitate the comparison between the three variables with different ranges, the color indicates the z score according to the color bar below each panel.
performance. Both have similar but antagonistic effects on memory performance: All three features of the ROC curves (R, F, AUC) are impaired by noise, whereas scaling has the opposite effect (Figure 8). These dependencies are not surprising because both, increased scaling and reduced noise, lead to improved robustness of memory retrieval. Crucially, the R and F parameters are somewhat differently affected by noise and scaling. The R parameter is more sensitive: It decreases more rapidly with increased noise and reduced scaling. At certain values, R drops abruptly to zero (Figure 8B), mostly because of the small nonzero false alarm rate for the most conservative criterion, a phenomenon also observed by previous experimental studies (Bird, 2017).

Our results show that neither noise nor scaling can be viewed as a direct analog of the DPSD model parameters R or F. The sensitivity of the R parameter, together with the more gradual change of the F parameter, enables us to obtain different combinations of these parameters by changing the scaling and noise (Figure 8). This effect could be used to account for “dissociations” of recognition and familiarity in experimental studies. Suppose that a manipulation such as increased retention time (Wais et al., 2006) or shallow encoding (Yonelinas, 2001) can be modeled as an increase of memory noise in our model. This manipulation would affect R more than F (Figure 9A), thus making it appear as if familiarity and recognition were modulated differentially (Yonelinas, 2001).

Arguably, the strongest evidence for a dissociation of familiarity and recollection comes from lesion studies. Specifically, hippocampal lesions have been reported to eliminate the y intercept in recognition ROC curves while leaving the curvilinearity largely intact (Aggleton et al., 2005; Yonelinas et al., 2002; Figure 9C). We modeled hippocampal lesions by comparing simulations with high scaling values (intact hippocampus) to simulations where no scaling was employed (lesioned hippocampus). As expected, the y intercept of the ROC curve (R) shows a notable decrease (Figure 9B). Although curvilinearity (F) is also affected by this manipulation, the effect is much smaller than the change in y intercept. Given the strong influence of noise on the y intercept, we next asked whether a reduction of noise would compensate for the reduced memory robustness in the lesioned case and

Figure 9. Dissociation of the estimated parameters. (A) Memory noise affects the y intercept of the ROC curves the most (left), as reflected by the largest change in the R parameter (right). (B) Simulations of hippocampal lesions. Compared with a control subject (green symbols), a hippocampal lesion leads to a lower scaling parameter p (solid purple lines). The reduction of memory robustness affects the R parameter more than F. However, the y intercept can be recovered, even when memory robustness is compromised, by reducing the memory noise (dashed purple lines), for example, by shortening the study and retention times. (C) Experimental findings. H refers to hippocampal patients, C refers to the controls, and H-matched is the condition where the list length and retention time were shortened for the patients. Reproduced from Wais et al. (2006) and Yonelinas et al. (2002).
found that it can indeed (Figure 9B, dashed purple lines). This is exactly the pattern observed in experiments, where the study list and the retention interval were shortened for amnesic patients (Wais et al., 2006; Figure 9C). The (re)emergence of the nonzero \( y \) intercept in amnesic patients cannot be explained by dual-process models, because the latter assume that the \( y \) intercept arises because of a qualitatively different memory process in the hippocampus. In summary, even without having two different memory processes or making assumptions about the memory strength distributions, we obtain ROC curves that appear to suggest a dissociation between two processes.

**List Strength Effect**

The list strength effect is typically not observed in recognition memory experiments (Hirshman, 1995; Ratcliff, Sheu, & Gronlund, 1992; Murnane & Shiffrin, 1991). Nevertheless, recent studies report list strength effects under special experimental conditions (Osth et al., 2014; Buratto & Lamberts, 2008; Norman, 2002; Murnane & Shiffrin, 1991). Therefore, we examined how our model performs when some of the target items are strengthened either by storing multiple traces of these items or by applying less noise to them. We tested the performance on the items in lists either containing strong (strengthened) items only, weak (nonstrengthened) items only, or in mixed lists containing half of each type. We studied the ROC curves for the strengthened and nonstrengthened items in strong (S), weak (W), and mixed lists (MS = strong items tested, MW = weak items tested).

There is a clear difference in the hit rates between pure and mixed lists. The hit rate of strengthened items is higher in the mixed lists than in the pure strong lists, whereas the opposite is true for the weak items (Figure 10A, B, middle). This finding is well aligned with experimental studies (Hirshman, 1995; Murnane & Shiffrin, 1991; Ratcliff et al., 1990) and has a straightforward explanation in our model. The more weak items there are in a list, the more the width of the target distributions increases (compare to the histograms in Figure 5). Because the criteria are adjusted based on the distance distribution of targets (see Equation 7), the mixed list results in stricter, that is, more stringently distributed criteria compared with the pure weak list and more lenient ones in comparison to the pure strong list (Figure 10A, B, left). This pattern does not depend on the strengthening method. As a result, hit rates are lower for weak items in mixed lists than in pure weak lists and higher for strong items in mixed lists than in pure strong lists (Figure 10A, B, middle). The changes in the false alarm rate are less predictable. When strengthening is achieved by storing multiple traces of the same item, the false alarm rate increases and does not differ much across conditions (Figure 10A, right), putting more emphasis on the differences in the hit rates (Figure 10A, middle). This discrepancy leads to a small list strength effect in the multiple-trace condition (Figure 10C). As with the list length condition, the differences are more pronounced in the F parameter than in the AUC. By contrast, when strengthening is modeled by noise reduction, false alarms are low in the conditions with higher proportion of strengthened items (Figure 10, right), consistent with the criterion changes (Figure 10B, left). Correspondingly, the ROC curves of pure and mixed lists are very similar, indicating no list strength effect (Figure 10D).

Although this pattern need not hold across all experimental paradigms, our results suggest that manipulations that affect the false alarm rate, for example, use of related lures, strengthening through repetition in different sentences, can increase the likelihood of observing a list strength effect, as observed in experimental studies (Buratto & Lamberts, 2008; Norman, 2002; Murnane & Shiffrin, 1991).

**The Influence of Decision Criteria on ROC Parameters**

In the signal detection theory, the shape of the ROC curve depends entirely on the distance between the noise and signal distributions, whereas the decision criterion merely determines the position on the ROC curve. This is generally true in our model as well, and the ROC curves are virtually identical for different values of the decision bias \( \lambda \) (Figure 11A, left). However, the exact positioning of the data points along the curve, especially at low false alarm rates, influences the R and F parameter fits to the DPSD model (Figure 11A, right). The F parameter decreases moderately, and the R parameter increases dramatically. The antagonistic R and F parameters are not surprising, because the overall performance remains stable and recognition performance in the DPSD model is the sum of a familiarity and a recollection process. In our model, we can see clearly that differences in parameter fits to ROC curves are artifacts of criterion placement and model fitting, suggesting that differences in R and F parameter fits to experimental data might have been incorrectly interpreted as increased involvement of recollection and decreased contribution of familiarity.

Component processes do not act in isolation to drive recognition performance, but rather interact dynamically. It is therefore important to study how the interplay between the components affects recognition memory. For instance, the combination of criterion placement with changes in memory quality generates a complex effect that might be able to account for the diversity of findings after frontal lobe lesions. A number of studies have reported a liberal response bias (Bastin et al., 2006) and selective impairment of the F parameter (Aly et al., 2011; MacPherson et al., 2008) in frontal patients. By contrast, a rodent experiment reported selective R impairment (Farovik et al., 2008). We modeled the liberal bias of frontal patients by increasing the parameter \( \lambda \), which results in a liberal shift for the entire criterion set. We then examined how this
liberal bias affects the ROC parameters in different noise conditions.

Our simulations show that, if the memory quality is not too different between the lesioned and control systems, the liberal shift in responses could lead to an increase or slight decrease of the y-offset, whereas the curvilinearity decreases earlier and more consistently (Figure 11B, $\omega \leq 0.35$). However, frontal lesions can also lead to dramatic drops in the R parameter when more liberal criteria are paired with poor memory quality (Figure 11B, $\omega \geq 0.4$). We examined different combinations of the decision bias $\lambda$ and memory noise $\omega$ and successfully replicated the specific pattern of experimental outcomes. In Figure 11C, D, we observe a larger decrease of F, consistent with the results of MacPherson et al. (2008) and Aly et al. (2011), whereas in Figure 11E, the opposite pattern is seen. Our model can also account for the recognition findings from a perirhinal patient (Figure 11F), who was reluctant to use the middle range of the response scale, hinting at a liberal use of high-confidence responses (Bowles et al., 2007). Because of paucity of selective perirhinal lesions, it is not clear whether this response strategy was unique to the tested individual or to the lesion type, although liberal response bias has also been observed in a patient with entorhinal lesion (Brandt, Eysenck, Nielsen, & von Oertzen, 2016).

Thus, differences in the memory quality between the controls and frontal patients could, in principle, arise because of compromised encoding or damage to the memory processes themselves or because of impaired
interactions with the temporal lobes. The exact effect of the frontal lesions would therefore depend on the lesion site and extent. In summary, these results show that the effects of frontal lesion on recognition ROC curves can also be explained without invoking familiarity and recollection processes as is often done.

Another interaction can occur between criterion placement and target–lure similarity. For instance, several studies have shown more linear ROCs when switched plurality lures are used (Buratto & Lamberts, 2008; Arndt & Reder, 2002; Norman, 2002; Rotello et al., 2000), typically attributed to higher recruitment of recollection because of high task demands. We tested whether linear ROC curves could also arise without employing a special memory process. Although the stimuli we are using are pictorial and may differ from the representations of plurality reversed words, the question we are asking is more general and concerns the relationship between stimulus similarity, decision criteria, and the shape of the ROC curves. In our model, more linear ROC curves were observed when higher target–lure similarity interacted with a more liberal criterion placement (Figure 12). The similar lures are usually tested together with unrelated lures, and if the decision bias is not adjusted to more conservative values (i.e., smaller $\lambda$), the ROC

Figure 11. The effect of criterion placement on model parameter fits. (A) Shifting the decision criterion (by manipulating the parameter $\lambda$) has no visible effect on the shape of the ROC curves or the AUC measure, but changes the F parameter estimates mildly and the R parameter dramatically. $\omega = 0.35, p = 1.8$, in this example. (B) The features of the ROC curve are affected differently by the decision bias $\lambda$ for different levels of memory noise $\omega$. In the right panel, $\Delta R (\Delta F)$ represents the difference between the R (F) parameter between the reference value ($\lambda = 0.006, \omega = 0.25$, green ROC curve) and other conditions ($\lambda = 0.01$, variable $\omega$). While the F parameter decreases consistently for most values of $\omega$, the R parameter increases, remains unchanged, or decreases. After $\omega = 0.35$, there is a sudden drop in $\Delta R$, corresponding to the large change in the $y$-offset in the left panel. (C–F) Accounting for experimental results of frontal or perirhinal lesion relative to controls. A more liberal response bias (increasing the decision criterion parameter $\lambda$) combined with a decrease of memory quality can account for dramatic differences in experimental observations.
curves may look more linear. Hence, we predict that meta-
knowledge about the task, such as whether or not similar
lures will be presented during test, can affect the shape of
the ROC curve. Experimental studies could test this hy-
pothesis by systematically decision bias across different
conditions (e.g., similar vs. unrelated lures).

**DISCUSSION**

We have proposed a novel account of recognition memory,
which systematically scrutinized the contributions of stimu-
lus statistics, memory, and decision-making to recognition
memory judgments. Despite its parsimonious nature, our
results indicate that the shape of the recognition ROC curve
and ensuing parameter fits reflect a much more complex
interplay of these factors than appreciated previously.

Specifically, we propose that stimulus statistics can affect
the performance directly, for example, through decreased
discriminability when similar lures are used, and indirectly,
through influence on the decision-making process, for
example, when the awareness of the similar lures leads to
criterion adjustments. Similarly, memory robustness can
lead to better performance by directly affecting the hit rate,
but also indirectly decreasing the false alarm rate because of
strength-based criterion changes. Finally, personal bias
and knowledge about the task structure were identified as
crucial factors influencing the experimental outcomes.

Taken together, our results challenge the traditional inter-
pretations of several findings in recognition memory, which
focus narrowly on the memory component.

**Why Stimulus Statistics Matters**

One of the advantages of our approach is the use of natu-
realistic images that are processed by a biologically inspired
network. This allows us to model recognition memory for
concrete stimuli, a relevant feature given the importance of
the stimulus statistics for recognition memory. Although
the exact nature of stimulus representations in the brain
is not clear, our model attempts to exclude idiosyncrasies
that might arise from using artificial feature vectors with
ad hoc statistics, as it is often the case in other computa-
tional models of recognition memory (Greve, Donaldson,
& van Rossum, 2010; Norman, 2002; Clark & Gronlund,
1996). Hence, our work supports the growing recognition
of the relevance of stimulus, statistics when examining
various memory phenomena (Rust & Mehrpour, 2020;
Johns & Jones, 2010; Lacroix, Murre, Postma, & Herik,
2006; Howard & Kahana, 2002; Kahana & Sekuler, 2002).

By using face morphs of varying similarity, we showed that
the stimulus statistics had a direct influence on the memory
performance, for example, increasing target–lure similarity
negatively affected the discrimination. The ability to distin-
guish similar stimuli has been attributed to pattern separa-
tion in the hippocampus. Typically, pattern separation refers
to the process of assigning distinct representations to similar
inputs (Yassa & Stark, 2011; Leutgeb, Leutgeb, Treves,
Moser, & Moser, 2004; Treves & Rolls, 1994). Several oper-
ations, such as sparse coding, input expansion, and decor-
relation, have been suggested to enable pattern separation
(Cayco-Gajic & Silver, 2019). In addition to input orthog-
onalization, the term “pattern separation” is also used to re-
fer to the ability to behaviorally distinguish similar sensory
inputs, such as targets from related lures (Stark, Yassa,
Lacy, & Stark, 2013; Kirwan et al., 2012). In our model, be-
havioral pattern separation was achieved by increased mem-
ory robustness. This is due to the fact that the difference
between similar items lies in small details in the input space;
therefore, if a memory system can retrieve the representa-
tions with high precision, the fidelity of the representations
would still allow to distinguish the fine details and to relate
the items to each other. We thus predict that people with
compromised hippocampus will be impaired both in recog-
nizing the similarity between the items (Olsen et al., 2015)
and distinguishing them (Görler, Wiskott, & Cheng, 2020;
Bayley et al., 2008; Westerberg et al., 2006; Holdstock
et al., 2002). The models by Norman and O’Reilly (2003)
and Greve et al. (2010) also account for the decreased
discrimination in hippocampal patients; however, they
implement different architectures and/or different retrieval
processes for different brain regions. Our approach is more parsimonious because we observe the same effect using a single memory mechanism.

Although we have only used face stimuli, it is in principle possible to extend the stimulus space to different categories, including scene images. This would allow us to directly test whether some of the dissociations, such as intact face and impaired scene recognition in healthy (Sato & Yoshikawa, 2013) and amnesic patients (Bird, Vargha-Khadem, & Burgess, 2008; Taylor et al., 2007; Cipolotti et al., 2006), can be explained by stimulus statistics alone and whether stimulus type has an effect on list length and target–lure similarity manipulations. Faces share similar features, such as overall shape and relationship between their constituents. On the one hand, high similarity may make faces harder to differentiate, thus impairing memory performance in some conditions, such as list length simulations. On the other hand, there is a relatively low number of features that need to be attended to (eyes, mouth, nose, color), making the processing of the faces less demanding. By contrast, scenes and words have higher interstimulus variation, making them easier to distinguish whenever gist information can be used, and harder, when individual features become relevant, for instance, when target–lure similarity is tested. Our future work aims to thoroughly study these aspects. It is straightforward to apply our model to nonvisual stimuli, such as words or sounds, by replacing the HMAX model with another model that generates an appropriate stimulus embedding. Such a change would not affect our findings on the dependence of the recognition ROC on stimulus statistics, because our results depend on the similarity between stimulus representations and on their precise form.

As already mentioned, input properties seem to influence the list length effect. Early studies reported decreased performance with longer lists (Gronlund & Elam, 1994; Yonelinas, 1994); however, these studies did not control for confounds such as differences in retention time, attention, displaced rehearsal, and context reinstatement (Kinnell & Dennis, 2011; Dennis & Humphreys, 2001; Murdock & Kahana, 1993). When these confounds are eliminated, the list length effect either decreases dramatically (Kinnell & Dennis, 2011; Cary & Reder, 2005) or disappears altogether (Kinnell & Dennis, 2011; Buratto & Lamberts, 2008). One exception seems to be the case where stimuli with high interitem similarity are used. In this case, there is a list length effect mainly driven by false alarms (Kinnell & Dennis, 2012; Dennis & Chapman, 2010). The role of the stimulus statistics is further highlighted by the findings of a study using two-alternative forced-choice task. A list length effect was observed with similar stimuli, but not with dissimilar stimuli (Brandt, Zaiser, & Schnuerch, 2019).

Our model can naturally account for the list length results. Namely, we find that the list length effect is driven by false alarms; therefore, it is more likely to be observed when stimuli with overlapping properties, such as faces, are used. This hypothesis is different from previous attempts to explain the presence or the absence of the list length effect. In particular, many models, which take a single-process approach, calculate the global similarity between the cue and the stored items (Clark & Gronlund, 1996). In these models, each item in the list is an additional source of noise; therefore, longer lists increase the variance of the memory strength distributions, leading to increased overlap and impaired discrimination. Other models, such as the BCDMEM (bind cue decide model of episodic memory), predict null list length effect, because the main factor influencing recognition of a specific item is the interference from previous contexts (Dennis & Humphreys, 2001). Yet, other models combine elements from both approaches, for example, Brandt et al. (2019) modified the global matching model by incorporating the influence of the extra list items. One disadvantage of assuming interference because of previous study contexts or items is the requirement that the test items are familiar, whereas recognition tests often use unfamiliar images.

Finally, there is a well-known effect of word frequency in recognition literature: Uncommon words tend to be recognized better as targets and more readily rejected as lures, creating a frequency-based mirror effect (Glanzer, Adams, Iverson, & Kim, 1993; Glanzer & Adams, 1990). Although we did not explicitly model this phenomenon, our model is not in conflict with it. If we assume that low-frequency words have less orthographic or semantic overlap than high-frequency words, our model would predict that low-frequency words would lead to lower false alarm rate if used as lures and higher memory accuracy if used as targets. In addition, the unexpectedness of low-frequency words might lead to better encoding because of increased attention, which in turn leads to higher memory precision, resulting in higher hit rate for the targets. Such an elevated attention hypothesis has been suggested before and tested experimentally (Criss & Malmberg, 2008; Malmberg & Nelson, 2003).

**Generic Memory Retrieval Is Sufficient to Explain ROC Curves**

The straightforward implementation of the memory component in our model allows us to examine the fundamental properties of recognition memory with minimal assumptions. In particular, instead of designing memory systems with different architectures or retrieval processes (Greve et al., 2010; Norman & O’Reilly, 2003), we only used one generic module for memory storage and retrieval. Although we acknowledge the differences between MTL structures, such as the hippocampus and the perirhinal cortex, it remains open whether these areas implement separate memory systems, and if so, whether these memory systems differ quantitatively or qualitatively. By implementing a single memory store corresponding to the integrity of the MTL in our model, we explored the possibility that a single memory process can give rise to rich behavior the diverse results in recognition memory tasks by using different parameter settings, representing the combined
contributions of different structures. Although biologically inspired models may offer insights that we may have missed, the parsimony of the model allows us to avoid idiosyncrasies that may be introduced by more complex models.

The feature of a single memory process may seem similar to global matching models (Shiffrin & Steyvers, 1997; Clark & Gronlund, 1996; Hintzman, 1984). However, these models base the recognition judgments on the summed similarity between a cue and all stored items, whereas our model retrieves the stored item most similar to the cue and then makes the decision based only on the distance between the two. Memory retrieval in our model more closely resembles attractor network dynamics, where the activity settles into the nearest attractor representing a stored item (Greve et al., 2010; Hopfield, 1982). We therefore believe that this process is more efficient in computational terms and hence more realistic in a biological network than global matching. To study the role of different brain structures in recognition memory, we associated the memory robustness parameter \( p \) with the integrity of the MTL. We assume that the hippocampus contributes more to memory robustness because of the strong recurrent connections in CA3, leading to deep basins of attraction (Rolls, 2007; Treves & Rolls, 1994). The attractor dynamics can also arise in cortex either because of local recurrent connections or, potentially, because of reciprocal connections between cortical areas (Rolls, 2010). However, the attractors here may be shallower or even spurious because of the absence of strong recurrent connections.

Using a simple memory retrieval process and without making assumptions about memory strength distributions, we successfully replicated several major findings in the recognition memory literature. To facilitate comparison, the ROC curves generated by our model were fit to the widely used dual-process model to obtain parameter estimates for what have been termed the “recollection” and “familiarity” processes, even though our model does not assume the existence of two qualitatively different memory processes. Nevertheless, our model can account for a number of results cited as evidence for a dissociation of recollection and familiarity (Koen & Yonelinas, 2011; Sauvage, Beer, & Eichenbaum, 2010; Aggleton et al., 2005; Fortin et al., 2004; Yonelinas et al., 1998, 2002; Yonelinas, 2001), strongly suggesting that recognition ROC curves are not sufficient to provide evidence for the different memory processes. Furthermore, our model can account for observations that are difficult to reconcile with the dual-process model. Because, in our model, the nonzero \( y \)-offset of the ROC curves is not tied categorically to certain memory processes or brain regions, our model allows for the ROC curves of the hippocampal patients to have a \( y \)-offset (Bird et al., 2008; Cipolotti et al., 2006; Aggleton et al., 2005) and reach the performance levels of controls in simplified tasks (Wais et al., 2006). Although single-process models can account for these findings, too, our model differs from the former by design (see above). That is, if the ROC curves generated by our model were fit by a single-process model, the resulting parameters might support the interpretations of that model, for example, unequal variances for targets and lures, even though our model did not impose those conditions. Such a situation could occur because the underlying assumptions of the single-process model, such as Gaussian distance distributions and fixed decision criteria, are violated in our model.

We have argued elsewhere (Hakobyan & Cheng, 2020) that attractor dynamics can lead to the phenomenology of familiarity and recollection, as well. In particular, we assumed that successful retrieval of an item can trigger the retrieval of additional details from the neocortex, as discussed in the hippocampal indexing theory (Fang et al., 2018; Teyler & Rudy, 2007). Because of deeper attractor basins, the indices are more likely to lead to the retrieval of veridical details from the neocortex (recollection), compared with the indices from the shallow or even spurious attractors in the cortex. The implications of this idea will be discussed in future work, including studies of source and associative recognition.

Finally, although our model was sufficient to successfully replicate several major findings in recognition memory, our results do not imply that a model with a single memory store is necessarily superior to models with two or more memory stores, especially if those model are equipped with visual representation and decision-making modules. However, in the range of cases that we studied here, it was not necessary to include more than one memory store in our model to account for the data.

**Memory Strength Influences Decision Criteria**

In addition to the direct effect of memory strength on recognition performance, we posit an indirect way of influence through decision-making. Namely, the decision criteria in our model were adjusted to the range of the target distance distribution. Consequently, if the target memory is strong, the distance distribution is narrower and closer to zero, leading to more stringent decision criteria. In turn, the false alarm rate decreases, whereas a high hit rate is observed because of high memory fidelity. This leads to the widely observed mirror effect in recognition memory, that is, the increase of hit rates and the decrease of the false alarms with manipulations improving the memory strength (Glanzer et al., 1993). A famous example of the mirror effect comes from the manipulations of the list strength. Our model accounts for findings that increased list strength leads to stricter criteria, higher hit rate and lower false alarm rates (Hirshman, 1995). Consistent with the experimental findings (Hirshman, 1995; Ratcliff et al., 1992; Murnane & Shiffrin, 1991), our model does not predict a list strength effect when strengthening is applied by reducing the noise. Although still small, a more noticeable effect was seen when multiple traces were stored by repeated presentations of the targets. Murnane and Shiffrin (1991) indeed showed a list strength effect
when words were repeated in different sentences, supposedly encouraging the storage of multiple traces. In our model, the list strength effect in the repetition condition was due to the fact that the false alarms were similar across weak and strong conditions, which allowed the differences in the hit rates to affect overall performance. This result predicts that disproportional change in hits or false alarms can lead to a small list strength effect, for instance, if similar stimuli are used. This prediction has recently been verified in studies where related lures were used (Buratto & Lamberts, 2008; Norman, 2002) and in a study where the use of (hard to distinguish) fractal stimuli resulted in list strength effect, whereas the use of (easier to distinguish) face and scene stimuli did not (Osth et al., 2014).

The suggested changes in criterion placement across different conditions have been hypothesized previously in the literature (Verde & Rotello, 2007; Stretch & Wixted, 1998; Hirshman, 1995). However, there are also competing accounts, such as the differentiation hypothesis (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). The latter hypothesis assumes that the target items become less confusable when the memory is strong, thereby increasing the chances of matching the target cue and mismatching a lure. Although we agree in principle with the notion that better memory reduces interference, our model does not use the global matching mechanism proposed in differentiation models and does not assume changes of the lure distance distribution per se. Despite some recent attempts (Kantner, Vettel, & Miller, 2015; Criss, Wheeler, & McClelland, 2013; Criss, 2010), it remains a challenge to unambiguously dissociate between the criterion shift and differentiation accounts.

Finally, we assume that human participants can access their memory strength before being tested. For instance, when the study test lag is short or the items were presented repeatedly, they may be confident in their memory strength and may adopt a more stringent criterion placement. In our model, we used the higher and lower ends of the target distribution to achieve such placement, and we assume that such a process may occur automatically, not necessarily with much conscious effort. One could possibly use the mean of the target distribution rather than the endpoints to set the criteria, which would probably be more intuitive when comparing to assessments of human confidence about their memory strength. It is, however, possible that the endpoint for the most confident response is determined by recalling a few items that have stronger memory, as suggested by Hirshman (1995). In either case, there is experimental evidence that memory strength leads to criterion adjustments. In fact, if participants study mixed lists of strong and weak items and are tested on the weak items only, their bias is more conservative compared with the condition where only weak items were studied (Hirshman, 1995, Experiment 4). Another line of evidence in favor of pretesting criterion placement comes from a study by Cox and Dobbins (2011), where the hit and false alarm rates as well as confidence ratings were largely similar in standard or pure lists (targets or lures only), even if the participants knew about these contingences. However, this does not mean that the participants do not respond to the realities at test. For instance, there was a small tendency of more conservative or liberal responses when either pure target or pure lure lists were used, respectively. Hence, meta-knowledge about the task can lead to readjustments of the criterion placement during the test. Experimental evidence seems to suggest that this happens early and stays relatively unchanged throughout the test phase. Such adjustments as well as general personal bias were modeled by the parameter $\lambda$ in our model (the Personal Bias and Meta-knowledge Shift Decision Criteria section follows to discuss this issue further). Note that, in any case, our approach is compatible with experimental reports of between-list criterion changes (Singer & Wixted, 2006; Stretch & Wixted, 1998; Hirshman, 1995) and does not currently implement the more contentious within-list item-by-item criterion adjustments (Franks & Hicks, 2016).

**Personal Bias and Meta-knowledge Shift Decision Criteria**

Decision-making is a complex process and can be influenced by multiple factors such as personal bias or preexperimental knowledge. We can capture the effect of these factors in our model by shifting the decision bias. This manipulation leads to prominent changes in the ROC shape and parameters, which are generally attributed to familiarity and recollection processes. Personal bias has been shown to be relatively stable across many conditions (Kantner & Lindsay, 2012). Given the importance of the decision bias described in our study, this factor is especially relevant in interpreting case studies. For instance, Bowles et al. (2007) tested the patient NB with selective perirhinal lesion and concluded that she was selectively impaired in familiarity. The authors explicitly mentioned the liberal use of the high-confidence responses by the patient and the refusal to employ certain points on the confidence scale. Although they concluded that the patient’s bias would not have had an effect on the curvilinearity and the $y$-offset, our results clearly challenge that assumption and call for a more careful interpretation of the model estimates. Note, however, that the authors report consistent results when other methods such as the remember/know paradigm and response deadline procedure were used (see Köhler & Martin, 2020, for a summary on the patient NB).

Furthermore, patients with frontal lesions are known to have an abnormally liberal response bias (Bastin et al., 2006; Swick & Knight, 1999). When this aspect is incorporated into recognition memory judgments, the conflicting findings from frontal lesion studies can be reconciled. According to our results, when more liberal criterion is combined with lower levels of noise, the curvilinearity of the ROC curves decreases faster than the $y$-offset, consistent with experimental results (Aly et al., 2011; MacPherson et al., 2008). By contrast, the $y$-offset may drop rapidly when the
noise increases, consistent with the results of Farovik et al. (2008). Based on the site and the extent of the frontal lesions, memory noise can be attributable to encoding impairments because of compromised working memory and attention or to memory retrieval itself. Similar reasons could underlie the impaired curvilinearity and preserved y-offset following lesions of the amygdala (Farovik, Place, Miller, & Eichenbaum, 2011).

We also considered more explicit sources of bias such as knowledge about the task structure. There are multiple studies showing that the awareness of target and lure contingencies affects recognition performance. For instance, if participants know that the probability of targets at test is high, they employ liberal criterion, while being more conservative, if the probability is low (Franks & Hicks, 2016; Aminoff et al., 2012). Another task domain, where top–down decision-making can have an effect is item recognition with similar lures. Although people tend to be more stringent when tested on similar lures compared with dissimilar ones (Benjamin & Bawa, 2004), this aspect is underappreciated when interpreting recognition memory findings. For instance, multiple studies report a decrease of curvilinearity in ROC curves when lures are the switched-plurality forms of the targets (Buratto & Lamberts, 2008; Arndt & Reder, 2002; Norman, 2002; Rotello et al., 2000). The increased linearity of the ROC curves in this condition has often been explained by assuming a higher involvement of recollection because the familiarity for both forms of the words is likely to be equal. We simulated a similar experiment using highly similar morph images instead of switched plurality lures and found that the ROC curves looked more curvilinear when the decision criterion was strict and became increasingly linear when the criteria were relaxed. Experimental designs informing the participants about the presence of highly similar lures could in principle lead to more conservative criterion setting (Rotello et al., 2000). However, to our knowledge, this has not resulted in more curvilinear ROC curves, contrary to our prediction. Crucially though, it is worth noting that the related lures are typically tested together with unrelated lures, which might discourage the participants to shift the default criterion. In fact, people seem to be reluctant to change their criterion within an experimental session (Verde & Rotello, 2007). Therefore, our hypothesis could be examined in an experiment where only targets and related lures are tested.

Our results unambiguously demonstrate that the decision bias can affect the estimates of the ROC curve parameters and can therefore lead to incorrect interpretations of the experimental outcomes. It is an intriguing empirical question whether gradual change of test contingencies would result in the pattern observed here.

Conclusions

We acknowledged the complexity of recognition memory judgments by proposing a multicomponent model, which allowed us to systematically scrutinized the role of perceptual, mnemonic, and decision-making in recognition memory. Our results showed that the interplay of these components determine the shape and the parameters of the ROC curve, a widely used tool in recognition memory. We therefore conclude that many divergent findings in the recognition memory literature might be due to a failure in controlling for all the components that affect the recognition ROC curve. In turn, this failure often stems from a misleading focus on attributes of the involved memory processes.

Acknowledgments

This work was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft; project number 122679504 – SFB 874, B2). We would like to thank Eva Yezerets and Mariam Jabara for their assistance in developing the sensory processing module of our model.

Reprint requests should be sent to Sen Cheng, Institute for Neural Computation, Ruhr University Bochum, Universitätsstraße 150, 44801 Bochum, Germany, or via e-mail: sen.cheng@rub.de.

Funding Information

Deutsche Forschungsgemeinschaft (http://dx.doi.org/10.13039/501100001659), grant number: 122679504 – SFB 874, B2.

Diversity in Citation Practices

A retrospective analysis of the citations in every article published in this journal from 2010 to 2020 has revealed a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the Journal of Cognitive Neuroscience (JoCN) during this period were M(an)/M = .408, W(oman)/M = .335, M/W = .108, and W/W = .149, the comparable proportions for the articles that these authorship teams cited were M/M = .579, W/M = .243, M/W = .102, and W/W = .076 (Fulvio et al., JoCN, 33:1, pp. 3–7). Consequently, JoCN encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article’s gender citation balance.

REFERENCES


