A NEW MEASURE OF MNEMONIC DISCRIMINATION APPLICABLE TO RECOGNITION MEMORY TESTS WITH CONTINUOUS VARIATION IN NOVEL STIMULUS INTERFERENCE

by

Simon Léger

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
July 2024

Table of Contents

List of Tables

List of Figures

Abstract

Background: Mnemonic discrimination (MD) involves distinguishing new stimuli from memories of highly similar "lure" items or events, and is a putative indirect probe of dentate gyrus functioning. MD is impaired in the elderly and in individuals with hippocampal lesions, schizophrenia, major depressive disorder, and Alzheimer's disease. The gold-standard MD test, called the mnemonic similarity task (MST), is rarely used in clinical research. We thus aimed to demonstrate convergent validity between the MST's MD measure and a novel analysis method that extracts information about MD that could be used in widely clinically used recognition memory tests that do not have categorical distinctions between "lures" and "foils."

Methods: By fitting a logistic function to the relationship between stimulus interference and the probability of classifying a stimulus as novel, at the single participant level, we derived participant-level indices of MD (λ) and overall recognition memory performance (Δ). We applied the novel measures to MST data from two independent datasets (N=18; N=67), and to synthetic MST data. Using linear mixed-effects modelling, we sought to confirm that λ predicts the MST's *lure discrimination index* (LDI), while Δ predicts the MST's overall recognition memory index (REC).

Results: Across all datasets, λ predicted LDI (β=0.76, 95% CI [0.62-0.91], p<0.001), but not REC (β=0.06, 95% CI [-0.03-0.15], p=0197), while Δ predicted REC (β=0.93, 95% CI [0.83-1.02], p<0.001), but not LDI (β=-0.06, 95% CI [-0.20-0.09], p=0.438). The λ and Δ indices were not correlated. Simulations suggest that λ may be more robustly estimated in participants with stronger overall recognition performance.

Conclusion: Our novel measure accurately indexes MD, without correlating with overall recognition memory performance. Future studies should apply it to large clinical datasets with widely used recognition memory tests.

Keywords

Recognition Memory, Mnemonic Discrimination, Mnemonic Similarity Task, MST

List of Abbreviations and Symbols Used

Δ	Novel recognition index	6
λ	Novel mnemonic discrimination index	6
CVLT	California Verbal Learning Test	1
CVMT	Continuous Visual Memory Test	3
ENIGMA	Enhancing NeuroImaging Genetics through Meta-Analysis	1
HVLT	Hopkins Verbal Learning Test	1
LDI	Lure Discrimination Index	3
MST	Mnemonic Similarity Task	1
RAVLT	Rey Auditory Verbal Learning Test	1
REC	Recognition Index	3

Chapter 1

Introduction

Mnemonic discrimination is the ability to distinguish new stimuli from highly similar items stored in memory. Although there are other processes that contribute to mnemonic discrimination [1–4], successful mnemonic discrimination is thought to be largely supported by the hippocampus, specifically the hippocampal dentate gyrus [5–8], such that measurements of mnemonic discrimination putatively serve as an indirect probe of dentate gyrus function. Importantly, poor mnemonic discrimination is associated with aging [9,10], impaired hippocampal function [5,11–13], cognitive impairment [9], and several psychiatric disorders, such as schizophrenia [14], and depression [15–17]. More specifically, mnemonic discrimination impairments were seen in patient BL whose lesions primarily involved the dentate gyrus, in spite of BL's intact performance on other hippocampal-dependent tasks [5,18,19]. Therefore, mnemonic discrimination assessment is clinically relevant, and may allow us to indirectly study dentate-gyrus-dependent disease mechanisms. However, our ability to study mnemonic discrimination in clinical populations is currently limited by available samples of patient data using the gold-standard measure of mnemonic discrimination, the Mnemonic Similarity Task (MST) [20].

In research on clinical neuropsychiatric populations, memory testing is commonly done using standard clinical batteries, including tests of memory such as the Rey Auditory Verbal Learning Test (RAVLT), Hopkins Verbal Learning Test (HVLT), and California Verbal Learning Test (CVLT) [21]. Indeed, in a recent collaboration by the Enhancing NeuroImaging Genetics through Meta-Analysis (ENIGMA) consortium, more than 10,000 verbal learning test results were pooled for analysis, comprising healthy controls and patients with traumatic brain injury [22]. Many more verbal learning test records are also available for analysis from other clinical populations. Other recognition memory task data are also available from existing clinical studies, such as continuous visual memory test data from patients with bipolar disorder [23]. If we could use these recognition memory paradigms to probe mnemonic discrimination, then we would have access to large, clinically relevant samples to allow us to study how mnemonic discrimination is affected by various neurological and psychiatric disorders, with higher degrees of ecological validity. Furthermore, clinically applicable recognition memory tests, such as those from the verbal learning tests discussed above, are also often collected alongside clinical trials. By extracting measures of mnemonic discrimination performance from these clinically common recognition memory tests, we may enable many highly powered studies of mnemonic discrimination as a biomarker of treatment outcomes [24,25]. Finally, since many common clinical recognition memory tests have also been provided to affected patients and their unaffected relatives [26], enabling measurement of mnemonic discrimination from these tests could facilitate large studies of the degree to which mnemonic discrimination impairments in clinical populations are either (A) familial and potentially represent trait vulnerability markers, or (B) possibly sequelae of the disease process itself. However, to our knowledge, there is no approach yet validated to examine mnemonic discrimination in standard recognition memory paradigms that are commonly used in clinical populations. Prior to implementing such applications, however, a measure must be developed and (A) validated against existing gold-standard approaches of assessing mnemonic discrimination, and (B) examined for its stability and robustness across different assumptions and testing conditions.

Therefore, the present study aims to take an important step toward developing a novel approach to extracting information about mnemonic discrimination performance based on recognition memory paradigms other than the MST. The MST is a delayed recognition memory paradigm in which participants study a set of images, called the *study list*, and are subsequently presented with a *test list* of images that contain (A) all of the study list images, here called “old” images, (B) images that are highly distinct from the old images, here called “foils”, and (C) images that are similar, but not identical, to the old images, which are here called “lures” [20]. Lure images are designed to introduce interference by virtue of perceptual similarity. Individuals with perfect mnemonic discrimination performance would be able to identify all lures as novel, regardless of the fact that lure images may be highly similar in appearance to old images on the study list. The critical difference between the MST and most existing, clinically common, recognition memory paradigms is that the MST has a *categorical* separation between old, lure, and foil stimuli, which facilitates measurement of overall recognition memory performance (called *REC*) and mnemonic discrimination, via the *lure discrimination index* (LDI). The MST REC measure is generally quantified as the probability that an old image is correctly classified as old, minus the probability that a foil image is misclassified as old. The MST LDI is generally quantified as the probability that a lure image is correctly classified as "similar," minus the probability that a foil image is mistakenly classified as "similar". Other analytical methods have been proposed for the MST, but they all require clear categorical distinctions between which images in the test set are lures and foils [27, 28]. However, in clinically common memory tests, the differences between testing stimuli may not be so significant that one group can be classified as "foils" while the others are "lures." Rather, these tests, such as the RAVLT [29] or the Continuous Visual Memory Test (CVMT) [23], may have test stimuli whose degree of similarity varies in a more continuous fashion.

To address this problem, we develop a novel approach to measuring mnemonic discrimination that does not require a categorical distinction between "lures" and "foils", but rather can quantify mnemonic discrimination in relation to some continuous variation in similarity between test stimuli and those in the study list. To accomplish this, we first develop a general mathematical formulation of mnemonic discrimination paradigms. We then motivate the intuition behind our novel approach to measuring recognition memory performance and mnemonic discrimination, which relies on fitting a 5-parameter logistic function to individual trial-by-trial data. We then show that our measures of recognition memory performance and mnemonic discrimination consistently track MST REC and LDI, respectively, as computed from two MST studies, demonstrating convergent [30] and divergent validity. Finally, we generated simulated MST data to examine conditions that may moderate the performance of our novel measures.

Chapter 2

Methods

2.1 Mathematical Generalization of Mnemonic Discrimination Paradigms

In order to develop a novel set of indices for mnemonic discrimination, we first sought to develop a generalization of mnemonic discrimination paradigms to facilitate the extraction of the essential components necessary for mnemonic discrimination measurement, in a task-agnostic fashion. Let X be a random variable defined on a discrete space consisting of tokens $\{1, 2, \ldots, K\}$, which in the recognition memory task context are different stimuli. Let x_i be the i'th realization of X, sampled with replacement from a distribution over X.

A recognition memory experiment consists of an initial study phase in which an agent is supplied with a sequence of N items of X, denoted $\overline{X} = \{x_1, x_2, \ldots, x_N\}$. For some index n, where $1 < n < N$, we can divide the list of items $\overline{X}$ into a study and test list, calling $\overline{X}_S = \{x_1, \ldots, x_n\}$ the *study list*, and $\overline{X}_T = \{x_{n+1}, x_{n+2}, \ldots, x_N\}$ the *test list*. Given this division, let $Y^* = \{y_i^*\colon\ i \in 1, 2, \ldots, N-n\}$ be a vector denoting whether item x_{n+i} is novel: $y_i^* = \mathbb{I}[x_{n+i} \notin \overline{X}_S]$, which takes a value of 1 if the argument is true, and a value of 0 otherwise. When $y_i^* = 1$, then x_{n+i} is known as a "lure" or "foil" (meaning that it is novel), and when $y_i^* = 0$, it is known as a "target."

Let y_i be the agent's estimate of y_i^* (that is, its prediction of whether $x_{n+i} \notin \overline{X}_S$, which means that the test stimulus x_{n+i} is novel). We assume that y_i is generated by a function $f_\theta(x_{n+i};\ \overline{X}_S)$ with parameters θ. In other words, we assume that the agent's predictions of whether a given item is novel is a function governed by some parameters θ that characterize the agent, as well as the list of items studied. Note that the prediction y_i can be either discrete (such as for binary yes/no recognition, which is the case we consider here), or continuous (such as for probabilistic predictions or expressions of confidence).

A recognition memory paradigm becomes a mnemonic discrimination paradigm when we consider how the probability of identifying test list stimulus x_{n+i} as novel, which here we will denote as $p_\theta(x_{n+i})$, varies in relation to the degree of similarity between x_{n+i} and the most similar stimulus in the study list $\overline{X}_S(n)$. If we let $d(x_{n+i},\ \overline{X}_S(n))$ represent some measure of perceptual or semantic "dissimilarity" between the item x_{n+i} and study list $\overline{X}_S(n)$, such that $d(x_{n+i},\ \overline{X}_S(n)) = 0$ implies that x_{n+i} is an old item in the study list (no dissimilarity), then one should expect in general that $p_\theta(x_{n+i})$ increases monotonically with respect to $d(x_{n+i},\ \overline{X}_S(n))$. That is, as the stimulus x_{n+i} becomes more distinct from the stimuli in the study list, the more likely the participant is to classify it as novel. For an individual with excellent mnemonic discrimination ability, $p_\theta(x_{n+i})$ approaches its maximal value for even small values of $d(x_{n+i},\ \overline{X}_S(n))$, implying that the participant's recognition memory system is highly sensitive to even small differences between old and new stimuli.

2.2 A Novel Measure of Recognition and Mnemonic Discrimination Performance

Consider the range of perceptually or semantically relevant dissimilarities from novel to studied stimuli, $d(x_{n+1},\ \overline{X}_S(n))$, which we assume to be scaled from 0 (the dissimilarity of studied stimuli to themselves, describing a stimulus identical to one studied), to 1 (the dissimilarity between the study list and the most distinct novel stimulus in a test list, describing the most dissimilar stimulus). Recognition of the most distinct novel stimuli does not depend heavily on mnemonic discrimination. As dissimilarity to the studied stimuli declines, then successful recognition memory should be taxed to progressively higher degrees. Poor recognition should reflect poor performance across all similarity levels, whereas poor mnemonic discrimination should reflect poor recognition in relation to high similarity levels (i.e., lower levels of dissimilarity) in test stimuli. A mnemonic discrimination index should therefore capture the degree to which recognition memory performance declines as a function of interference by perceptual or semantic similarity of test stimuli.

These assumptions imply that the probability of classifying a stimulus as new, $p_\theta(x_{n+1})$, is related to stimulus dissimilarity, $d(x_{n+1},\ \overline{X}_S(n))$. In other words, the probability of classifying a stimulus as new increases monotonically with respect to the dissimilarity of that stimulus from those in the study list. For each participant, we can imagine a performance curve representing the probability of classifying stimuli as new, $P_{NEW}(dissimilarity)$, as being strictly increasing with respect to dissimilarity. Indexing recognition in such a curve would involve calculating the difference between $P_{NEW}\,(1)$, correct identification of the most dissimilar items, and $P_{NEW}\,(0)$, incorrect identification of *OLD* items as being novel. We call this the recognition index $\Delta =$ $P_{NEW}(1)$ - $P_{NEW}(0)$. To index mnemonic discrimination, we need to capture how far P_{NEW}(dissimilarity) deviates from $P_{NEW}(1)$, in relation to dissimilarity. For instance, perfect mnemonic discrimination would imply that $P_{NEW}(dissimilarity) \approx P_{NEW}(1)$ for all dissimilarity values greater than 0. The mnemonic discrimination index should be maximal in those cases. Conversely, the mnemonic discrimination index should be low if it only reaches its maximum value at $P_{NEW}(dissimilarity) \approx P_{NEW}(1)$ at high values of dissimilarity. To capture mnemonic discrimination we need to capture an index that is inversely related to $P_{NEW}(1) \; - P_{NEW}(dissimilarity)$ across all dissimilarities of novel stimuli in the paradigm. To obtain this index we must model P_{NEW} (dissimilarity) as a performance curve fit to recognition data. We modeled the relationship P_{NEW} (dissimilarity) using the following sigmoidal function,

$$P_{\text{NEW}}(\text{dissimilarity}) = d + \frac{a - d}{\left(1 + \left(\frac{\text{dissimilarity}}{c}\right)^{b}\right)^{e}} \tag{2.1}$$

which is graphically depicted in Figure 2.1. The parameter a adjusts the lower asymptote of the curve, which will mostly influence the probability that a participant will *incorrectly* classify an *OLD* image as *NEW*. The parameter b corresponds to the slope of the curve. The parameter c shifts the curve horizontally. The parameter d alters the upper asymptote of the curve, which will mostly influence the probability that a maximally dissimilar image would be correctly classified as *NEW*. The parameter e enables asymmetry of the sigmoidal function to add flexibility in participant-level fits.

With a performance curve representing P_{NEW} (dissimilarity), we can calculate its deviation from $P_{NEW}\,(1)$ by taking the area between the performance curve and its maximum. Since having a larger Δ would also mean having a larger area between the curve and its maximum, we scale the output by Δ to reduce the colinearity of the two measures. Our mnemonic discrimination index, denoted λ, is thus defined as

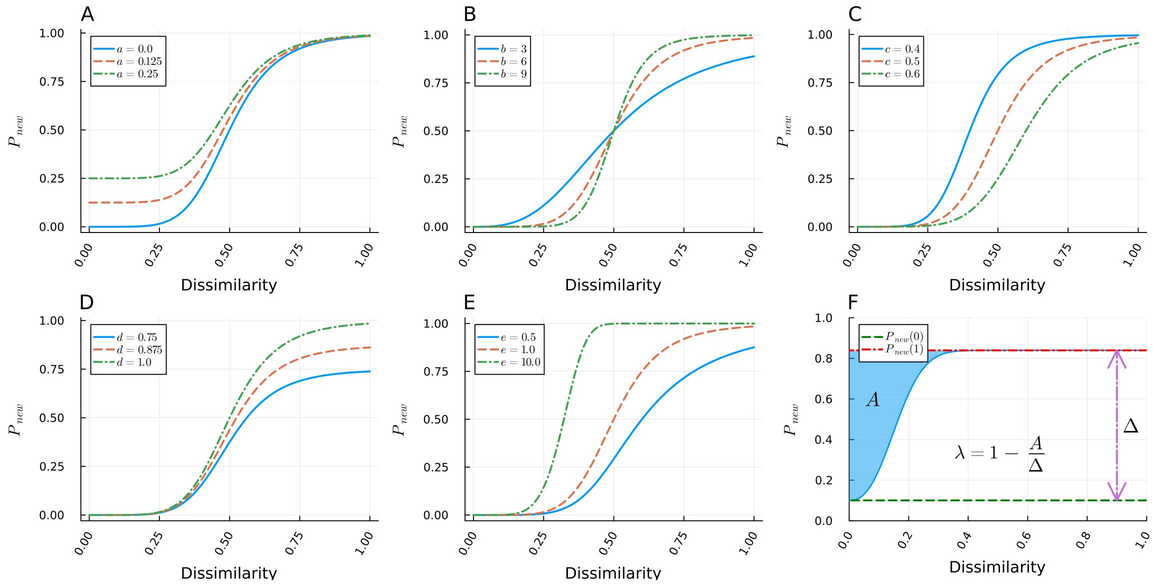

Figure 2.1: Consequences of altering parameters on the 5-parameter logistics function. ***Panel A***: Parameter a primarily corresponds to the lower asymptote, which will mostly influence the probability of misclassifying an old image as new. ***Panel B***: Parameter b changes the steepness of the curve. ***Panel C***: Parameter c shifts the curve horizontally. ***Panel D***: Parameter d primarily corresponds to the upper asymptote, which will mostly influence the probability of classifying the most distinct new image as new. ***Panel E***: Parameter e adjusts the asymmetry of the curve. ***Panel F***: Example of a fitted curve. The dashed green line shows $P_{NEW}(0)$, representing the probability of classifying an old image as new. The dot-dashed red line shows $P_{NEW}(1)$, representing the probability of classifying *the most distinct* (semantically dissimilar) new image as new. The difference between $P_{NEW}(1)$ and $P_{NEW}(0)$ defines an overall measure of recognition memory performance, Δ. The λ index of mnemonic discrimination is calculated by examining the degree to which P_{NEW} declines with increasingly similar stimuli, specifically by calculating the inverse of the area between the curve and its upper asymptote and dividing it by Δ.

follows (Figure 2.1F):

$$\lambda = 1 - \frac{A}{\Delta}, \quad \text{where} \quad A = \mathrm{P}_{\mathrm{NEW}}(1) - \int_0^1 \mathrm{P}_{\mathrm{NEW}}(x)\,dx \tag{2.2}$$

A is the integral defining the area between the sigmoidal function P_{NEW}(dissimilarity) and the maximal value of $P_{NEW}(1)$, the latter of which represents the correct recognition rate of the most distinct novel items.

2.3 Experimental Validation

Dataset 1: Lee & Stark 2023

The first dataset is from a study that aimed to develop new cognitive models based on the multinomial processing tree framework to be applied to the MST [27]. We analyzed data from 21 participants (13 female, mean age: 21, age range: 18-24) who were given the standard 3-way classification (old-similar-new) version of the MST [20]. Participants were students recruited through the Sona Systems experimental management system at the University of California at Irvine. The study received ethics board approval, as noted in the original study, and their data were made freely available online. The participants studied 128 images while tasked with judging the pictures as either containing "indoor" or "outdoor" items, in order to elicit incidental encoding (that is, to have participants attentively encode the pictures without knowing that their memory of these images would be subsequently tested). Participants then were tested on 192 images made up of 64 old images, 64 new images, and 64 lure images. Three participants were excluded from the analysis: two due to missing responses that exceeded two standard deviations above the mean, and the other for having answered randomly with respect to stimulus type.

Dataset 2: Wahlheim et al. 2022

The second dataset is from a study investigating the relationship between functional connectivity as measured using fMRI between brain areas within the default mode network and performance on the MST [31]. It included data from 36 younger adults (20 female, mean age: 22.21, age range: 18–32) and 36 older adults (20 female, mean age: 69.82, age range: 61–80). Participants were recruited from the greater Greensboro, North Carolina community. Ethics board approval for this study was noted in the original manuscript and their data were made freely available online. Participants were given the old-similar-new version of the MST. Participants studied 72 test images while making indoor/outdoor classification decisions on each item, and then were tested on their recognition of 108 images (36 old, 36 new, 36 lures). Five participants were excluded in the analysis: four due to missing responses that exceeded two standard deviations above the mean, and one for apparent random responses relative to the stimulus type.

Datasets - Lure Bins

Each lure image trial in both datasets has an associated lure bin value ranging from 1 to 5 which corresponds to the image's similarity to a test set image (a lure bin of 1 indicates highly similar lures, whereas images with lure bin 5 are the least similar to old images). We translate lure bins to dissimilarities in our analysis, as described below in the Statistical Analysis section. Lure bin sets are image pair groupings empirically determined to be similarly discriminable [9, 32]. They were formed by evaluating the rates in which participants taking the MST mistakenly identified lures as being old for each pair of stimuli in the MST. Image pairs with similar false alarm rates are grouped together in ordinal bins corresponding to their observed false alarm rate. That is, real-world participants most often misclassify lures identified as belonging to lure bin 1 as "old", whereas they most rarely misclassify lure bin 5 images as "old". Lure images in the test phase in the Wahlheim et al. [31] dataset only belonged to one of the 3 most similar lure bin sets (i.e. lure bins 1-3).

Statistical Analysis

For each study, the data were pre-processed such that the ordinal lure bins were standardized into a "dissimilarity measure" that ranged between 0 (old stimuli) and 1 (completely new stimuli; the most dissimilar possible). For the lure stimuli, this was done as follows: (LureBin)/6, where *LureBin* is a given target image's lure bin. This facilitates the computation of $p_\theta(x)$ for some image x, representing the probability that the participant will identify the test image as novel.

To model the influence of dissimilarity on the response metric we fit the 5-parameter logistic function to each participant's combination of stimulus dissimilarity values and old/new/similar responses (Participant responses were coded as 0 for "old" images, and 1 for "similar" and "new" responses). This was done using nonlinear least squares in the *LsqFit.jl* package for the Julia programming language (https://github.com/JuliaNLSolvers/LsqFit.jl). From these fitted curves, we extracted the λ and Δ measures for each participant. We used the following linear mixed effects model, presented here in R syntax (for the *lme4* package [33]), to determine whether λ and Δ were colinear: $\lambda \sim \Delta + (1|Study)$.

For each dataset, we also calculated the original MST LDI and REC scores for each participant. The MST's LDI score is calculated by taking the proportion of lures correctly identified as being lures and subtracting from this the proportion of foils incorrectly identified as being lures. The MST's REC score is similarly calculated by taking the proportion of old items correctly identified as being old and subtracting the proportion of foils incorrectly identified as being old items.

We sought to evaluate whether λ and Δ could explain variation in the original MST LDI and REC, respectively. Secondly, we sought to demonstrate that Δ was not associated with the original MST LDI, and that λ was not associated with the original MST REC. To do this, we used linear mixed effects modelling, with study-level random intercepts. The following model sets the original MST LDI as the dependent variable, and the λ and Δ values as independent variables: $LDI \sim \lambda + \Delta + (1|Study)$. This model would provide convergent validity if the MST LDI associates with λ, and would provide divergent validity if LDI does not associate with Δ. The second model sets the original MST REC as the dependent variable, with λ and Δ as independent variables: $REC \sim \lambda + \Delta + (1|Study)$. This model would provide convergent validity if the MST REC associates with Δ, and would provide divergent validity if REC does not associate with λ. The older and younger subgroups in the Wahlheim et al. [31] study were treated as separate studies in an additional analysis.

Participants with missing responses exceeding two standard deviations above the mean in the dataset were excluded in the analysis. One participant in dataset 1 was short of the cutoff and was deemed an influential point in the analysis (the cutoff was 46.55, and the participant had 43 missing responses). We thus reexamined the relationships again after removing this participant in an additional analysis.

Sensitivity Analysis - Low Interference Tests

A concern that arises in wanting to extract information about mnemonic discrimination from tests other than the MST is that traditional tests of recognition memory are not explicitly designed to elicit large false alarm rates by employing lures (images specifically chosen to be perceptually similar to those in the study list), as is done in the MST. While this issue may prevent some recognition memory tests from being used to index mnemonic discrimination, we sought to examine the performance of our measures in situations when most test list items are highly dissimilar from old items (i.e., in situations where the test set is not designed to have high interference). To do this, we reanalyzed the Lee & Stark [27] dataset while excluding trials involving the most similar lures. Specifically, lure trials with lure bins of 3 or below were excluded when extracting the λ and Δ indices.

Sensitivity Analysis - Semantic Perceptual Similarity derived from Deep Neural Networks

We sought to trial an alternative dissimilarity measure to lure bins to address a potential circularity induced by the fact that lure bins are used to determine discrimination ability, yet are themselves derived from human participants' discrimination performance. Therefore, we conducted an additional analysis using the Wahlheim et al. [31] dataset, employing deep neural networks to quantify the dissimilarity of test items in the MST. These models enable the transformation of raw images onto embeddings (analogous to coordinates on a map) that represent images' perceptually salient high-level features. Much like points on a map, the embeddings of the MST images can be compared to one another to obtain distances that can represent the perceptual dissimilarity of the two images.

To obtain image embeddings of MST images, we leveraged a deep learning model from the *MetalHead.jl* package for the Julia programming language, *ResNet-152* [34]. This model was trained to recognize images from the ImageNet database. When presenting this trained model with MST images, we used the model's highest-level representation of each image to compare the images with one another. Specifically, the model transformed each MST image into a continuous vector by extracting the model's high-level embeddings. These embeddings can be thought of as high-level abstract representations of the contents of an image. With a vector representation of each MST image, we then defined the neural network-derived "perceptual dissimilarity" between image pairs as the cosine distance between their respective image vectors.

Since each participant did not study the same set of images in the study phase, we calculated each test trial's dissimilarity independently for each participant. A given test trial image's dissimilarity was thus defined as the neural network-derived "perceptual distance" between that test trial image and its least "perceptually dissimilar" studied image. A comparison between each trial image's neural network-derived dissimilarity measure and the trial image's lure bin is depicted in Figure S1. With this key difference in how dissimilarity is defined in the statistical analysis plan, the rest of the procedure was conducted as described in the Statistical analysis.

2.4 Synthetic Data Experiments

Synthetic data generation

Since MST data from human participants may represent only a small fraction of all possible MST task performances, we sought to generate synthetic data with multinomial processing tree models as per Lee & Stark [27] to further validate that our novel measures track LDI and REC. Modeling synthetic agents that undergo the MST allows for studying the conditions under which our measurement approach breaks down, and identifying the edge cases where our novel measures may not track LDI and REC. Additionally, we sought to establish a simulated ground truth of lure discriminability to which we can investigate and compare how well the novel λ index and LDI can extract its information based on the simulated agent's behaviour.

In our model, synthetic agents respond to stimuli categorized as old, new, or lure based on predefined probabilities (graphically depicted in Figure 2.2A). Each agent has a probability ρ of recognizing an old stimulus. If the stimulus is not recognized, the agent guesses its category—old, similar, or new—based on probabilities γ^O, γ^S, and γ^N, respectively, where $\gamma^O + \gamma^S + \gamma^N = 1$.

For new stimuli, agents have a probability ψ of recognizing that the item was not previously studied. If this recall fails, the agent again resorts to guessing among the three categories with the probabilities γ^O, γ^S, and γ^N.

When encountering a lure stimulus, the agents have a probability ρ of recognizing the related studied item. Subsequently, there is a probability δ_l that the agent successfully discriminates the lure as being similar, rather than old. If recognition fails, guessing occurs as previously described.

The agents' probability δ of discriminating lures from studied items depends on

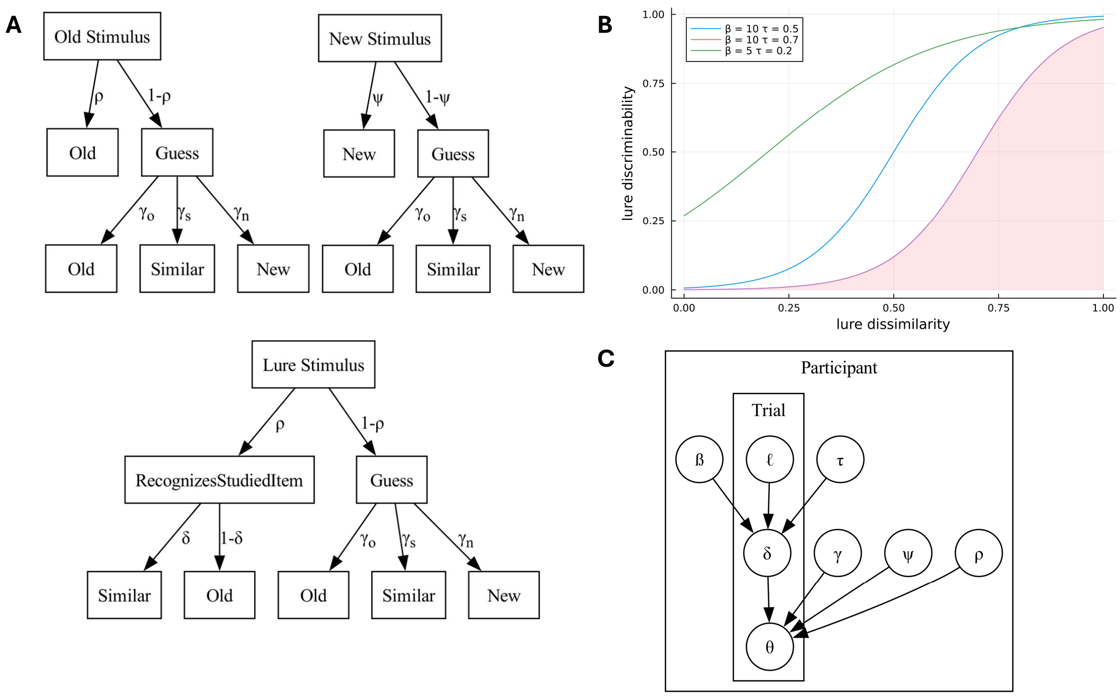

Figure 2.2: Summary plots of the synthetic data experiment methodology. ***Panel A***: Visual depiction of the synthetic agent's decision trees on a test trial. Each simulated agent has set values of ρ, γ, ψ, β, and τ that dictates the decisions the agents take when presented with either an old, new or lure stimulus. The β and τ parameters alongside the relevant stimulus' lure similarity, ℓ, govern the δ parameter used in discriminating lure stimulus. ***Panel B***: Graphical depiction of the two-parameter logistic function with sample parameters used to characterize the relation between lure discriminability (δ) and lure dissimilarity (ℓ) for each simulated synthetic agent. The β parameter adjusts the steepness of the curve, while the τ parameter shifts the curve horizontally. The shaded area beneath the red curve corresponds to that curve's ground truth λ. ***Panel C***: Plate diagram summary of the synthetic data experiment. At the participant level each agent is characterized by the parameters ρ, γ, ψ, τ, and β. The parameters τ and β, together with the specific lure trial's lure similarity level ℓ, determine the agent's lure discriminability at that trial. Given these parameters, each agent at each trial then extracts values of θ^O, θ^S, and θ^N, where $\theta^O + \theta^S + \theta^N = 1$. These θ values represent the probability that the agent will determine the specific trial stimulus as being either old, new, or similar, of which the determination itself is represented graphically as the decision node.

the level of similarity ℓ (represented as a value between 0 and 1) of the lure stimuli to a studied item in a manner that is unique to each agent. To this end, we implement a hierarchical extension as per Lee & Stark [27] which specifies additional parameters per agent, β and τ, such that each agent has a model of the relationship between discriminability (δ) and lure dissimilarity (ℓ) for lure stimuli only (See Figure 2.2B for a graphical depiction of the relationship):

$$\delta_\ell = \frac{1}{1 + \exp\{-\beta(\ell - \tau)\}} \quad \text{where} \quad 0 < \ell < 1 \tag{2.3}$$

This curve characterizes how discriminability increases as lure dissimilarity decreases, the β parameter adjusts the steepness of the curve, while the τ parameter shifts the curve horizontally. Note that the above function is also a logistic function similar to the one used to extract our λ and Δ indices. The key distinctions are that the function features fewer parameters and is only applied to lure stimuli (i.e., values of the function at similarity levels of 0 and 1 are not used anywhere). In this synthetic experiment, the above-described participant-level relation between discriminability and lure similarity contains the relevant information about mnemonic discrimination that both LDI and the novel λ measure aim to extract. To validate said efforts, we calculate each agent's ground truth λ by taking the area under the δ_ℓ curve.

Given the above structure, graphically summarized in Figure 2.2C, we simulated 1500 synthetic agents that were each assigned random values between 0 and 1 of ρ, γ, ψ, and τ, and a random value between 0 and 50 for β. Each agent is then given a simulated test list that contains 16 old, 16 new, and 16 lure stimuli (lure stimuli have an even distribution of each possible lure similarity values) to which the agents make discrete decisions on whether the stimuli are old, new, or similar.

Statistical analysis - Synthetic data

Each synthetic agent's LDI and REC scores were calculated in the same manner as with the empirical data: The LDI score is the proportion of correctly identified lures minus the proportion of foils incorrectly identified as lures. The REC score is the proportion of correctly identified old items minus the proportion of foils incorrectly identified as old items.

To calculate our novel indices, we used each synthetic agent's trial responses and the trials' dissimilarities to fit the previously described 5-parameter logistic function to the synthetic participant-level data. The fitted functions were then used to extract each agents' λ and Δ indices.

We considered that poor overall recognition could impact the degree to which the λ index could track both LDI, and the ground truth λ. To this end, we further examined these relations by subsetting the dataset by the agents' overall recognition. We conducted two additional analyses of these relations where the data only included synthetic agents with an Δ index of at least 0.4 for one analysis and 0.6 for the other (both are values non-excluded human participants generally exceed in the MST).

Chapter 3

Results

Empirical data

Table 3.1 and Figure 3.1C demonstrate a lack of collinearity between λ and Δ. The fixed effects coefficient of our linear mixed model did not show a statistically significant effect (β=0.21, 95% CI [-0.01, 0.43]; p = 0.061). The model exhibited a marginal R^2 of 0.042 and a conditional R^2 of 0.099, with an ICC of 0.06, suggesting consistency across the two studies. However, when considering the older and younger subgroups of the Wahlheim et al. [31] study as separate studies, this association strengthened (β=0.22, 95% CI [0.01, 0.43]; p = 0.039; Table S1 and Figure S2C). Table S2 and Figure S3C show the results of repeating these analyses after excluding one participant due to substantial missing responses (albeit falling just below the 2 standard deviation threshold we pre-specified in our methods). After excluding this participant, the fixed effects coefficient was similar in magnitude (β=0.20, 95% CI [-0.02, 0.42]; p = 0.077).

Table 3.1 also shows the results of our mixed effects modeling approach to verifying concordance between the original MST performance indices and our novel λ and Δ measures. As hypothesized, our λ measure showed a statistically significant association with the MST LDI index (fixed effects β=0.76, 95% CI [0.62-0.91], $p < 0.001$; Table 3.1 and Figure 3.1A), while not being significantly associated with the MST REC (fixed effects β=0.06, 95% CI [-0.03, 0.15], p = 0.197; Table 3.1 and Figure 3.1E). Similarly, our Δ measure showed significant association with the MST REC (fixed effects β=0.93, 95% CI [0.83, 1.02], $p < 0.001$; Table 3.1 and Figure 3.1B), while not being significantly associated with the MST LDI (fixed effects β=-0.06, 95% CI [-0.20, 0.09], p = 0.438; Table 3.1 and Figure 3.1D). The association between λ and LDI was highly consistent across studies. The association between Δ and REC showed a marginal R^2 of 0.830 and a conditional R^2 of 0.849. The intraclass correlation coefficient for the Δ and REC relationship was 0.11, suggesting that the

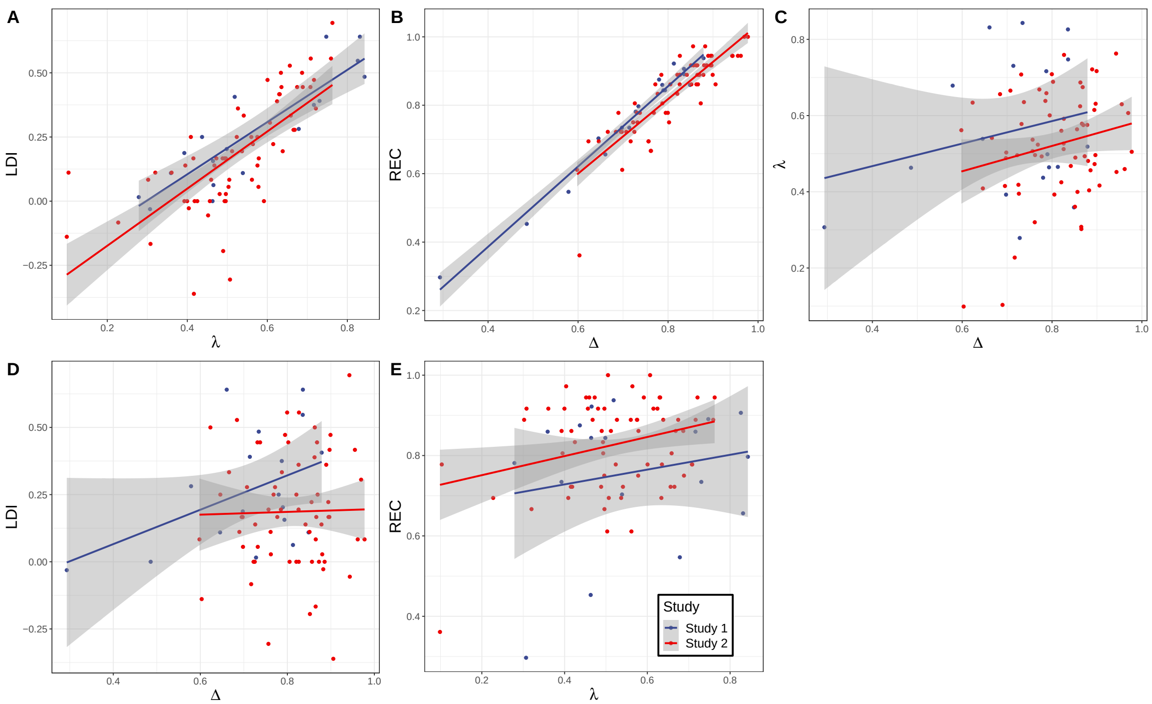

Figure 3.1: Pairwise comparisons of the original mnemonic similarity task (MST) measures and our novel measures. Colors indicate the study the data is taken from: Lee & Stark [27] in blue, Wahlheim et al. [31] in red. ***Panel A:*** The MST's lure discrimination index (LDI) and the novel λ measure. ***Panel B:*** The MST's recognition measure (REC) and the novel Δ measure. ***Panel C:*** The novel Δ and the novel λ measure. ***Panel D:*** The MST's LDI and the novel Δ measure. ***Panel E:*** The MST's REC and the novel λ measure.

association was relatively consistent between the two studies.

Tables S3 and S4 outline our approach using mixed-effects modeling to examine the concordance between the original MST performance indices and our λ and Δ measures, after the exclusion of an influential participant with considerable missing responses. For our λ measure, the results reveal a significant association with the MST LDI index (fixed effects β=0.77, 95% CI [0.62, 0.91], p $<$ 0.001; see Table S3 and Figure S3A). The Δ measure is not significantly associated with the MST LDI with a fixed effects coefficient of β=-0.08 (95% CI [-0.23, 0.07], p = 0.277; see Table S3 and Figure S3D). The association between λ and LDI remains highly consistent across studies, given a between-study variance of 0. Our Δ measure exhibited a significant association with the MST REC (fixed effects β=0.92, 95% CI [0.82, 1.01], p $<$ 0.001; see Table S4 and Figure S3B). The λ measure is not significantly associated with the MST REC, showing a fixed effect coefficient of β=0.06 (95% CI [-0.03, 0.15], p = 0.205; see Table S4 and Figure S3E). The association between Δ and REC demonstrated a marginal R^2 of 0.812 and a conditional R^2 of 0.836. An ICC of 0.13 suggests that this association was relatively consistent between the two studies.

The sensitivity analysis comparing linear models of LDI in the Lee & Stark [27] dataset with and without exclusion of the most similar lures in the λ and Δ calculations performed similarly to one another. The model without any lure exclusion (R^2 = 0.826) demonstrated a slightly stronger R^2 than the model with the excluded lures (R^2 = 0.783; See Table S5). The sensitivity analysis of the Wahlheim et al [31] dataset comparing neural network-derived dissimilarity measures with the standard MST lure bins was similar for both approaches in the association between the λ index and the Lure Discriminability Index (LDI). When using neural network-derived distances as the dissimilarity measure, the relationship between the λ index and LDI demonstrated a slightly stronger explanatory power (R^2 = 0.563, see Figure S4A and Table S6) compared to using lure bin values (R^2 = 0.531).

Synthetic data

Our analyses with the synthetic data confirmed a strong association between the Δ index and REC, exhibiting an R^2 of 0.923 (see Figure S5 and Table S7). The association between the λ index and LDI was initially moderate but statistically significant in the full dataset (β=0.55, 95% CI [0.51 – 0.59], $p < 0.001$). However, this association strengthened when analyzing only synthetic participants with higher Δ scores ($\Delta \geq 0.6$), where the effect size of λ on LDI increased to β=0.76 (95% CI [0.71 – 0.81], $p < 0.001$), as illustrated in Figure 3.2A. The relationship between the λ index and ground truth λ, a measure of ground truth mnemonic discrimination, similarly improved with higher Δ thresholds. In the subset with higher recognition capabilities, λ's predictive power on ground truth λ markedly increased from β=0.64 (95% CI [0.60 – 0.68], $p < 0.001$) in the full dataset to β=0.90 (95% CI [0.87 – 0.94], $p < 0.001$) in the high Δ subset, indicating that as the overall recognition capability of the agents improved, the novel λ index's ability to track both the LDI and the ground truth λ also improved (see Figure 3.2B). In line with the empirical findings, the synthetic analysis additionally confirmed the lack of a significant direct relationship between λ and Δ in both the full dataset and the subset with higher Δ scores (Table S7).

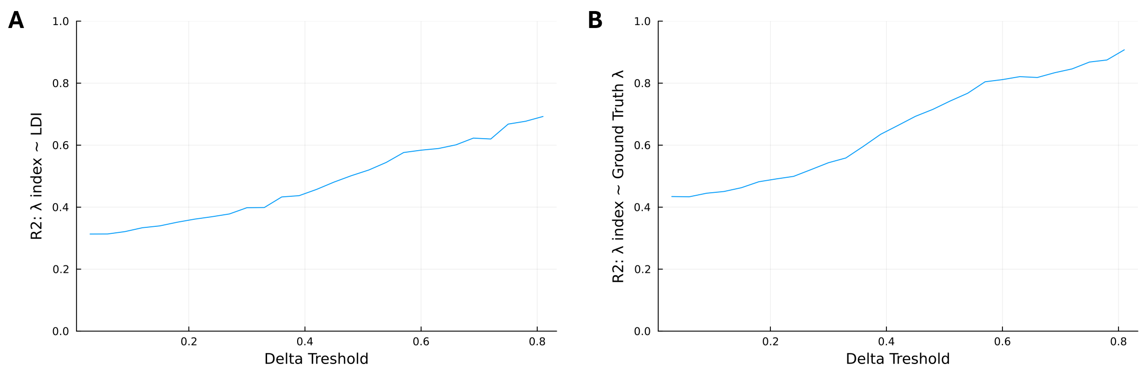

Figure 3.2: R^2 values of the relationships between the synthetic agent's λ index and **A)** the MST's lure discrimination index (LDI), and **B)** the ground truth λ calculated from the synthetic agent's two-parameter logistic function. Models were subsetted to only include synthetic agents with a Δ index above the specified Δ thresholds.

Table 3.1: *Mixed effects model results of Δ predicting λ scores, λ and Δ predicting mnemonic similarity task (MST) lure discrimination index (LDI) scores, and of λ and Δ predicting MST recognition index (REC) scores.*

Model 1: $\lambda \sim \Delta + (1\|study)$			
Predictors	*Estimates*	*CI*	*p*
Intercept	0.08	-0.35 – 0.50	0.713
Δ	0.21	-0.01 – 0.43	0.061
Random Effects			
σ^2	0.96		
τ_{00} *(study)*	0.06		
ICC	0.06		
N (study)	2		
Observations	85		
Marginal R^2 / Conditional R^2	0.042 / 0.099		
Model 2: $LDI \sim \lambda + \Delta + (1\|study)$			
Predictors	*Estimates*	*CI*	*p*
Intercept	-0.00	-0.14 – 0.14	1.000
λ	0.76	0.62 – 0.91	**<0.001**
Δ	-0.06	-0.20 – 0.09	0.438
Random Effects			
σ^2	0.44		
τ_{00} *(study)*	0.00		
Marginal R^2 / Conditional R^2	0.565 / NA		
Model 3: $REC \sim \lambda + \Delta + (1\|study)$			
Predictors	*Estimates*	*CI*	*p*
Intercept	0.05	-0.18 – 0.28	0.658
λ	0.06	-0.03 – 0.15	0.197
Δ	0.93	0.83 – 1.02	**<0.001**
Random Effects			
σ^2	0.16		
τ_{00} *(study)*	0.02		
ICC	0.11		
N (study)	2		
Observations	85		
Marginal R^2 / Conditional R^2	0.830 / 0.849		

Chapter 4

Discussion

Our study presents a novel measure of mnemonic discrimination, λ, which allows the study of mnemonic discrimination in recognition memory tasks without relying on categorical distinctions between "lures" and "foils" in test lists. A central motivation for this new metric lies in the fact that data are available for analysis from several recognition memory tests that have been administered within clinical populations [22, 35]. However, these tasks do not categorically distinguish lures and foils in their test images, precluding the application of the traditional MST analyses that calculate LDI and REC. Furthermore, if one were to impose categorical divisions of lures and foils onto test lists whose stimuli vary continuously in terms of perceptual or semantic interference, statistical power would be lost [36–46]. By developing a mnemonic discrimination measure that allows continuous variation among novel items in a recognition memory test list, we may address some of these challenges and facilitate the study of mnemonic discrimination using many readily available recognition memory tasks. However, prior to applying the λ index to each candidate task, future studies should assess the variance in the λ measure for each recognition memory test, through normative simulations such as those employed in the present study for the MST. This variance should be evaluated to (A) inform whether the test has the requisite number of trials to ensure sufficient power to find individual differences of a given effect, and (B) assess how many participants are needed to find a group difference of a given effect.

The primary finding in this study is the demonstration of both convergent and divergent validity for our new measures [30]. Specifically, we posited that λ and MST LDI should be strongly related, as should the Δ and MST REC measures. Our results reveal that λ is significantly associated with MST LDI but not with MST REC. Similarly, Δ is significantly associated with MST REC but not with MST LDI. This, in combination with the fact that λ and Δ are uncorrelated, suggests that they provide valid and well-separated markers of mnemonic discrimination (λ), and overall recognition memory performance (Δ).

The synthetic experiments further demonstrated the strong association between the Δ and MST REC, and λ and MST LDI measures, respectively. Interestingly, the strength of the relationship between λ and MST LDI was particularly high when the synthetic agent's Δ index was high. These results suggest that the reliability of λ increases with the magnitude of Δ.

While our measures offer an avenue for calculating mnemonic discrimination performance from recognition memory tasks beyond the MST, there are opportunities to test their potential further. For instance, given that clinical neuropsychiatric populations often undergo memory tests such as the CVLT, RAVLT, and HVLT, future investigations should consider applying our mnemonic discrimination measure λ to large clinical datasets that contain responses from such recognition memory tests [22]. Verbal learning tests in particular are important since impairments in this domain are common in clinical populations, such as in patients with bipolar disorder [47], and are predictive of functional outcomes [48, 49]. Harnessing data from established tools, such as the CVLT, could facilitate understanding mnemonic discrimination in clinically diverse and large sample populations, improving generalizability. It may also be useful for future studies to evaluate ways to adapt our method of analysis to forced-choice recognition tests.

Although common recognition memory tests are not explicitly designed to elicit high false alarm rates like the MST, variations in dissimilarity within these tests can still reveal information about mnemonic discrimination, conditional upon an absence of ceiling effects. We also demonstrated that our mnemonic discrimination measurement approach was robust to removing the most similar lures from the test set in MST data. Specifically, we found that despite excluding the 60% most similar lures (i.e. those with the highest false alarm rates) in the analysis, the λ index continued to show convergent validity with the MST's LDI among the participants in the Lee & Stark [27] dataset.

A limitation of our novel measures is that their performance will depend on the stimulus dissimilarity metric. In the present study, we rescaled the MST lure bins to reflect the dissimilarity of test items from those already viewed during encoding.

Here, these lure bins have been determined empirically based on difficulty of discrimination from old images [32]. This results in a "dissimilarity measure" that is ordinal, rather than perfectly continuous. However, we also showed that our measurement approach resulted in appropriate convergent and divergent validity when evaluated under an alternative continuous measure of dissimilarity between the images in the MST. Specifically, we calculated a "perceptual dissimilarity" between images in the MST by extracting the high-level abstract representations of those images from a deep-learning model pre-trained on a large corpus of natural images. Continuous dissimilarity between image embeddings in this abstract space may be more reflective of perceptually relevant differences between images, and could provide a viable alternative dissimilarity measure for estimating λ and Δ. When used as the relevant dissimilarity measure in the extraction of λ and Δ, this alternative measure of dissimilarity yielded similar results to the analysis using lure bins.

Continuous measures of dissimilarity may also be applied to verbal recognition memory tasks like the CVLT using language models to reflect semantic dissimilarity, but also simpler measures such as phonological and orthographic distance, or a combination of these measures to form an overall word dissimilarity index. What is most important in selecting a dissimilarity metric is that the metric used should reflect the discriminability of the stimuli. This may depend on the nature of the task. For example, in a verbal recognition memory task where the words are read aloud, a dissimilarity measure that particularly reflects phonological dissimilarity should be evaluated, in addition to dissimilarity measures reflective of semantic differences between words. While the current study utilized an ordinal dissimilarity measure, future research should prioritize continuous metrics that capture perceptually relevant differences between stimuli, tailored to the specific nature of the task.

Another limitation of the current study is the inclusion of only two real-world datasets, both of which had relatively small sample sizes. However, there is a paucity of freely available MST datasets. Given this, our novel analytical approach should enable analysis of many other existing recognition memory datasets, which although not from the MST proper, may now be amenable to analysis of mnemonic discrimination. Despite the limited availability of MST datasets, our estimates were highly consistent across the datasets utilized. Given this limited heterogeneity, we believe it is likely that our estimates would remain consistent if additional datasets were added. Nevertheless, as a further confirmatory step, it would be useful for future studies to evaluate our approach on additional MST data to further test the convergent and divergent validity of λ and Δ.

To summarize, our introduction of the λ index of mnemonic discrimination offers an alternative for analyzing mnemonic discrimination abilities using data from traditional recognition memory tasks that eliminate the need for categorical distinctions in participants' responses between lures and foils. Our findings demonstrate the λ index's convergent and divergent validity. The λ index has the potential to be more widely applicable, for example, to examine mnemonic discrimination using large datasets from clinical samples for a more comprehensive understanding of mnemonic discrimination in these populations.

Bibliography

[1] Patrick S R Davidson, Petar Vidjen, Sara Trincao-Batra, and Charles A Collin. Older adults' lure discrimination difficulties on the mnemonic similarity task are significantly correlated with their visual perception. *The Journals of Gerontology: Series B*, 74(8):1298–1307, Oct 2019.

[2] Chris M. Foster and Kelly S. Giovanello. Domain general processes moderate age-related performance differences on the mnemonic similarity task. *Memory*, 28(4):528–536, 2020.

[3] Adelaide Jensen, Galit Karpov, Charles A Collin, and Patrick S R Davidson. Executive function predicts older adults' lure discrimination difficulties on the mnemonic similarity task. *The Journals of Gerontology: Series B*, 78(10):1642–1650, Oct 2023.

[4] S. Pishdadian, N. V. Hoang, S. Baker, M. Moscovitch, and R. S. Rosenbaum. Not only memory: Investigating the sensitivity and specificity of the mnemonic similarity task in older adults. *Neuropsychologia*, 149:107670, Dec 2020.

[5] Stevenson Baker, Paula Vieweg, Fuqiang Gao, Asaf Gilboa, Thomas Wolbers, Sandra E. Black, and R. Shayna Rosenbaum. The human dentate gyrus plays a necessary role in discriminating new memories. *Current Biology*, 26(19):2629–2634, Oct 2016.

[6] Arnold Bakker, C. Brock Kirwan, Michael Miller, and Craig E. L. Stark. Pattern separation in the human hippocampal ca3 and dentate gyrus. *Science*, 319(5870):1640–1642, Mar 2008.

[7] David Berron, Hartmut Schütze, Anne Maass, Arturo Cardenas-Blanco, Hugo J. Kuijf, Dharshan Kumaran, and Emrah Düzel. Strong evidence for pattern separation in human dentate gyrus. *Journal of Neuroscience*, 36(29):7569–7579, Jul 2016.

[8] Jenna L. Klippenstein, Shauna M. Stark, Craig E. L. Stark, and Ilana J. Bennett. Neural substrates of mnemonic discrimination: A whole-brain fmri investigation. *Brain and Behavior*, 10(3):e01560, Feb 2020.

[9] Shauna M. Stark, Michael A. Yassa, Joyce W. Lacy, and Craig E.L. Stark. A task to assess behavioral pattern separation (bps) in humans: Data from healthy aging and mild cognitive impairment. *Neuropsychologia*, 51(12):2442–2449, Oct 2013.

[10] Michael A. Yassa, Joyce W. Lacy, Shauna M. Stark, Marilyn S. Albert, Michela Gallagher, and Craig E. L. Stark. Pattern separation deficits associated with increased hippocampal ca3 and dentate gyrus activity in nondemented older adults. *Hippocampus*, 21(9):968–979, 2011.

[11] Arnold Bakker, Gregory L. Krauss, Marilyn S. Albert, Caroline L. Speck, Lauren R. Jones, Craig E. Stark, Michael A. Yassa, Susan S. Bassett, Amy L. Shelton, and Michela Gallagher. Reduction of hippocampal hyperactivity improves cognition in amnestic mild cognitive impairment. *Neuron*, 74(3):467–474, May 2012.

[12] C. Brock Kirwan, Andrew Hartshorn, Shauna M. Stark, Naomi J. Goodrich-Hunsaker, Ramona O. Hopkins, and Craig E.L. Stark. Pattern separation deficits following damage to the hippocampus. *Neuropsychologia*, 50(10):2408–2414, Aug 2012.

[13] Annika Hanert, Anya Pedersen, and Thorsten Bartsch. Transient hippocampal ca1 lesions in humans impair pattern separation performance. *Hippocampus*, 29(8):736–747, 2019.

[14] Tanusree Das, Elena I. Ivleva, Anthony D. Wagner, Craig E.L. Stark, and Carol A. Tamminga. Loss of pattern separation performance in schizophrenia suggests dentate gyrus dysfunction. *Schizophrenia Research*, 159(1):193–197, Oct 2014.

[15] Nicolas Déry, Malcolm Pilgrim, Martin Gibala, Jenna Gillen, J. Martin Wojtowicz, Glenda MacQueen, and Suzanna Becker. Adult hippocampal neurogenesis reduces memory interference in humans: opposing effects of aerobic exercise and depression. *Frontiers in Neuroscience*, 7, 2013.

[16] Xue Han, Yingga Wu, Yanfeng Zhong, and Suzanna Becker. Recovery of high interference memory in spite of lingering cognitive deficits in a longitudinal pilot study of hospitalized depressed patients. *Frontiers in Psychiatry*, 11, 2020.

[17] Don J. Shelton and C. Brock Kirwan. A possible negative influence of depression on the ability to overcome memory interference. *Behavioural Brain Research*, 256:20–26, Nov 2013.

[18] Hannah Marlatte, Zorry Belchev, Madison Fraser, and Asaf Gilboa. The effect of hippocampal subfield damage on rapid temporal integration through statistical learning and associative inference. *Neuropsychologia*, 193:108755, Jan 2024.

[19] Helena Shizhe Wang, R. Shayna Rosenbaum, Stevenson Baker, Claire Lauzon, Laura J. Batterink, and Stefan Köhler. Dentate gyrus integrity is necessary for behavioral pattern separation but not statistical learning. *Journal of Cognitive Neuroscience*, 35(5):900–917, May 2023.

[20] Shauna M. Stark, C. Brock Kirwan, and Craig E. L. Stark. Mnemonic similarity task: A tool for assessing hippocampal integrity. *Trends in Cognitive Sciences*, 23(11):938–951, Nov 2019.

[21] Dean C Delis. *CVLT-II manual*. Pearson, Bloomington, Minn., 2000.

[22] Eamonn Kennedy, Shashank Vadlamani, Hannah M. Lindsey, Pui-Wa Lei, Mary Jo-Pugh, Maheen Adamson, Martin Alda, Silvia Alonso-Lana, Sonia Ambrogi, Tim J. Anderson, Celso Arango, Robert F. Asarnow, Mihai Avram, Rosa Ayesa-Arriola, Talin Babikian, Nerisa Banaj, Laura J. Bird, Stefan Borgwardt, Amy Brodtmann, Katharina Brosch, Karen Caeyenberghs, Vince D. Calhoun, Nancy D. Chiaravalloti, David X. Cifu, Benedicto Crespo-Facorro, John C. Dalrymple-Alford, Kristen Dams-O'Connor, Udo Dannlowski, David Darby, Nicholas Davenport, John DeLuca, Covadonga M. Diaz-Caneja, Seth G. Disner, Ekaterina Dobryakova, Stefan Ehrlich, Carrie Esopenko, Fabio Ferrarelli, Lea E. Frank, Carol Franz, Paola Fuentes-Claramonte, Helen Genova, Christopher C. Giza, Janik Goltermann, Dominik Grotegerd, Marius Gruber, Alfonso Gutierrez-Zotes, Minji Ha, Jan Haavik, Charles Hinkin, Kristen R. Hoskinson, Daniela Hubl, Andrei Irimia, Andreas Jansen, Michael Kaess, Xiaojian Kang, Kimbra Kenney, Barbora Keřková, Mohamed Salah Khlif, Minah Kim, Jochen Kindler, Tilo Kircher, Karolina Knížková, Knut K. Kolskår, Denise Krch, William S. Kremen, Taylor Kuhn, Veena Kumari, Jun Soo Kwon, Roberto Langella, Sarah Laskowitz, Jungha Lee, Jean Lengenfelder, Spencer W. Liebel, Victoria Liou-Johnson, Sara M. Lippa, Marianne Løvstad, Astri Lundervold, Cassandra Marotta, Craig A. Marquardt, Paulo Mattos, Ahmad Mayeli, Carrie R. McDonald, Susanne Meinert, Tracy R. Melzer, Jessica Merchán-Naranjo, Chantal Michel, Rajendra A. Morey, Benson Mwangi, Daniel J. Myall, Igor Nenadić, Mary R. Newsome, Abraham Nunes, Terence O'Brien, Viola Oertel, John Ollinger, Alexander Olsen, Victor Ortiz García de la Foz, Mustafa Ozmen, Heath Pardoe, Marise Parent, Fabrizio Piras, Federica Piras, Edith Pomarol-Clotet, Jonathan Repple, Geneviève Richard, Jonathan Rodriguez, Mabel Rodriguez, Kelly Rootes-Murdy, Jared Rowland, Nicholas P. Ryan, Raymond Salvador, Anne-Marthe Sanders, Andre Schmidt, Jair C. Soares, Gianfranco Spalleta, Filip Španiel, Alena Stasenko, Frederike Stein, Benjamin Straube, April Thames, Florian Thomas-Odenthal, Sophia I. Thomopoulos, Erin Tone, Ivan Torres, Maya Troyanskaya, Jessica A. Turner, Kristine M. Ulrichsen, Guillermo Umpierrez, Elisabet Vilella, Lucy Vivash, William C. Walker, Emilio Werden, Lars T. Westlye, Krista Wild, Adrian Wroblewski, Mon-Ju Wu, Glenn R. Wylie, Lakshmi N. Yatham, Giovana B. Zunta-Soares, Paul M. Thompson, David F. Tate, Frank G. Hillary, Emily L. Dennis, and Elisabeth A. Wilde. Bridging big data: Procedures for combining non-equivalent cognitive measures from the enigma consortium. page 2023.01.16.524331, Apr 2023.

[23] Dhanyaasri Madanlal, Christian Guinard, Vanessa Pardo Nuñez, Suzanna Becker, Julie Garnham, Anouar Khayachi, Simon Léger, Claire O'Donovan, Selena Singh, Shani Stern, Claire Slaney, Thomas Trappenberg, Martin Alda, and Abraham Nunes. A pilot study examining the impact of lithium treatment and responsiveness on mnemonic discrimination in bipolar disorder. *Journal of Affective Disorders*, 351:49–57, Apr 2024.

[24] Martin Alda, Margaret McKinnon, Ryan Blagdon, Julie Garnham, Susan MacLellan, Claire O'Donovan, Tomas Hajek, Cynthia Nair, Serdar Dursun, and Glenda MacQueen. Methylene blue treatment for residual symptoms of bipolar disorder: Randomised crossover study. *The British Journal of Psychiatry*, 210(1):54–60, Jan 2017.

[25] Katherine E. Burdick, Caitlin E. Millett, Manuela Russo, Martin Alda, Ney Alliey-Rodriguez, Amit Anand, Yokesh Balaraman, Wade Berrettini, Holli Bertram, Joseph R. Calabrese, Cynthia Calkin, Carla Conroy, William Coryell, Anna DeModena, Scott Feeder, Carrie Fisher, Nicole Frazier, Mark Frye, Keming Gao, Julie Garnham, Elliot S. Gershon, Kara Glazer, Fernando S. Goes, Toyomi Goto, Gloria J. Harrington, Petter Jakobsen, Masoud Kamali, Marisa Kelly, Susan Leckband, Else Marie Løberg, Falk W. Lohoff, Adam X. Maihofer, Michael J. McCarthy, Melvin McInnis, Gunnar Morken, Caroline M. Nievergelt, John Nurnberger, Ketil J. Oedegaard, Abigail Ortiz, Megan Ritchey, Kelly Ryan, Martha Schinagle, Candice Schwebel, Martha Shaw, Paul Shilling, Claire Slaney, Emma Stapp, Bruce Tarwater, Peter Zandi, and John R. Kelsoe. The association between lithium use and neurocognitive performance in patients with bipolar disorder. *Neuropsychopharmacology*, 45(10):1743–1749, Sep 2020. 2020 The Author(s), under exclusive licence to American College of Neuropsychopharmacology.

[26] Stephanie A. Cardenas, Layla Kassem, Melissa A. Brotman, Ellen Leibenluft, and Francis J. McMahon. Neurocognitive functioning in euthymic patients with bipolar disorder and unaffected relatives: A review of the literature. *Neuroscience & Biobehavioral Reviews*, 69:193–215, Oct 2016.

[27] Michael D. Lee and Craig E. L. Stark. Bayesian modeling of the mnemonic similarity task using multinomial processing trees. *Behaviormetrika*, 50(2):517–539, Jul 2023.

[28] Rita E. Loiotile and Susan M. Courtney. A signal detection theory analysis of behavioral pattern separation paradigms. *Learning & Memory*, 22(8):364–369, Aug 2015. Company: Cold Spring Harbor Laboratory PressDistributor: Cold Spring Harbor Laboratory PressInstitution: Cold Spring Harbor Laboratory PressLabel: Cold Spring Harbor Laboratory Presspublisher: Cold Spring Harbor LabPMID: 26179230.

[29] Jessica Bean. *Rey Auditory Verbal Learning Test, Rey AVLT*, page 2174–2175. Springer New York, New York, NY, 2011.

[30] Kevin D. Carlson and Andrew O. Herdman. Understanding the impact of convergent validity on research results. *Organizational Research Methods*, 15(1):17–32, Jan 2012.

[31] Christopher N. Wahlheim, Alexander P. Christensen, Zachariah M. Reagh, and Brittany S. Cassidy. Intrinsic functional connectivity in the default mode network predicts mnemonic discrimination: A connectome-based modeling approach. *Hippocampus*, 32(1):21–37, Jan 2022.

[32] Joyce W. Lacy, Michael A. Yassa, Shauna M. Stark, L. Tugan Muftuler, and Craig E. L. Stark. Distinct pattern separation related transfer functions in human ca3/dentate and ca1 revealed using high-resolution fmri and variable mnemonic similarity. *Learning & Memory*, 18(1):15–18, Jan 2011. Company: Cold Spring Harbor Laboratory PressDistributor: Cold Spring Harbor Laboratory PressInstitution: Cold Spring Harbor Laboratory PressLabel: Cold Spring Harbor Laboratory Presspublisher: Cold Spring Harbor LabPMID: 21164173.

[33] Douglas Bates, Martin Machler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67:1–48, 2015.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 770–778, Las Vegas, NV, USA, Jun 2016. IEEE.

[35] George C. Nitzburg, Armando Cuesta-Diaz, Luz H. Ospina, Manuela Russo, Megan Shanahan, Mercedes Perez-Rodriguez, Emmett Larsen, Sandra Mulaimovic, and Katherine E. Burdick. Organizational learning strategies and verbal memory deficits in bipolar disorder. *Journal of the International Neuropsychological Society: JINS*, 23(4):358–366, Apr 2017.

[36] DG Altman and P Royston. The cost of dichotomising continuous variables. *BMJ*, 332(7549):1080, May 2006.

[37] J Cohen. The cost of dichotomization. *Applied Psychological Measurement*, 7(3):249–253, Jun 1983.

[38] GJ Fitzsimons. Death to dichotomizing. *Journal of Consumer Research*, 35(1):5–8, Jun 2008.

[39] LG Humphreys. Doing research the hard way: Substituting analysis of variance for a problem in correlational analysis. *Journal of Educational Psychology*, 70(6):873–876, 1978.

[40] LG Humphreys. Research on individual differences requires correlational analysis, not anova. *Intelligence*, 2(1):1–5, Jan 1978.

[41] LG Humphreys and A Fleishman. Pseudo-orthogonal and other analysis of variance designs involving individual-differences variables. *Journal of Educational Psychology*, 66(4):464–472, Aug 1974.

[42] JR. Irwin and GH McClelland. Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research*, 40(3):366–371, 2003.

[43] RC MacCallum, S Zhang, KJ Preacher, and DD Rucker. On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7(1):19–40, 2002.

[44] P Royston, DG Altman, and W Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, 25(1):127–141, 2006.

[45] DD. Rucker, BB McShane, and KJ Preacher. A researcher's guide to regression, discretization, and median splits of continuous variables. *Journal of Consumer Psychology*, 25(4):666–678, Oct 2015.

[46] DL Streiner. Breaking up is hard to do: The heartbreak of dichotomizing continuous data. *The Canadian Journal of Psychiatry*, 47(3):262–266, Apr 2002.

[47] Breda Cullen, Joey Ward, Nicholas A. Graham, Ian J. Deary, Jill P. Pell, Daniel J. Smith, and Jonathan J. Evans. Prevalence and correlates of cognitive impairment in euthymic adults with bipolar disorder: A systematic review. *Journal of Affective Disorders*, 205:165–181, Nov 2016.

[48] Colin A Depp, Brent T Mausbach, Alexandrea L Harmell, Gauri N Savla, Christopher R Bowie, Philip D Harvey, and Thomas L Patterson. Meta-analysis of the association between cognitive abilities and everyday functioning in bipolar disorder. *Bipolar Disorders*, 14(3):217–226, 2012.

[49] Samson Tse, Sunny Chan, King Lam Ng, and Lakshmi N Yatham. Meta-analysis of predictors of favorable employment outcomes among individuals with bipolar disorder. *Bipolar Disorders*, 16(3):217–229, 2014.

Appendix - Supplementary Materials

Spread Analysis

To examine the robustness of the measures to the number of trials, additional simulations were conducted where the number of trials was systematically altered. At each set number of trials, 20 different sets of 50 agents with identical parameters were generated. Then, for each set of agents in which the agents have the same parameters, but through random discrete choices have variance in outcomes between them, we examined their variance in several outcome measures in relation to the number of trials the agents underwent. This analysis was repeated with a subsetted group of agents who had a Δ index of at least 0.6.

Supplementary Results

Our investigation into the associations between various simulation parameters and the novel indices revealed the following: the Δ index is primarily associated with the parameters ρ (probability of recognizing old stimuli) and ψ (probability of remembering that an item was not studied), showing its sensitivity to changes in recognition memory performance (see Figure S6). The λ index showed a primary association with the parameter τ (the threshold for discriminability in relation to lure similarity), but it did not show a significant relationship with β (which determines the steepness of the logistic function relating discriminability to lure similarity), suggesting that λ's sensitivity is more attuned to the discriminability threshold than to variations in the steepness of discriminability across different levels of lure similarity (see Figure S7). The spread analysis showed the variation in results that can be seen among identical synthetic agents in relation to the number of trials the agents underwent (see Figures S8 and S9). Without any participant exclusions, all measures except the λ index showed similar decreases in spread with increasing number of trials (Figure S8). The λ index in contrast demonstrated considerably less dropoff in spread with increasing trials. Interestingly, when subsetting the agents to allow only those with moderately high recognition (Δ index of at least 0.6; Figure S9), all measures across all number of trials showed less variation, and the λ index showed a similar dropoff to LDI with increasing trials.

Supplementary Figures

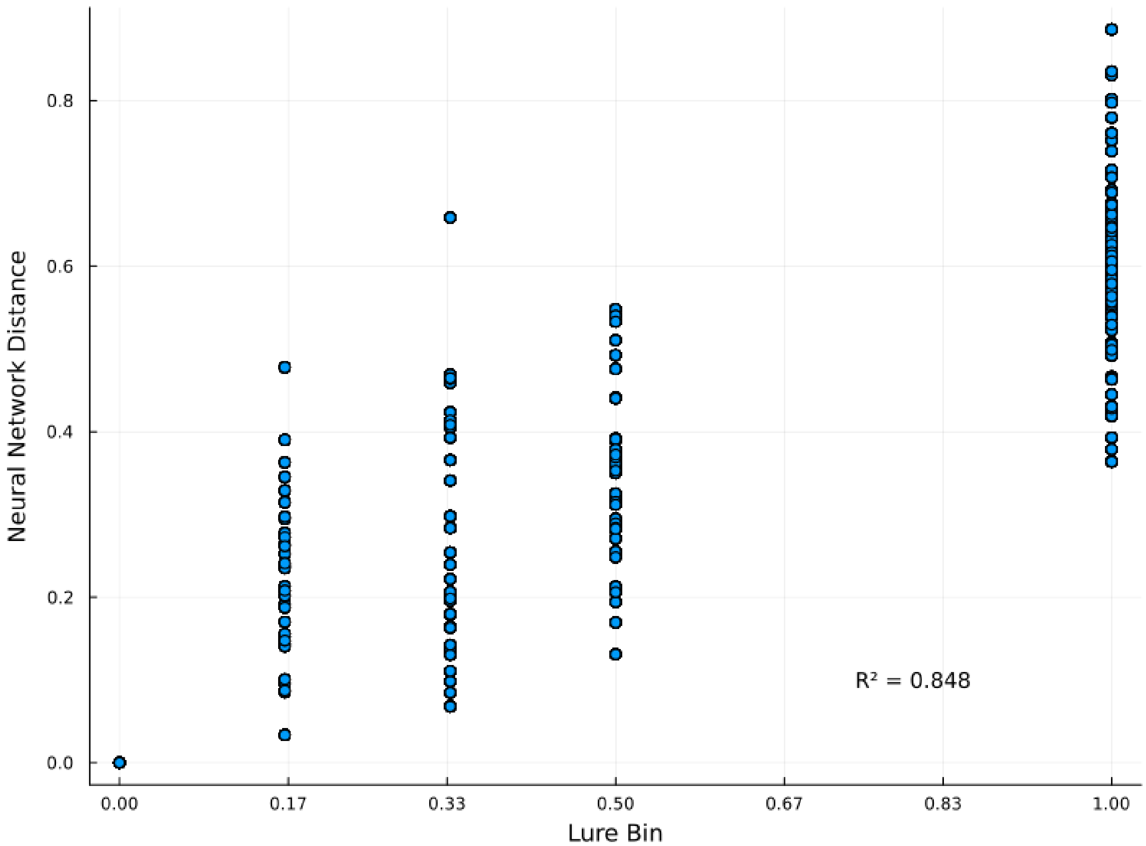

Figure S1: Neural network-derived distance measure compared to the original lure bin in the Wahlheim et al. [31] mnemonic similarity task dataset. Both measures are scaled to be between 0 and 1.

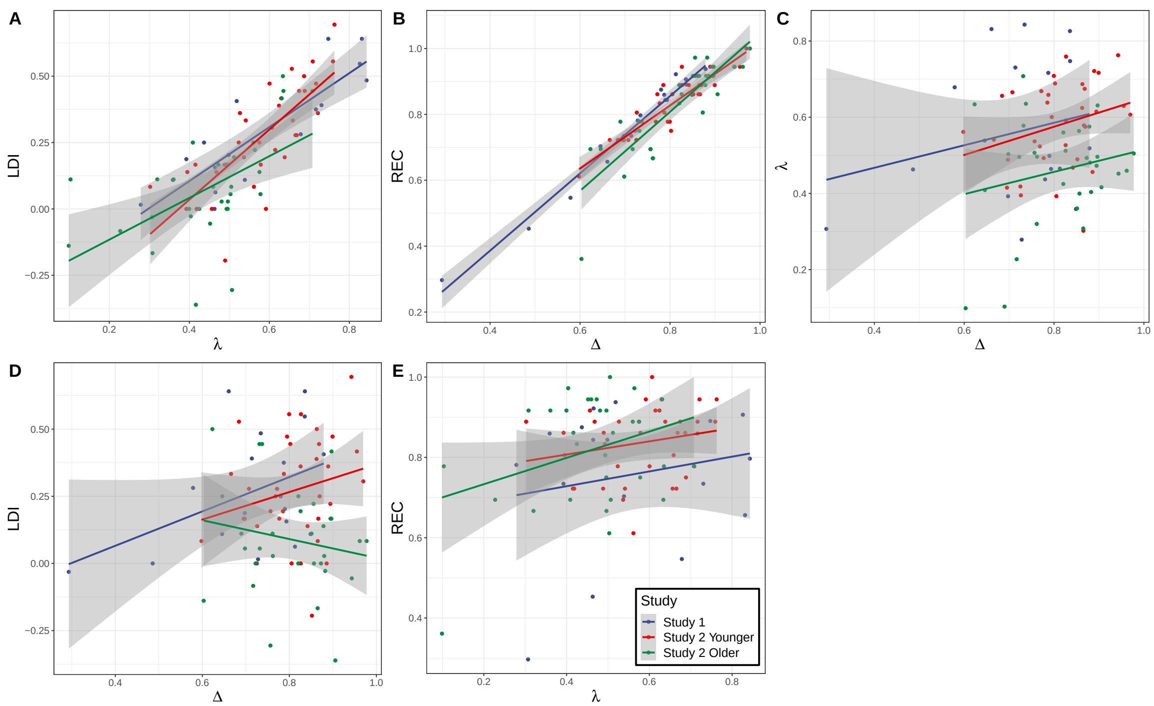

Figure S2: Pairwise comparisons of the original mnemonic similarity task (MST) measures and our novel measures. Older and Younger subgroups in the Wahlheim et al. [31] study were treated as separate studies. Colors indicate the study the data is taken from: Lee & Stark [27] in blue, Wahlheim et al. [31] in red (Younger cohort) and green (Older cohort). ***Panel A:*** The MST's lure discrimination index (LDI) and the novel λ measure. ***Panel B:*** The MST's recognition measure (REC) and the novel Δ measure. ***Panel C:*** The novel Δ and the novel λ measure. ***Panel D:*** The MST's LDI and the novel Δ measure. ***Panel E:*** The MST's REC and the novel λ measure.

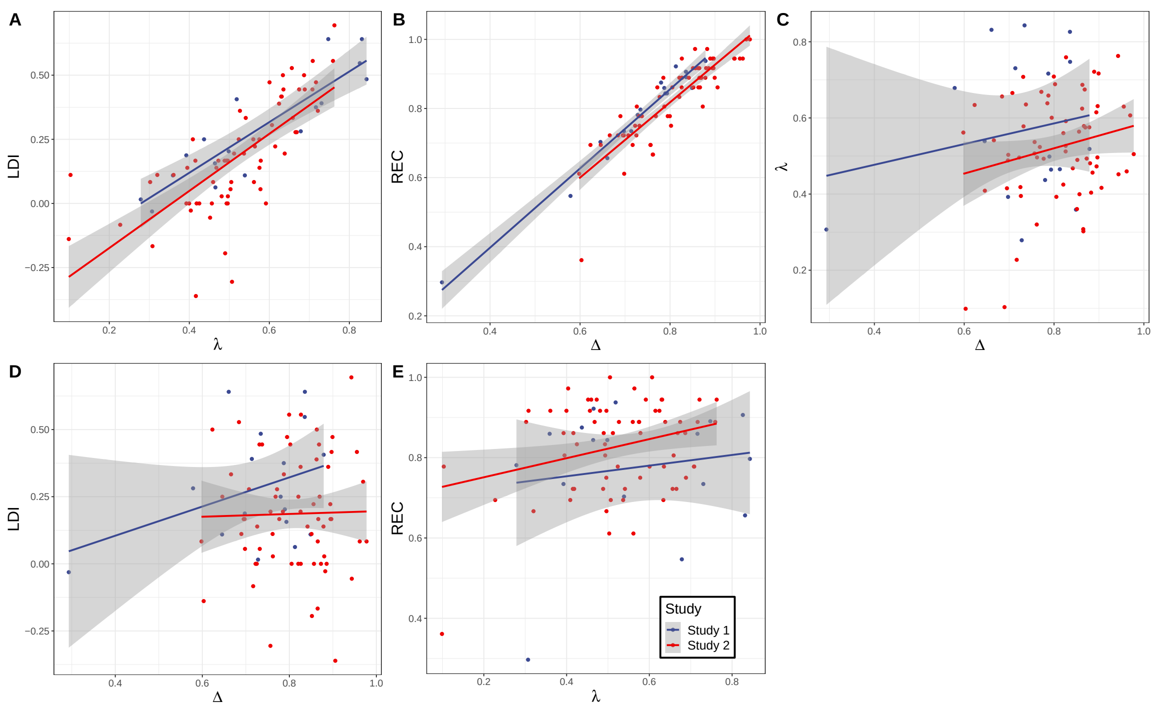

Figure S3: Pairwise comparisons of the original mnemonic similarity task (MST) measures and our novel measures with an additional participant excluded in the analysis due to considerable missing responses. Colors indicate the study the data is taken from: Lee & Stark [27] in blue, Wahlheim et al. [31] in red. ***Panel A:*** The MST's lure discrimination index (LDI) and the novel λ measure. ***Panel B:*** The MST's recognition (REC) and the novel Δ measure. ***Panel C:*** The novel Δ and the novel λ measure. ***Panel D:*** The MST's LDI and the novel Δ measure. ***Panel E:*** The MST's REC and the novel λ measure.

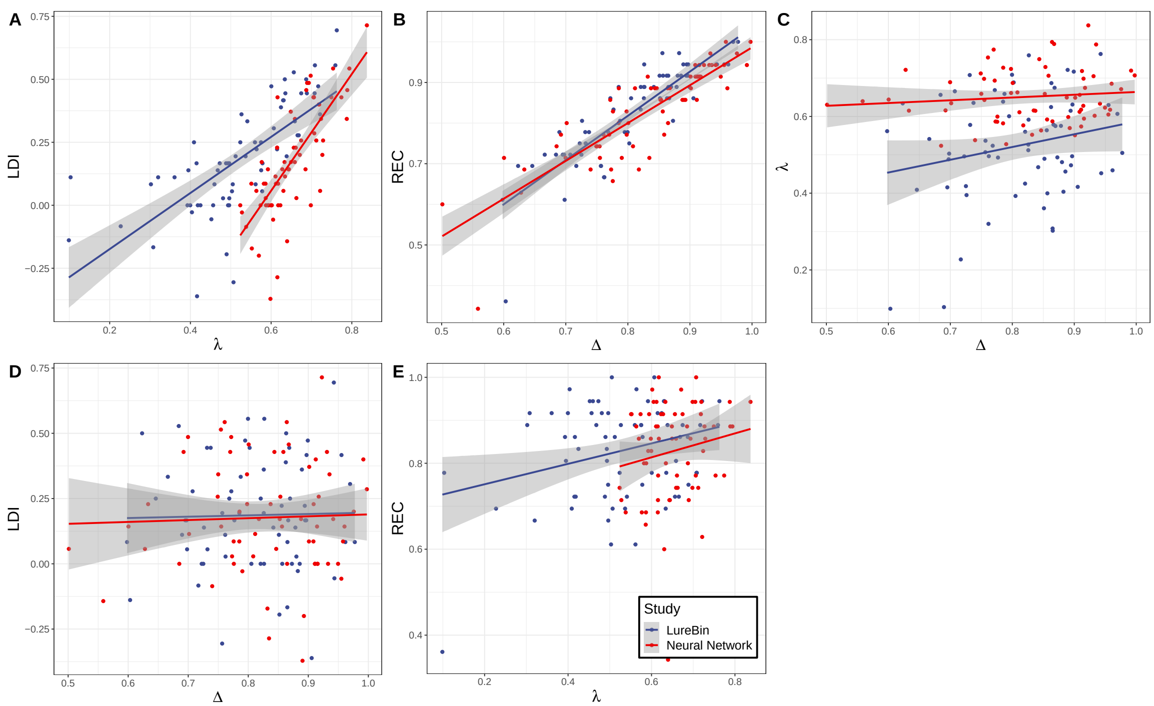

Figure S4: Pairwise comparisons of the original mnemonic similarity task (MST) measures and our novel measures using the Wahlheim et al. 2021 dataset using two different dissimilarity measures. Colors indicate the dissimilarity measure that was used to fit the logistic function. ***Panel A:*** The MST's lure discrimination index (LDI) and the novel λ measure. ***Panel B:*** The MST's recognition measure (REC) and the novel Δ measure. ***Panel C:*** The novel Δ and the novel λ measure. ***Panel D:*** The MST's LDI and the novel Δ measure. ***Panel E:*** The MST's REC and the novel λ measure.

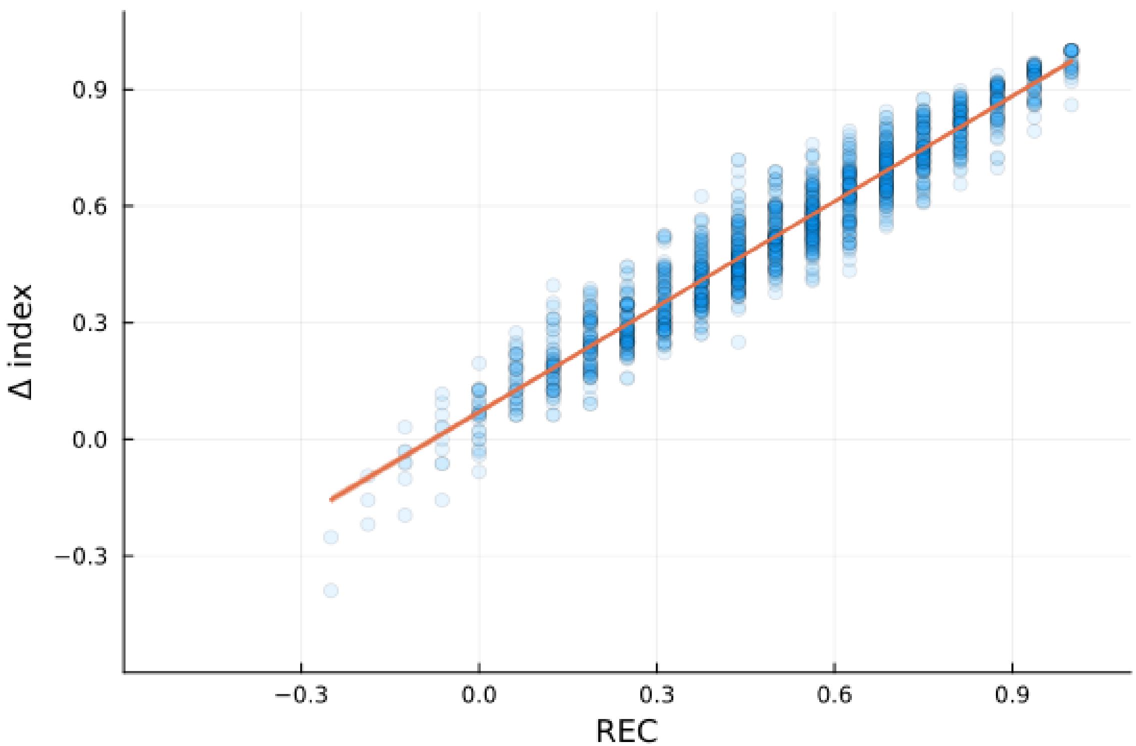

Figure S5: Relationship between the novel Δ index and the mnemonic similarity task's recognition measure (REC) in the synthetic data experiment.

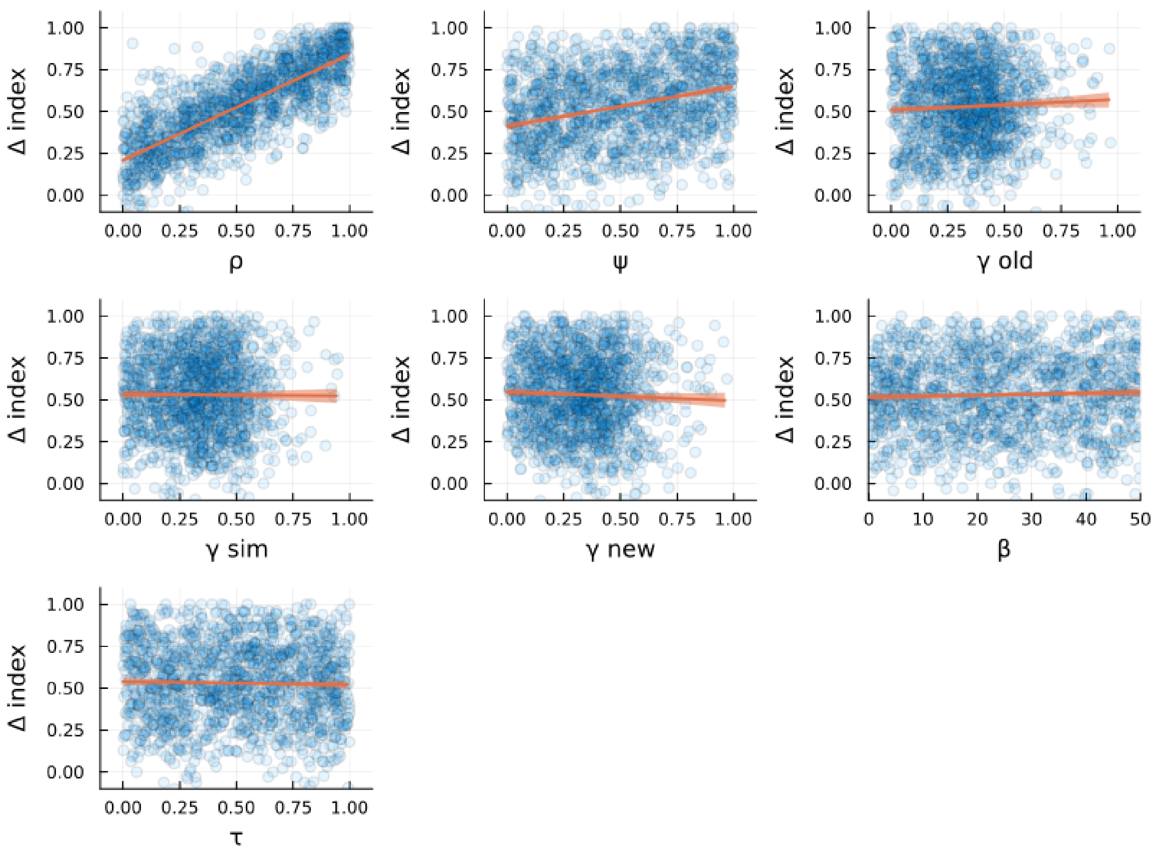

Figure S6: Associations between the Δ index all the synthetic agent's parameters.

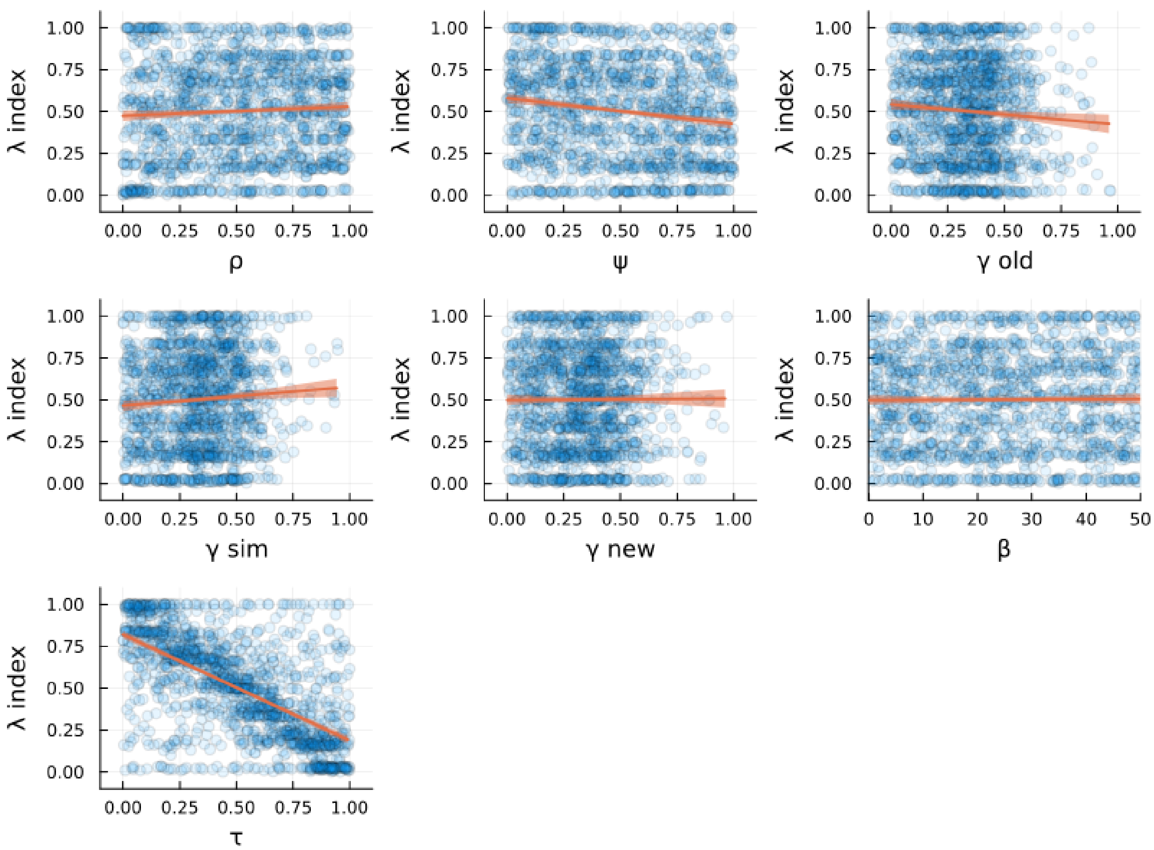

Figure S7: Associations between the λ index all the synthetic agent's parameters.

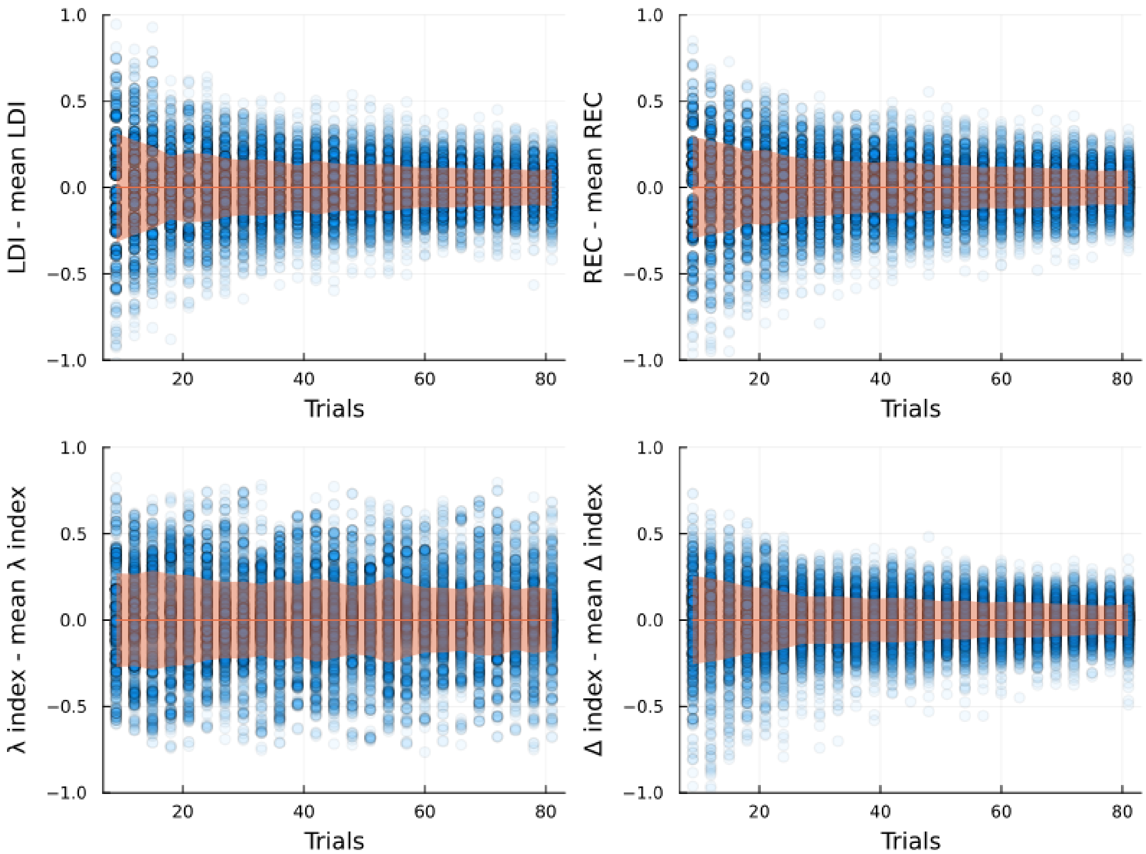

Figure S8: Variance of the lure discrimination index (LDI), recognition (REC), λ, and Δ measures in relation to the number of test trials undergone by synthetic agents. 50 copies of 20 unique synthetic agents underwent the simulated mnemonic similarity task-like experiment at varying number of trials. The number of trials ranged from 6 to 81 at increments of 3. Each plot point represents a specific agent's score compared to the average of its copies who underwent the same number of test trials. Ribbons show the standard deviation among the agents who underwent the same number of test trials.

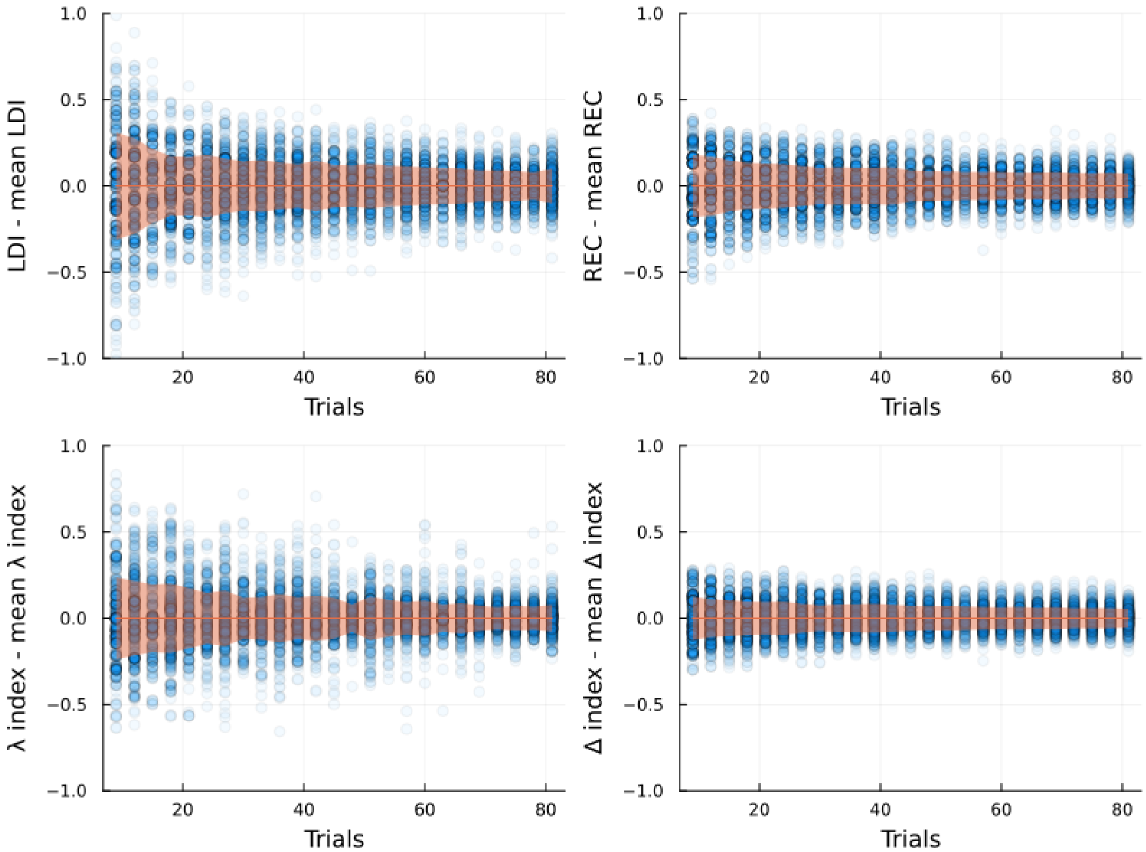

Figure S9: Variance of the lure discrimination index (LDI), recognition (REC), λ, and Δ measures in relation to the number of test trials undergone by synthetic agents with a Δ index of at least 0.6. 50 copies of 20 unique synthetic agents underwent the simulated mnemonic similarity task-like experiment at varying number of trials, with those who obtained a Δ index below 0.6 removed. The number of trials ranged from 6 to 81 at increments of 3. Each plot point represents a specific agent's score compared to the average of its copies who underwent the same number of test trials. Ribbons show the standard deviation among the agents who underwent the same number of test trials.

Supplementary Tables

Table S1: *Mixed effects model results of Δ predicting λ scores, λ and Δ predicting mnemonic similarity task (MST) lure discrimination index (LDI) scores, and of λ and Δ predicting MST recognition index (REC) scores. Older and Younger subgroups in the Wahlheim et al. [31] study were treated as separate studies.*

Model 1: $\lambda \sim \Delta + (1\|study)$			
Predictors	*Estimates*	*CI*	*p*
Intercept	0.03	-0.53 – 0.58	0.926
Δ	0.22	0.01 – 0.43	0.039
Random Effects			
σ^2	0.84		
τ_{00} *(study)*	0.20		
ICC	0.19		
N (study)	3		
Observations	85		
Marginal R^2 / Conditional R^2	0.045 / 0.231		
Model 2: LDI $\sim \lambda + \Delta + (1\|study)$			
Predictors	*Estimates*	*CI*	*p*
Intercept	0.00	-0.19 – 0.19	0.970
λ	0.74	0.59 – 0.89	**<0.001**
Δ	-0.05	-0.19 – 0.10	0.540
Random Effects			
σ^2	0.43		
τ_{00} *(study)*	0.01		
ICC	0.03		
Marginal R^2 / Conditional R^2	0.548 / 0.560		
Model 3: REC $\sim \lambda + \Delta + (1\|study)$			
Predictors	*Estimates*	*CI*	*p*
Intercept	0.01	-0.14 – 0.16	0.850
λ	0.05	-0.04 – 0.15	0.273
Δ	0.92	0.83 – 1.01	**<0.001**
Random Effects			
σ^2	0.16		
τ_{00} *(study)*	0.01		
ICC	0.06		
N (study)	3		
Observations	85		
Marginal R^2 / Conditional R^2	0.834 / 0.845		

Table S2: *Mixed effects model results of Δ predicting λ scores with an additional participant excluded in the analysis due to considerable missing responses.*

Model: $\lambda \sim \Delta + (1\|study)$			
Predictors	*Estimates*	*CI*	*p*
Intercept	0.08	-0.35 – 0.51	0.705
Δ	0.20	-0.02 – 0.42	0.077
Random Effects			
σ^2	0.97		
τ_{00} *(study)*	0.06		
ICC	0.06		
N (study)	2		
Observations	84		
Marginal R^2 / Conditional R^2	0.037 / 0.095		

Table S3: *Mixed effects model results of λ and Δ predicting original MST LDI scores with an additional participant excluded in the analysis due to considerable missing responses.*

Model: $LDI \sim \lambda + \Delta + (1\|study)$			
Predictors	*Estimates*	*CI*	*p*
Intercept	0.00	-0.14 – 0.15	0.996
λ	0.77	0.62 – 0.91	**<0.001**
Δ	-0.08	-0.23 – 0.07	0.277
Random Effects			
σ^2	0.44		
τ_{00} *(study)*	0.00		
N (study)	2		
Observations	84		
Marginal R^2 / Conditional R^2	0.568 / 0.568		

Table S4: *Mixed effects model results of λ and Δ predicting original MST REC scores with an additional participant excluded in the analysis due to considerable missing responses.*

Model: $REC \sim \lambda + \Delta + (1\|study)$			
Predictors	*Estimates*	*CI*	*p*
Intercept	0.06	-0.19 – 0.31	0.640
λ	0.06	-0.03 – 0.15	0.205
Δ	0.92	0.82 – 1.01	**<0.001**
Random Effects			
σ^2	0.17		
τ_{00} *(study)*	0.03		
ICC	0.13		
N (study)	2		
Observations	84		
Marginal R^2 / Conditional R^2	0.812 / 0.836		

Table S5: *Linear models applied to the Lee & Stark [27] dataset with the novel λ and Δ indices derived from the dataset with and without exclusion of the most similar lure trials (lure bins 1, 2, and 3). Note that both models use an LDI calculated without any lure exclusions.*

Model: $LDI \sim \lambda + \Delta$, All lures present			
Predictors	*Estimates*	*CI*	*p*
Intercept	0.00	-0.22 – 0.22	1.000
λ	0.82	0.59 – 1.06	**<0.001**
Δ	0.24	0.01 – 0.48	**0.044**
Observations 18			
Marginal R^2 / Conditional R^2 : 0.826 / 0.803			
Model: $LDI \sim \lambda + \Delta$, Lurebins 1-3 removed			
Intercept	-0.00	-0.25 – 0.25	1.000

λ	0.80	0.54 – 1.06	**<0.001**
Δ	0.27	0.00 – 0.53	**0.047**
Observations 18			
Marginal R^2 / Conditional R^2 0.783 / 0.754			

Table S6: *Linear models applied to the Wahlheim et al [31] dataset using two different distance measures for lure stimulus: Lure Bin (left), Neural network-derived embeddings (right).*

Model: LDI $\sim \lambda + \Delta$, Distance measure = Lure Bin			
Predictors	*Estimates*	*CI*	*p*
Intercept	0.00	-0.17 – 0.17	1.000
λ	0.75	0.57 – 0.92	**<0.001**
Δ	-0.14	-0.32 – 0.03	0.111
Observations 67			
Marginal R^2 / Conditional R^2 : 0.531 / 0.517			
Model: LDI $\sim \lambda + \Delta$, Distance measure: Neural Net			
Intercept	0.00	-0.16 – 0.16	1.000
λ	0.75	0.59 – 0.92	**<0.001**
Δ	-0.05	-0.21 – 0.12	0.574
Observations 67			
Marginal R^2 / Conditional R^2 0.563 / 0.549			

Table S7: *Linear model results of synthetic data experiment with and without the removal of low Δ synthetic participants.*

Model: LDI $\sim \lambda + \Delta$			
Predictors	*Estimates*	*CI*	*p*
Intercept	0.00	-0.04 – 0.04	1.000
λ	0.55	0.51 – 0.59	**<0.001**
Δ	0.24	0.20 – 0.28	**<0.001**

Observations 1488			
Marginal R^2 / Conditional R^2 : 0.352 / 0.352			
Model: LDI $\sim \lambda + \Delta$, *Subset:* $\Delta \geq 0.6$			
Intercept	0.00	-0.04 – 0.04	1.000
λ	0.76	0.71 – 0.81	**<0.001**
Δ	0.12	0.07 – 0.17	**<0.001**
Observations 609			
Marginal R^2 / Conditional R^2 : 0.599 / 0.597			

Model: Baseline $\lambda \sim \lambda + \Delta$			
Predictors	*Estimates*	*CI*	*p*
Intercept	-0.00	-0.04 – 0.04	1.000
λ	0.64	0.60 – 0.68	**<0.001**
Δ	0.04	0.00 – 0.08	**0.043**
Observations 1488			
Marginal R^2 / Conditional R^2 : 0.409 / 0.408			
Model: Baseline $\lambda \sim \lambda + \Delta$, *Subset:* $\Delta \geq 0.6$			
Intercept	0.00	-0.03 – 0.03	1.000
λ	0.90	0.87 – 0.94	**<0.001**
Δ	0.03	-0.01 – 0.06	0.154
Observations 609			
Marginal R^2 / Conditional R^2 : 0.812 / 0.811			

Model: REC $\sim \lambda + \Delta$			
Predictors	*Estimates*	*CI*	*p*
Intercept	-0.00	-0.01 – 0.01	1.000
λ	0.00	-0.01 – 0.02	0.805
Δ	0.96	0.95 – 0.97	**<0.001**
Observations 1488			
Marginal R^2 / Conditional R^2 : 0.923 / 0.923			
Model: REC $\sim \lambda + \Delta$, *Subset:* $\Delta \geq 0.6$			
Intercept	-0.00	-0.04 – 0.04	1.000

λ	-0.12	-0.16 – -0.08	**<0.001**
Δ	0.85	0.81 – 0.89	**<0.001**
Observations 609			
Marginal R^2 / Conditional R^2 : 0.740 / 0.739			

Model: $\lambda \sim \Delta$

Predictors	*Estimates*	*CI*	*p*
Intercept	-0.00	-0.05 – 0.05	1.000
Δ	-0.03	-0.08 – 0.02	0.207
Observations 1488			
Marginal R^2 / Conditional R^2 : 0.001 / 0.000			

Model: $\lambda \sim \Delta$, *Subset:* $\Delta \geq 0.6$

Intercept	0.00	-0.08 – 0.08	1.000
Δ	0.01	-0.07 – 0.09	0.845
Observations 609			
Marginal R^2 / Conditional R^2 : 0.000 / -0.002			