Response time modeling provides stable and mechanistically interpretable measures of individual differences in behavioral pattern separation.

Nidhi V. Banavar (nbanavar@uci.edu)
Aaron M. Bornstein (aaron.bornstein@uci.edu)
Department of Cognitive Sciences, University of California, Irvine
Irvine, CA 92617 USA

Abstract

The incredible specificity and fidelity of human memory encoding is thought to be supported by a process known as *pattern separation* (Marr, 1971). Behaviorally, this is typically inferred via performance in the Mnemonic Similarity Task (MST; (Stark, Kirwan, & Stark, 2019)), an object recognition task with added similar "lure" images, from which a key metric, the Lure Discrimination Index (LDI) is calculated. Supported by an extensive literature validating its predictive power, this measure is gaining increasing use as a diagnostic of cognitive decline and neurological dysfunction. It is however unclear the exact mechanism through which this behavioral measure of pattern separation reflects the underlying neural computations. In particular, choices alone cannot in principle distinguish the degree to which a given behavior results from signal-based discrimination of the object in question (i.e. the putative separated patterns) versus a more general tendency to inhibit or excite responses (e.g. response caution). Here, we distinguish these potentially co-contributing factors by modeling response times using a sequential sampling framework that identifies independent contributions to choices made by signal-noise discrimination and response thresholding. Across two independent datasets encompassing a lifespan sample (total $N = 307$, ages 8-89), we find evidence that both factors reliably contribute to response behavior, but that signal discrimination is both more strongly correlated with Lure and Foil discrimination and more stable within-individual than response thresholding, suggesting that this model-derived parameter may be a more specific and reliable measure of the underlying trait of interest in studies of pattern separation.

Keywords: memory and discrimination; evidence accumulation; recognition

Introduction

How do individuals encode objects in memory, and how does the distinctiveness of encoding affect behavioral expressions of recognition? These functions are thought to be supported by a process known as *pattern separation*, whereby similar sensory or latent input patterns are projected into higher-dimensional space to create highly distinct patterns that support later discrimination among fine degrees of difference (Stark et al., 2019). Traditionally, this process has been attributed to the hippocampus, a critical brain structure for learning and memory (Long, Lee, & Kuhl, 2016; Marr, 1971; Stark et al., 2019). Computational models predict that the more distinct object representations are (i.e. the "better" an individual is at pattern separating), the better an individual will be able to *discriminate* between objects that were seen previously and those that weren't. In particular, people who are better at pattern separating should be less susceptible to interference when novel objects are similar to the previously seen objects.

The most widely used behavioral measure of pattern separation, known as the Lure Discrimination Index (LDI), stems from the 3AFC *Mnemonic Similarity Task* (MST), a modified object recognition task (Stark et al., 2019). In the typical version of this experiment, individuals first complete a learning phase where they study a collection of object pictures. Then, during the recognition phase, individuals see a series of objects of one of three types: *repeats*, or objects they had seen before during learning; *lures*, which vary in degrees of similarity to the repeats; and *foils*, which are totally new objects never seen before in the experiment. Thus responses on these three trials can be analyzed to quantify how sensitively an individual discriminates between what they have, and have not seen before. This measure, the LDI, has been shown to correlate with standard behavioral and physical measures of cognitive decline and neurological dysfunction (Stark et al., 2019).

It is however an open question as to what aspects of recognition memory behavior are measured by the LDI. Specifically, it is unclear to what degree LDI solely reflects the actual "separation" of the underlying memory representations (in Signal Detection Theory terms, the separation between signal and noise distributions), versus more general response selection processes (e.g. the threshold for response execution). To the extent that LDI is indeed a measure consistent with hippocampal pattern separation, we would predict the latter: that it would correspond with an increase in signal to noise ratio (Long et al., 2016).

Sequential sampling models of response time provide an excellent method to assess these separable influences on recognition memory. This family of models, specifically the Linear Ballistica Accumulator which we use in this paper, robustly distinguishes separable contributions to behavior of both signal-noise separation (as *drift rate*) and response execution (as *threshold/boundary* or *starting point*) (Brown & Heathcote, 2008).

Here, we model response times to examine the relationship between LDI and components of the recognition memory process. We find evidence for both processes contributing to measured LDI, examine their relative contributions to choices, and assess their ability to predict behavior out-of-sample. Our results support the suggestion that LDI can be decomposed to isolate a stable, separable signal-based measure of memory discrimination. This measure may further improve the reliability and precision with which clinical practitioners can assess a key transdiagnostic process underlying a wide array of disorders and neurological conditions.

Methods

Data and Experiments

We model two data sets of individuals that completed the Mnemonic Similarity Task (MST). In this task, participants initially completed an "encoding" phase where they categorized unique objects as either belonging indoors or outdoors. They were also told that they would be subsequently tested on their memory of these objects.

Then, participants made a sequence of recognition choices during the "test" phase where they identified each object as either a repeat (seen before during the encoding phase), lure (similar to an object seen during encoding), or foil (a brand new object). Participants saw $\frac{1}{3}$ repeated objects, $\frac{1}{3}$ lures, and $\frac{1}{3}$ foils. There was no feedback after each choice (i.e. participants were not informed if their choice was accurate or not) and subjects had up to 10s to make a choice. The presentation order was fully randomized.

Experiment 1 We model $n = 223$ adult subjects (ages $18-89$, median = 41, 141 female). Subjects saw 128 trials during the encoding phase, and made 192 recognition judgements during the test phase. The data was collected in two modalities: online via Amazon mTurk ($n = 173$) and in person ($n = 72$).

Experiment 2 We model $n = 84$ subjects (ages $8-25$, median = 15, 53 female). Subjects saw 64 trials during the encoding phase, and made 96 recognition judgements during the test phase. The data was all collected online via Amazon mTurk. All participant ages in Experiment 2 were verified using photographs of government-issued identification cards.

Choice Behavior Measures

To quantify memory discriminability, we compute the Lure Discrimination Index (LDI) as in (Stark et al., 2019).

$$\begin{aligned} LDI = P(\ &Lure\ Response \mid Lure\ Trial\ ) \\ &- P(\ Lure\ Response \mid Foil\ Trial\ ) \end{aligned} \tag{1}$$

The LDI provides a sensitive measure of how reliably an individual distinguishes object photographs that were seen during the encoding phase from similar 'lures' presented during the test phase. This measure is typically interpreted as robust in that the more distinctly an individual encodes a previously seen object, the less they will subject to interference from *both* similar lures and unrelated foils. We further compute an individual's Recognition Score (RS), which quantifies how well someone remembers previously seen objects:

$$\begin{aligned} RS = P(\ &Repeat\ Response \mid Repeat\ Trial\ ) \\ &- P(\ Repeat\ Response \mid Foil\ Trial\ ) \end{aligned} \tag{2}$$

Response Time Modeling

We model response times (RT) using a Linear Ballistic Accumulator model (LBA) (Brown & Heathcote, 2008). The LBA is a powerful sequential sampling model that differs from other sequential sampling models in the following critical ways: a) it can fit n responses (nAFC), b) it assumes that evidence in favor of each alternative is accrued independently, and c) that evidence accumulation itself is linear and noiseless. The LBA does remarkably well in fitting response times and recovers standard patterns in RT data (Brown & Heathcote, 2008).

We use the R package *rtdists* (Singmann et al., 2018) to implement the LBA. We adhere to the assumptions of the most simple LBA in that we allow each individual to have the same starting point bias (A), evidence boundary (b, with $b > A$), and non-decision time (t_0). However, we allow for the drift rates to vary by accumulator (3 accumulators for 3 response types) and apply the scaling constraint that all drift rates must sum to 1 (i.e. $\Sigma_{i=1}^{3} v_i = 1$). Drift rates are drawn from a Normal distribution which has a common standard deviation (sv) across all three accumulators. We use Maximum Likelihood Estimation (MLE) to fit all parameters to individual subjects.

Results

In *Experiment 1*, we excluded a total of 20 subjects (13 for below chance accuracy, 7 for LDI scores below zero) resulting in a total of 255 subjects with valid data. In *Experiment 2*, we excluded a total of 10 subjects (5 for below chance accuracy, 5 for LDI scores below zero) resulting in a total of 74 subjects with valid data.

Choice Behavior

In *Experiment 1*, individuals chose the correct response 71% of the time. They were most often correct on *Repeat* trials (40% of correct responses) and *Foil* trials (38 %), followed by *Lure* trials (22%). In *Experiment 2*, individuals also chose the correct response 71% of the time. They were most often correct on *Repeat* trials (39% of correct responses) and *Foil* trials (38 %), followed by *Lure* trials (23%). LDIs were comparable across experiments ($median_{E1,E2} = 0.37(.3)$, Figure 1). Recognition scores were similarly comparable ($median_{E1} = 0.78(.16), median_{E2} = 0.78(.19)$).

Response Time Modeling

In *Experiment 1*, median (IQR) RTs for each response type were as follows: $Repeat = 1.14(0.42)$, $Lure = 1.29(0.47)$, and $Foil = 1.16(0.46)$. In *Experiment 2*, median (IQR) RTs for each response type were as follows: $Repeat = 1.07(0.43)$, $Lure = 1.29(0.43)$, and $Foil = 1.12(0.45)$.

Our LBA parameter inferences are presented in Table 1.

Both experiments show the highest median drift rate on the Repeat accumulator, followed by the Foil accumulator, and lastly the Lure accumulator. Both experiments show that subjects have the same median *response caution*, which is often defined as the difference between the boundary and starting point ($b - A$, $median = 0.28$).

We next confirmed qualitatively that our model had good descriptive adequacy. To do this, we overlaid predicted RT quantiles on observed RT quantiles

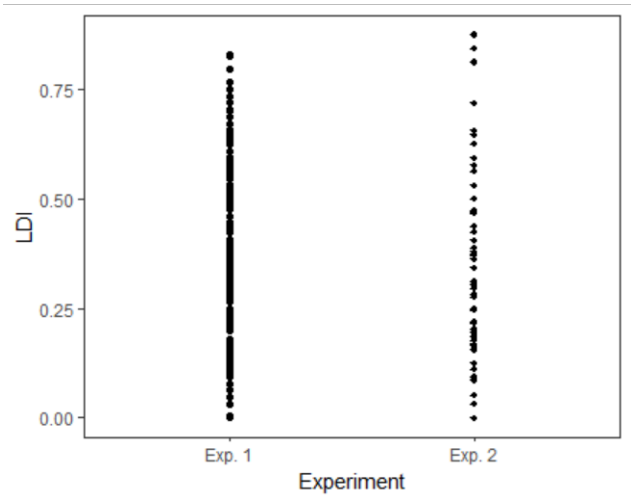

Figure 1: Lure Discrimination Indices for both experiments *median(IQR)* = 0.37(0.3).

Parameter	Exp. 1	Exp. 2
Starting Point (A)	0.42(0.21)	0.45(0.28)
Boundary (b)	0.70(0.21)	0.73(0.27)
Non Decision Time (t_0)	0.45(0.22)	0.39(0.20)
Drift: v_{Repeat}	0.39(0.12)	0.39(0.13)
Drift: v_{Lure}	0.26(0.15)	0.27(0.13)
Drift: v_{Foil}	0.36(0.06)	0.34(0.06)
Drift: Standard Deviation	0.24(0.32)	0.24(0.27)

Table 1: Maximum Likelihood Estimates (*median(IQR)*) for LBA parameters for both experiments. We fit a total of 6 parameters and the seventh, drift rate for the Foil accumulator is $v_{Foil} = 1 - v_{Repeat} - v_{Lure}$.

$(10\%, 30\%, 50\%, 70\%, 90\%)$. We present an example of subject fits across ages and correct/incorrect responses in Figure 2, noting that most subjects were qualitatively well fit by the data.

Relating LBA to MST

As our key question of interest focuses on relating LBA parameters (components of an individual's memory retrieval and recognition processes – in particular drift rates and boundary) to how distinctly people encode memories, we assessed whether there were any correlations between the LBA parameters and behavioral scores (LDI and RS). We report Kendall's τ rank correlation coefficient in the following analyses and adjust for multiple comparisons using the Bonferroni-Holm correction.

We found significant correlations between drift rates and LDI as shown in Figure 3. In particular, we found a negative correlation between the drift rate for the Repeat Accumulator and LDI in *Experiment 1* ($\tau_{\text{Kendall}} = -0.276, p < 0.01$) and *Experiment 2* ($\tau_{\text{Kendall}} = -0.20, p < 0.05$) trials. We further

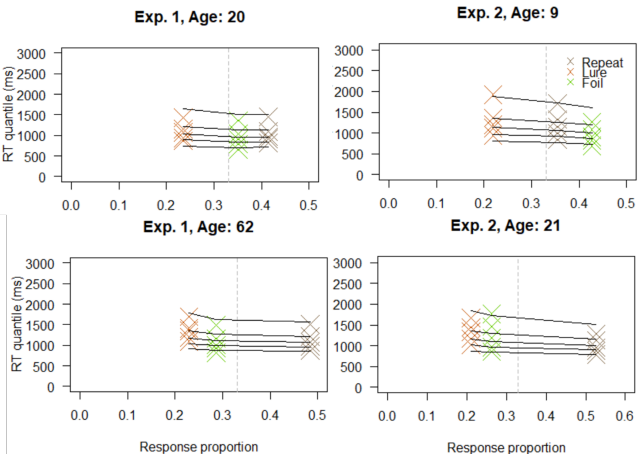

Figure 2: Example plots where observed quantiles are overlaid with predicted quantiles for subjects old and young, correct and incorrect. Purple markers are observed RT quantiles for repeat trials, red for lures, and green for foils. Black lines are predictions from LBA. The horizontal vertical line represents the true proportion of repeat, lure, and foil trials ($\frac{1}{3}$).

found a positive correlation between the Lure Accumulator drift rate and LDI in *Experiment 1* ($\tau_{\text{Kendall}} = 0.15, p < 0.01$). Finally, the correlations between drift rates for the Foil Accumulator and LDIs in *Experiment 1* or *Experiment 2* were not significant after adjusting for multiple comparisons.

We also observed a significant negative correlation between response caution ($b - A$) and LDI ($\tau_{\text{Kendall}} = -0.14, p < 0.05$) in *Experiment 1* only.

Correlation Strengths To compare correlation strengths, we used bootstrapping to resample the data and calculate Kendall's τs and the differences between each pair of τs (e.g. $\tau_A - \tau_b$). We then examined whether the bootstrapped 95% confidence interval distributions of the differences between each pair of correlations included zero. If they did not include zero, we interpreted this as evidence in favor of a non-zero difference between the correlations compared.

Critically, we found that in *Experiment 1*, all three of the bootstrapped distributions of correlation differences between LDI and boundary, and LDI and the three accumulator drift rates did not include zero: boundary-Repeat $(0.0973, 0.282)$, boundary-Lure $(-0.492, -0.218)$, boundary-Foil $(-0.412, -0.1479)$, *Figure* 4. We note that the CIs go in opposite directions for the Repeat vs Lure and Foil accumulators because of the negative correlation between LDI and Repeat accumulator drift rates. These results also held when we compared correlation strengths between response caution and the three accumulator drift rates: response caution-Repeat $(0.054, 0.265)$, response caution-Lure $(-0.456, -0.137)$, response caution-Foil $(-0.397, -0.139)$. In *Experiment 2*, however, all of the CIs contained zero: boundary-Repeat $(-0.139, 0.298)$, boundary-Lure $(-.451, 0.052)$, boundary-Foil $(-0.350, 0.105)$. Again, the

same held for response caution: response caution-Repeat $(-0.142, 0.272)$, response caution-Lure $(-0.312, 0.101)$, response caution-Foil $(-0.345, 0.125)$

We also found that the correlations between the drift rate accumulators and LDIs were significantly different in *Experiment 1*. Specifically, the LDI-repeat accumulator thresholds were stronger than the LDI-lure accumulator drift (-0.680, -0.326) and the LDI-foil accumulator drift (-0.565,-0.309). We further found that the correlation between LDI-lure accumulator drift was stronger than the LDI-foil accumulator drift (0.026,0.357). In *Experiment 2*, we only found that the LDI-repeat accumulator drift correlation was significantly greater than the LDI-foil accumulator drift (-0.381,-0.043).

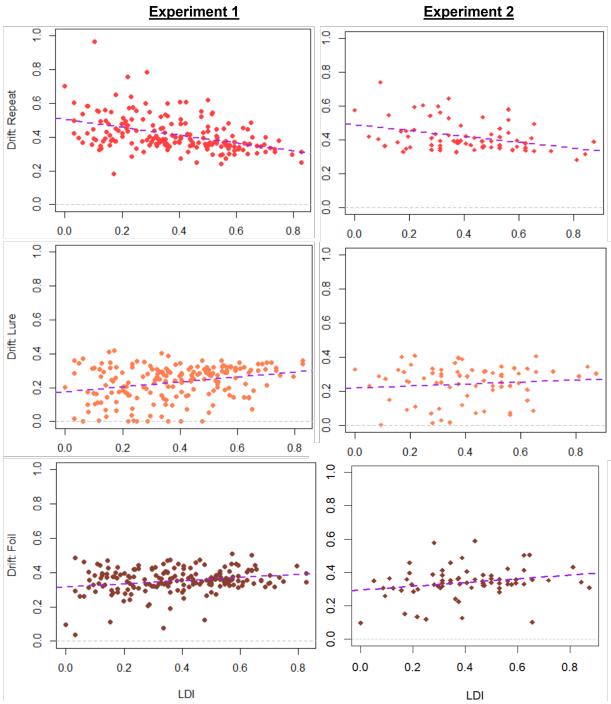

Figure 3: Correlations between Accumulator drift rates and the LDI across both experiments. We find statistically significant correlations between the drift rate of the Repeat accumulator and LDI in both experiments ($\tau_{E1} = -0.276$, $\tau_{E2} = -0.26$). We further find a significant correlation between the drift rate of the Lure accumulator and LDI in Experiment 1 ($\tau_{E1} = 0.15$)

Stability of Measures

Given the correlation between LDI and drift rates in both experiments, we wanted to see if the drift rate may in fact be a more stable behavioral measure than LDI. To evaluate the stability of the fit parameters and behavioral measures, we performed a split-halves analysis. Specifically, for each subject, we separately estimated each parameter and metric of interest on randomly selected halves of trials. We then computed the Mean Square Error for all parameters fit (both in the response time modeling and in choice behavior), Table 2. Specifically,

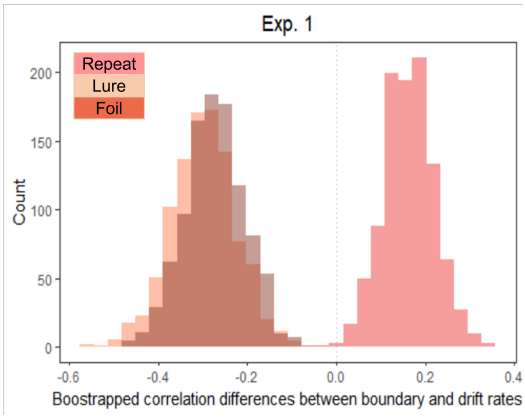

Figure 4: Boostrapped correlation differences between boundary and LDI, and drift rate and LDI for the three different accumulators in *Experiment 1*. All three 95% CIs do not include zero: boundary-Repeat $(0.0973, 0.282)$, boundary-Lure $(-0.492, -0.218)$, boundary-Foil $(-0.412, -0.1479)$.

we calculated LBA measures, LDI, and Recognition Scores twice for for all odd numbered trials, and all even numbered trials separately.

Parameter	Exp.1	Exp. 2
Starting Point (A)	0.049(0.014)	0.045(0.023)
Boundary (b)	0.028(0.011)	0.034(0.020)
Non Decision Time (t_0)	0.042(0.013)	0.112(0.036)
Drift: v_{Repeat}	0.009(0.006)	0.026(0.017)
Drift: v_{Lure}	0.010(0.007)	0.019(0.015)
Drift: v_{Foil}	0.007(0.005)	0.021(0.016)
Drift: Standard Deviation	0.04(0.013)	0.067(0.028)
Lure Discrimination Index	0.017(0.008)	0.034(0.020)
Recognition Score	0.008(0.006)	0.018(0.014)

Table 2: Mean square errors (Standard Error) for all parameters estimated by the LBA model and (below the line) for standard behavioral measures derived from the MST.

Supporting the hypothesis that signal discrimination is a stable measure within-individual, we found that the MSE of the drift rates for all the accumulators were the lowest in both experiments. We note that the degree of stability is an order of magnitude greater than all the other parameters in *Experiment 1*, the larger dataset with more trials per subject. To quantify differences between MSE across LBA and behavioral parameters (i.e. stability in measurements), we use the non-parametric paired Wilcoxon Rank Sum test and again correct for multiple comparisons using the Bonferroni-Holm correction. We found that the drift rates were more stable than all other LBA parameters ($p < 0.01$) and both behavioral parameters (LDI, Recognition Score $p < 0.01$) in *Experiment 1*. In *Experiment 2*, we found that drift rates were significantly more stable than all the LBA parameters ($p < 0.01$) except

the non-decision time, which was trendingly significant after correcting for multiple comparisons ($0.05 < p < 0.08$). However, like in *Experiment 1*, the drift rates were more stable than both behavioral parameters ($p < 0.01$).

Discussion

We present one of the first model based analysis of response times in the Mnemonic Similarity Task (MST). We use a simple sequential sampling model, the Linear Ballistic Accumulator (LBA), where evidence is accumulated independently for all three possible responses.

Our approach decomposed responses for this task into separable components of response execution and signal detection, allowing us to assess the individual stability of these processes, across subjects. We hypothesized that either or both the response caution (either boundary, b, alone or boundary minus starting point, $b-A$) or drift rate,v_i, to lure or foil trials would be key variables of interest for behavioral discrimination performance. Specifically, if the LDI is indeed a measure of pattern separation, we would expect higher drift rates on Lure and/or Foil accumulators, suggesting a boosted signal. At the same time, to the extent LDI reflects individual variability in response caution, boundary, or starting point bias, then this would be reflected in these terms.

We found that, although both parameters were significantly correlated with LDI, the drift rates were both a stronger predictor of the standard behavioral measure and also a more stable within-subject measurement. The latter point is of considerable interest given the extensive evidence that MST is a useful individual difference marker, predicting neurological dysfunction and cognitive performance in a wide variety of clinical and laboratory measures (Stark et al., 2019).

The finding that LDI is strongly influenced by evidence strength supports the suggestion that MST measures the degree of pattern separation underlying these responses. Further, our findings may enhance the application of MST in several ways. First, the finding that drift rates are a more stable within-subject measure suggests that it could be used to more finely predict the same sorts of outcomes currently predicted by LDI. Future work should examine the correspondence of this drift rate to cognitive and neurological outcomes of interest. Second, the use of sequential sampling models can enable extracting trial-by-trial timeseries reflecting putative underlying computations that drive behavior, which should support analysis of more precisely defined functional neuroimaging measures (Long et al., 2016). Finally, the robust statistical frameworks often used to fit these sorts of models may allow further refinement of the approach, producing even more stable trait-level estimates by, e.g., incorporating informative priors and models of contaminant behavior.

In sum, we have provided initial evidence that joint modeling of choices and response times can improve inference of trait-level properties underlying a widely used clinical and laboratory assessment tool. Future work will examine the robustness of this new metric in the many settings in which the MST has been applied.

Acknowledgements

The authors are grateful to Sharon M. Noh for providing data for Experiment 1, and Nora C. Harhen for providing data for Experiment 2. This work was supported by NIMH P50MH096889, NIA R21AG072673, and a NARSAD Young Investigator Award from the Brain and Behavior Research Foundation (AMB).

References

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive psychology*, *57*(3), 153–178.

Long, N. M., Lee, H., & Kuhl, B. A. (2016). Hippocampal mismatch signals are modulated by the strength of neural predictions and their similarity to outcomes. *Journal of Neuroscience*, *36*(50), 12677–12687.

Marr, D. (1971). Simple memory: a theory for archicortex. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *262*(841), 23–81.

Singmann, H., Brown, S., Gretton, M., Heathcote, A., Voss, A., Voss, J., & Terry, A. (2018). *Package 'rtdists'*.

Stark, S. M., Kirwan, C. B., & Stark, C. E. (2019). Mnemonic similarity task: A tool for assessing hippocampal integrity. *Trends in cognitive sciences*, *23*(11), 938–951.