Brief Communication

A signal detection theory analysis of behavioral pattern separation paradigms

Rita E. Loiotile¹ and Susan M. Courtney^{1,2,3}¹Department of Psychological and Brain Sciences, ²Department of Neuroscience, Johns Hopkins University, Baltimore, Maryland 21218, USA; ³F. M. Kirby Research Center, Kennedy Krieger Institute, Baltimore, Maryland 21218, USA

Behavioral pattern separation (BPS) paradigms ask participants to discriminate previously encoded (old) stimuli from highly similar (lure) and categorically distinct (novel) stimuli. The lure-old discrimination, thought to uniquely reflect pattern separation in the hippocampal formation, is typically pitted against the traditional novel-old discrimination. However, BPS paradigms have measured lure-old discrimination neither consistently across studies nor in such a way that allows for accurate comparison to novel-old discrimination. Therefore, we advocate for signal detection theory (SDT) as a unified framework. Moreover, we compare SDT with previously used measures of lure-old discrimination, indicating how other formulas' inaccuracies can lead to erroneous conclusions.

Tasks measuring behavioral pattern separation (BPS) are distinguished from standard tests of memory recognition by the inclusion of stimuli that are highly similar to those previously encoded. Because each of these "lure" stimuli differs only slightly from its corresponding previously encoded "old" stimulus, successful discrimination of old from lure stimuli is thought to denote particularly detailed memory representations of the old stimuli. In contrast, it has been suggested that successful discrimination of old from "novel" stimuli (i.e., categorically new, as commonly used in tests of recognition) need not rely on detailed representations of the old stimuli, but rather can be achieved even if the old stimuli have only gist-based representations. Consistent with this distinction, several studies imply a dissociation between novel-old and lure-old discrimination. For example, using the BPS paradigm, researchers have argued that only detail-based lure-old discrimination is impaired in healthy aging (Stark et al. 2013) and enhanced by post-study caffeine (Borota et al. 2014). Claims from a more recent study (Reagh and Yassa 2014) support an even stronger, double dissociation: repeated-encoding benefits gist-based, novel-old discrimination (a finding known even to Ebbinghaus (1885)), but surprisingly impairs detail-based, lure-old discrimination. One major concern—particularly with findings such as these, where there appear to be differences between novel-old and lure-old discrimination—is that the measures used for lure-old discrimination vary from each other and from those used for novel-old discrimination. A unified measurement approach is needed if one is to reliably characterize lure-old discrimination as well as properly compare it to novel-old discrimination.

Here, we propose the use of signal detection theory (SDT) (Green and Swets 1966) for measuring lure-old discrimination. We reran the aforementioned BPS paradigm (Reagh and Yassa 2014), which tested effects of repeated encoding on both detail and gist-based memory representations. We demonstrate not only that SDT is an appropriate framework for measuring lure-old discrimination but that SDT formulas yield opposite results to those reported: lure-old discrimination, like novel-old discrimination, gets better with repeated encoding. Additionally, since the original, repeated-encoding study used a unique formula for calculating lure-old discrimination, we discuss potential failings of

other formulas more commonly used for calculating lure-old discrimination in BPS tasks.

Our task followed a very similar structure to the original. The twenty-two healthy volunteers (17 female; age: mean = 20.1, standard deviation = 2.4) provided written consent under a protocol approved by the Institutional Review Board of Johns Hopkins University. Participants incidentally encoded 450 object stimuli (from <http://darwin.bio.uci.edu/~cestark/BPSO/bpso.html> and <http://web.mit.edu/tkonkle/www/stimuli.html>), while performing an "indoor" or "outdoor" judgment task during their 2.5-sec presentation time (each with a 0.5-sec interstimulus interval). Of these encoded stimuli, half were presented only once (1-repeat) while the other half were presented three times, non-consecutively (3-repeat). Repeated stimuli were presented randomly throughout the encoding session so that the lag between presentations differed substantially across items. Participants were then given a surprise, untimed recognition task consisting of 150 old stimuli (half 1-repeat, half 3-repeat), 150 lure stimuli (half similar to untested 1-repeat old stimuli, half similar to untested 3-repeat old stimuli), and 150 novel stimuli. This created five test stimulus types: 1-repeat old, 3-repeat old, 1-repeat lure, 3-repeat lure, and novel. (Because lures were never shown during encoding, the term "repeat," as in 1-repeat lure, refers to the number of times the corresponding old stimulus was shown during encoding.) Consistent with the initial study, participants first provided an "old" or "new" recognition judgment and, unique to our study, subsequently provided a confidence judgment of "very sure," "somewhat sure," or "not at all sure." Participants were instructed to respond "new" to any stimulus that had not been previously seen during encoding—i.e., lure and novel stimuli. Across subjects, mean (and standard deviation of) response frequencies for each stimulus type are reported in Table 1.

First, we sought to replicate results from the original study using their formulas. For novel-old discrimination, we used SDT's d' , defined as the difference between the normalized proportion of old stimuli judged as "old" and the normalized

Corresponding author: rlotit1@jhu.eduArticle is online at <http://www.learnmem.org/cgi/doi/10.1101/lm.038141.115>.

© 2015 Loiotile and Courtney This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first 12 months after the full-issue publication date (see <http://learnmem.cshlp.org/site/misc/terms.xhtml>). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Table 1. Means and standard deviations of response frequencies per stimulus type

Stimulus type (total)	Mean (standard deviation) response frequency					
	Very sure old	Somewhat sure old	Not at all sure old	Not at all sure new	Somewhat sure new	Very sure new
Novel (150)	3.3 (6.9)	3.0 (4.2)	2.3 (3.4)	3.7 (6.3)	35.5 (41.7)	102.1 (45.7)
Old 1-repeat (75)	30.5 (15.1)	13.7 (11.6)	2.6 (4.6)	2.5 (2.9)	9.4 (8.2)	16.2 (14.7)
Old 3-repeat (75)	51.5 (18.8)	11.3 (14.3)	2.6 (7.5)	1.7 (2.1)	3.0 (3.5)	4.8 (10.6)
Lure 1-repeat (75)	9.9 (9.7)	9.3 (9.3)	2.2 (2.9)	2.8 (3.0)	18.1 (13.4)	32.5 (17.6)
Lure 3-repeat (75)	14.5 (13.6)	10.0 (10.6)	2.0 (2.5)	2.4 (2.9)	15.5 (13.9)	30.5 (17.9)

proportion of novel stimuli judged as “old”—i.e., $d' = z(\text{“old”}|\text{old}) - z(\text{“old”}|\text{novel})$. We found that repetition enhanced novel-old discrimination, $t_{(21)} = 18.68$, $P < 0.001$, Cohen’s $d = 3.98$. For lure-old discrimination, we subtracted proportion of old stimuli misses, pooled across repeat conditions, from lure correct rejections—i.e., 1-repeat lure-old discrimination = $p(\text{“new”}|\text{lure})_{1\text{-repeat}} - p(\text{“new”}|\text{old})_{\text{pooled}}$, and 3-repeat lure-old discrimination = $p(\text{“new”}|\text{lure})_{3\text{-repeat}} - p(\text{“new”}|\text{old})_{\text{pooled}}$. Using these formulas, we replicated a novel result of the original experiment: lure-old discrimination appeared to be significantly better for stimuli encoded once, rather than three times, $t_{(21)} = 3.22$, $P < 0.005$, Cohen’s $d = 0.69$. Taken together, these results suggest that repeated exposure to a stimulus enhances the stimulus’s gist representation and, conversely, degrades the stimulus’s detailed representation (Reagh and Yassa 2014).

Next, we examined whether it was appropriate to analyze our data using SDT. Because SDT assumes that each stimulus type’s probability density function (PDF), with respect to memory recognition, forms a normal distribution, it predicts that plots of zROC—i.e., the receiver operating characteristic in z-space—will form straight lines (Macmillan and Creelman 2005). Therefore, for each subject we plotted zROCs for lure-old discrimination—i.e., 1-repeat lure versus 1-repeat old, 3-repeat lure versus 3-repeat old—and for novel-old discrimination—i.e., novel versus 1-repeat old, novel versus 3-repeat old—from which we could individually test linear fit and hence the assumption of normal PDFs. Our analysis indicated that all zROCs were linear, with no systematic deviations, consistent with SDT’s assumption of normal PDFs. Mean (and standard deviation) coefficient of determination (R^2) values are as follows: lure-old 1-repeat = 0.94 (0.08), lure-old 3-repeat = 0.88 (0.12), novel-old 1-repeat = 0.85 (0.18), novel-old 3-repeat = 0.74 (0.24). On average, the assumptions of SDT fit lure-old discrimination just as well, if not better, than novel-old discrimination. This is worth noting as SDT has been the preferred framework for calculating novel-old memory sensitivity, independently of response-criteria, since its seminal introduction by Green and Swets (1966). Therefore, it seems more than reasonable to use SDT for calculating lure-old discrimination. Moreover, SDT has been used to measure lure-old discrimination in non-BPS paradigms (e.g., see Wixted and Stretch 2000).

We subsequently measured both lure-old and novel-old discrimination using SDT. We chose d_a ($= \sqrt{2/(1 + \text{zROC slope}^2)} \times \text{zROC intercept}$; Simpson and Fitter 1973) as an appropriate metric for discriminative sensitivity over the more commonly used d' . d' assumes equal variances of the normal PDFs, an assumption that empirical data often fails to validate (Macmillan and Creelman 2005; see Dougal and Rotello 2007 for an example of d' leading to a different, erroneous, conclusion from d_a). Therefore, d_a , which accounts for both the distance of the PDFs as well as their relative variances, is a more accurate measure of relative overlap between two stimulus type PDFs and hence of a participant’s ability to discriminate between those two stimulus types. (Note that $d' = d_a$ when PDF variances are equal.) Using slopes and intercepts of the best-fit lines from our zROC graphs for

each participant, we were able to calculate d_a as a measure of novel-old and lure-old discrimination for both 1-repeat and 3-repeat conditions. For example, 1-repeat lure-old d_a for each participant was calculated using the participant’s ordinary least squares line of best fit to 1-repeat lure-old zROC plots. Our analysis indicated significantly better discrimination of novel stimuli from old 3-repeat, rather than old 1-repeat stimuli, $t_{(21)} = 13.09$, $P < 0.001$, Cohen’s $d = 2.79$. Corresponding lure stimuli were also better discriminated from old 3-repeat than old 1-repeat stimuli, $t_{(21)} = 9.72$, $P < 0.001$, Cohen’s $d = 2.07$. The latter result is particularly remarkable as it differs in direction from that found in the original study, and reproduced by us, using the original paper’s formula. Therefore, encoding repetition was found to either enhance or degrade lure-old discrimination dependent on whether SDT or the original paper’s formulas, respectively, were used (Fig. 1).

As an “adjudicator” between the conflicting formulas, we gave participants a two-alternative forced choice (2AFC) recognition task immediately following the aforementioned old/new recognition task, for the stimuli that had not been tested in the old/new task. In each of the 150 trials (half 1-repeat, half 3-repeat), we displayed two stimuli: an old stimulus (not tested in the old/new recognition task) and a corresponding lure stimulus. We asked participants to pick the stimulus they had seen at encoding and to then indicate their confidence in that choice (i.e., “very sure,” “somewhat sure,” or “not sure”). We thought that the 2AFC paradigm was particularly appropriate for assessing lure-old discrimination since participants were quite literally tasked with discriminating each old stimulus from its corresponding lure stimulus. Moreover, because a 2AFC enables participants to make a relative recognition decision (i.e., which one of the two has more memory strength?) rather than an absolute recognition decision (i.e., does the recognition strength for this one stimulus exceed my criterion for calling it “old?”), 2AFC responses do not require SDT formulas for extracting memory sensitivity from criterion-confounded responses of “old.” Rather, the commonly accepted formula for measuring memory sensitivity in a 2AFC task is a simple proportion correct (= no. of old stimuli correctly chosen/total no. of trials).⁴ Consistent with our SDT analyses,

⁴If one wishes to remove any potential participant bias for choosing left-stimuli over right-stimuli, one can use SDT’s d_a in lieu of proportion correct (see Macmillan and Creelman 2005 for how to do so). In humans, there is rarely a strong left–right bias. Moreover, there is a reason not to convert a 2AFC proportion correct measure to an old/new recognition measure such as d_a : under SDT, the d_a formula assumes that the lure-old pairs are randomly distributed across both lure and old memory-strength distributions. In a paradigm where a participant must choose “which is old?” between an old image and a randomly picked lure, this assumption is likely met. However, in the BPS paradigm, a participant must choose between an old image and that old image’s matching lure, not any random lure. Therefore, the assumption of random distribution cannot be guaranteed. If lure-old pairs are randomly distributed with respect to memory strength and there is no left–right bias, then 2AFC proportion correct = area under the old/new lure-old ROC = $\Phi(d_a/\sqrt{2})$, where Φ represents the normal cumulative distribution function. See Migo et al. (2009) for an excellent discussion of this issue, Bayley et al. (2008) for evidence that the random-distribution assumption is met.

Figure 1. Repeated encodings' effect on memory discrimination. Graphs *A* and *B* analyze the same data in two different ways. (*A*) Gist-based novel-old discrimination, as measured by d' , is significantly enhanced by repeated exposure to old stimuli. Conversely, detailed lure-old discrimination, as measured by a formula specific to Reagh and Yassa (2014), is significantly impaired by repeated exposure to old stimuli. (*B*) Using SDT's d_a , both novel-old and lure-old discrimination demonstrate significant benefit from repeated exposure to old stimuli. (*C*) Results from an independent 2AFC test on the same participants. Enhanced lure-old discrimination in the 3-repeat condition corroborates our findings from SDT analysis of the old/new recognition test. All error bars represent standard error of the mean.

repetition at encoding produced significantly better lure-old discrimination, $t_{(21)} = 9.81$, $P < 0.001$, Cohen's $d = 2.09$. Thus, the 2AFC task provides an independent and compelling confirmation

of the SDT results demonstrating enhanced gist- and detail-based memory for repeatedly encoded items.

Then, what contributed to the original formula's "flipped" effect of enhanced detail-based memory for items that were only encoded once? For each condition, the original formula measures lure-old discrimination as the difference between memory strength for that condition's lure stimuli and *both* conditions' old stimuli. The face validity of this measurement seems questionable. For example, for 1-repeat lure-old discrimination, the formula equals the difference between 1-repeat lure memory strength and an average of 1- and 3-repeat old memory strength, rather than the arguably more informative difference between 1-repeat lure and 1-repeat old memory strength. Because this pooled response to old stimuli is identical for both 1-repeat and 3-repeat lure-old discrimination measures, the original formula reduces to measuring lure "correct rejections" = $p(\text{"new"}|\text{lure})$ for each stimulus type. On average, the memory strength distribution for the 3-repeat lures has a larger variance than that for 1-repeat lures (1-repeat $\sigma = 1.29$, 3-repeat $\sigma = 1.64$), although the distributions have roughly equivalent means (1-repeat $\mu = 0.88$, 3-repeat $\mu = 0.86$). That is, the 3-repeat condition has a higher proportion of lures with *slightly* greater (false) memory recognition than the 1-repeat condition (see Fig. 2). Therefore, any measure of lure-old discrimination that primarily characterizes the memory strength of lure stimuli—rather than the *difference* in memory strength between lure and old stimuli—would produce a "flipped" result for this particular data set and, possibly, other BPS paradigms that use such a formula.

It is important to underscore that one might reach different conclusions if one uses a formula that interrogates memory for lure stimuli, rather than memory for the lure stimuli *relative* to the old stimuli. In the present study, 1-repeat lures are more likely than 3-repeat lures to be correctly identified as "new." However, this increase in lure false alarming for the 3-repeat condition does not imply deteriorated memory for items that were repeatedly encoded. Indeed, both SDT analysis of the old/new recognition task and results from the independent 2AFC task demonstrate better discrimination of 3-repeat, rather than 1-repeat, old stimuli from novel and corresponding lure stimuli. Examining the SDT PDFs reveals the reason behind the seemingly counterintuitive dissociation between lure false alarm rate and discrimination ability: while repetition at encoding engenders slightly greater false recognition of analogous lure items, the much-increased recognition strength of the repeatedly encoded old items results in a net decrease in overlap between the 3-repeat lure and 3-repeat old PDFs. It is this decrease in lure-old overlap for the 3-repeat condition that engenders enhanced, rather than degraded, discrimination. Ultimately the formulas must reflect what one hopes to measure. For BPS studies, the question of interest is not typically "recognition strength" for lures. Rather, one wishes to assess the precision of memory representation for old stimuli by comparing that representation to that of lure stimuli.

Exactly why repeated encoding increases false alarming to lures is an interesting theoretical issue. Memory strength for lure items is based on encoding success of the corresponding old items. Therefore, a participant would likely respond "new" to a lure stimulus for two very different reasons: (1) if the corresponding old stimulus was encoded with very low fidelity (i.e., if the old item had not been encoded at all) and (2) if the corresponding old stimulus was encoded with very high fidelity (i.e., if recalled details of the old item allow the participant to reject the lure item as "old," commonly known as "recall-to-reject" (Kirwan and Stark 2007)). As such, the increased variance of the 3-repeat lure PDF relative to that of the 1-repeat lure PDF makes perfect sense: repeated encoding allows for decreased encoding-failures of old stimuli (in turn, begetting stronger memory for some lure stimuli)

Figure 2. A SDT modeling of recognition memory: probability density functions with respect to memory recognition strength for each of the five test stimulus types, averaged across participants. (Novel PDFs are standardized at $\mu = 0$, $\sigma = 1$. Parameters for each of the other four PDFs are determined by the line best fit to the novel-stimulus zROC; $\mu = \text{intercept/slope}$, $\sigma = 1/\text{slope}$.) The vertical line in both graphs represents participants' new/old decision criterion, such that all stimuli exceeding the criterion's recognition strength were deemed "old" at test. Comparing the two graphs, one can see that 1-repeat lures are more often labeled "new" than 3-repeat lures—i.e., the portion of the 1-repeat lure PDF falling to the left of the decision criterion is larger in area than that of the 3-repeat's. However, it is also evident that the 3-repeat lure and old distributions are substantially less overlapping than the 1-repeat lure and old distributions. This degree of separation is reflected in SDT's memory discrimination measure d_a . Therefore, repeated encoding of old items enhances their subsequent discrimination from highly similar items.

and for increasingly detailed encoding of the old stimuli (in turn, begetting weaker memory for other lure stimuli). The fact that SDT cannot distinguish between these two alternative reasons for lure memory strength does not preclude its usage. Rather, in allowing one to measure the "difference" in memory strength between lure and old stimuli (themselves an undoubtable mixture of different types of representations), SDT allows one to make conclusions on the precision of the underlying old stimulus representations. For example, if "new" responses to lures primarily reflect encoding failure, this encoding failure would also be evidenced in many "new" responses to the old stimuli and would, therefore, produce a small lure-old d_a with both lure and old PDFs highly overlapped with the novel PDF. Therefore, SDT can help to characterize these nuanced aspects of memory strength.

Because the "pooled" formula is unique to the original study, it behooves us to examine formulas more commonly used to determine lure-old discrimination in BPS paradigms. The majority

of BPS studies (e.g., Toner et al. 2009; Yassa et al. 2010; Stark et al. 2013; Borota et al. 2014) ask participants to characterize tested stimuli as "old," "similar," or "new" rather than just "old" or "new." Often lure-old discrimination is measured as $p(\text{"similar"}|\text{lure})$ or $p(\text{"similar"}|\text{lure}) - p(\text{"old"}|\text{lure})$ (e.g., Yassa, Mattfeld, et al. 2011). These tasks assume that a participant only ever responds "similar" when she/he is successfully performing recall-to-reject. Yet this cannot possibly be true, as "similar" is also given in response to both old and novel stimuli. It is likely, therefore, that a handful of "similar" responses are given as a kind of low-confidence "old" judgment. On the surface, our inability to characterize the "similar" response given to lures in an old/similar/new recognition task might seem analogous to the varied reasons why a participant might respond "new" to lures in an old/new recognition task. However, there are noteworthy differences that render the former problematic and the latter manageable. As discussed above, exactly why a participant responds "new" to lures becomes somewhat irrelevant as SDT allows one to compare lure with old stimuli. Critically, in the old/similar/new recognition task, no single formula captures both pieces of information. (Sometimes these analyses include independent estimates of responses to old items, but a direct comparison to lure items is lacking.) Therefore, as with the pooled and "correct rejection" formulas, "similar" formulas only use participants' responses to lure items as a measure of lure-old discrimination. To address this concern, one might be tempted to try a SDT analysis on an old/similar/new recognition task. However, this would require the researcher to decide a priori whether to treat "similar" as indicative of recall-to-reject (and hence group the responses with "new") or to treat "similar" as a low-confidence "old" judgment. Thus, while using the old/similar/new response set may be intuitively appealing, it is highly problematic.

Of these BPS lure-old discrimination formulas, one of the closest to our preferred measure of d_a is a type of response-bias corrected proportion correct $P_r = p(\text{"old"}|\text{old}) - p(\text{"old"}|\text{lure})$ (e.g., in Leal et al. 2014). Like SDT's d_a and d' , and unlike the aforementioned BPS formulas, P_r accounts for the relative overlap between lure and old memory strengths. This is a strength of the formula. Unlike SDT's d_a and d' , P_r assumes that lure and old PDFs have boxcar, rather than normal, distributions with respect to recognition strength. Moreover, like d' , P_r is a one-point measurement which assumes that lure and old distributions have equal-variances (though d' and P_r differ on the assumed shape of those distributions). As seen in Figure 3, the data better reflect the assumptions of d_a than P_r : d_a assumes curvilinear ROCs and linear zROCs (of slope = 1 and y -intercept = d'), while P_r assumes a linear ROC (of slope = 1 and y -intercept = P_r) and a concave-up curvilinear zROC (Macmillan and Creelman 2005). Across subjects, the ROC models implied by SDT's d_a fit the data significantly better than those implied by P_r ; root-mean-square error is significantly lower for ROC values predicted by d_a in the 1-repeat, $t_{(21)} = 5.99$, $P < 0.001$, Cohen's $d = 1.31$, and 3-repeat conditions, $t_{(21)} = 3.41$, $P < 0.005$, Cohen's $d = 0.74$. This worse fit of ROCs implied by P_r might be somewhat trivially expected, as P_r utilizes only one of the five cumulative proportions to characterize memory discrimination. Therefore, it may be more appropriate to compare P_r with SDT's d' . d' also provides a significantly better fit than P_r for both conditions: 1-repeat, $t_{(21)} = 2.65$, $P < 0.05$, Cohen's $d = 0.57$, and 3-repeat, $t_{(21)} = 3.17$, $P < 0.005$, Cohen's $d = 0.68$. It should be noted that using the P_r formula does not yield different results from our SDT analyses in this particular case: 3-repeat lure-old discrimination is significantly better than 1-repeat, $t_{(21)} = 6.76$, $P < 0.001$, Cohen's $d = 1.44$. However, this does not mean that P_r will *always* yield the correct result. Ultimately, the implied (z)ROC models reflect each formula's measure of memory discrimination across varying decision criteria. As such, a poor fit

Figure 3. Lure-Old ROCs and zROCs for 1-repeat (A,B) and 3-repeat (C,D) conditions. ROC (zROC) data points are cumulative proportions (z-scored cumulative proportions) of each response type—e.g., “very sure old”—for lure and old stimuli, across all subjects. Dark curves/lines represent the fitted SDT model, allowing variances between lure and old PDFs to be unequal; light lines/curves represent the P_r ’s implied model of memory discrimination. Better discrimination between lure and old stimuli in the 3-repeat condition is evidenced by the increased distance of the 3-repeat points (and models) from the line $y = x$. Comparing the fit from models assumed by the d_a and P_r formulas, it is evident that the former fits the data better for both the 1-repeat and 3-repeat conditions and, therefore, more accurately characterizes the magnitude of memory discrimination. More generally, the data appear consistent with SDT predictions, and inconsistent with high-threshold model predictions: the lure-old ROC data points appear curvilinear, rather than linear, and the zROC data points appear linear, rather than curvilinear and concave-up.

can lead to conclusions of differing memory discrimination between two conditions for which memory discrimination is the same or, conversely, to equivalent memory discrimination between two conditions for which there is a difference in memory discrimination. Therefore, although $p(\text{“old”}|\text{old}) - p(\text{“old”}|\text{lure})$ is one of the better formulas used in BPS paradigms to measure lure-old discrimination, it is also not without its flaws and can lead to erroneous conclusions.

It is worth mentioning that P_r is a specific instantiation of a high-threshold model for recognition memory. Dual-process theories (Yonelinas 1994) state that recognition memory is a combination of signal-detection and high-threshold processes (i.e., familiarity and recollection, respectively). Therefore, they predict that, because novel-old discrimination utilizes both processes, novel-old ROCs will be less curvilinear and more linear in parts of the ROC-plot corresponding to high confidence old judgments. Additionally, because lure-old discrimination is thought to utilize only recollection, dual-process accounts predict a completely linear lure-old ROC. This dual-process account is currently well-debated. We, as others (e.g., Wixted and Stretch 2004), do not

find evidence of linearity in ROCs (and curvilinearity in zROCs) and, as such, favor a signal-detection account for both novel-old and lure-old discrimination. If one wishes to remain neutral to this debate or agnostic to the assumption of normally distributed recognition strength for Novel, Lure, and Old items, one can simply take the area under the ROC (AUC), in lieu of d_a , for both novel-old and lure-old discrimination. Unlike d_a , AUC is calculated without fitting a normally distributed model to the data (Pollack and Hsieh 1969). When the assumption of normal distribution is met, $d_a = \sqrt{2}\Phi^{-1}(\text{AUC})$, where Φ^{-1} represents the inverse normal cumulative distribution function. (Note that in practice, even when the data are normally distributed, methods for calculating AUC frequently underestimate the true area.) Using AUC calculated via the trapezoidal rule, our d_a results were confirmed: novel-old discrimination, $t_{(21)} = 11.11$, $P < 0.001$, Cohen’s $d = 2.37$, and lure-old discrimination, $t_{(21)} = 8.60$, $P < 0.001$, Cohen’s $d = 1.83$, were both enhanced with repeated exposure.

In considering how lure-old discrimination is measured in BPS paradigms, it is imperative that we also reflect on the original motivation behind the paradigm and the measurement’s implied construct. As the task name suggests, behavioral lure-old discrimination is thought to measure the success of the hippocampal process of “pattern separation”—i.e., the orthogonalization of memory representations for similar stimuli (Yassa and Stark 2011). Many in the field have acknowledged that it is difficult to make a direct link between a behavioral outcome in a complex task and a specific neural computation in one brain structure (e.g., Yassa and Stark 2011; Santoro 2013; Stark et al. 2013; Reagh and Yassa 2014).

There has been some evidence, however, that aspects of performance in BPS paradigms do correlate with activation in the hippocampus, and that this varies in aging and disease (e.g., Toner et al. 2009; Yassa et al. 2010, 2011; Holden et al. 2012). In order to more precisely link behavioral performance to neural computations, a better understanding is first needed of the underlying behavioral representations. Pattern separation is measured by the amount of neural representational overlap between lure and old stimuli (Vazdarjanova and Guzowski 2004; Leutgeb et al. 2007). Therefore, at a minimum, the behavioral data should also be measured by a formula that characterizes the amount of recognition overlap between lure and old stimuli. This is the essence of SDT’s d_a (or even proportion correct in a 2AFC recognition task). Other formulas have either considered only the lure representation, and not the relationship between lure and old representations, or have poorly characterized this relationship.

Though d_a does reliably measure the degree of lure-old overlap, more conditions must be met for one to conclude that less behaviorally measured overlap in these representations is

specifically due to the hippocampal process of pattern separation during processing of lure stimuli. For example, our data demonstrate increased discrimination, and hence less lure-old overlap, in the 3-repeat than 1-repeat condition. But, can we conclude that this increased separation in behavioral recognition strength is specifically driven by increased hippocampal pattern separation? In beginning to answer this question, it is first important to note that the representation of old stimuli—as measured by novel-old discrimination—is different between the 1-repeat and 3-repeat conditions. Therefore, it seems that increased discrimination between lure and old stimuli with repeated encodings is driven not by increased pattern separation of the lures but by differences in the representation of the old stimuli. More generally, one might argue that strong evidence for greater behavioral pattern separation in condition A than condition B should satisfy two conditions: lure-old overlap should be smaller in condition A, while novel-old overlap stays the same for both conditions. Simply put: increased behavioral pattern separation should shift only the lure distribution. As mentioned above regarding dual-process theories of memory recognition, a broader consideration when using these paradigms, however, is that there may be multiple processes contributing to the behavior (Molitor et al. 2014). Therefore, behavioral paradigms alone cannot usually suffice for conclusions about neural processes (Santoro 2013), but a better behavioral measure can help to test whether there is indeed a relationship.

In conclusion, the BPS paradigm uniquely allows one to examine both gist- and detail-based recognition. However, the many current measures of lure-old discrimination in BPS paradigms fail to accurately characterize the fidelity of memory representation for the old stimuli. These formulas have either neglected to compare memory strength between old and lure stimuli, focusing only on the latter, or have imprecisely modeled the overlap in old and lure representations. Moreover, when BPS studies use different formulas for measuring novel-old and lure-old discrimination, it is impossible to directly compare the two types of discrimination. Therefore, SDT for old/new recognition tasks—or proportion correct for nAFC tasks (see Brady et al. 2011 for a good example)—provides an appropriate and unified framework for measuring both novel-old and lure-old discrimination, allowing stronger tests of hypotheses that predict differences between gist- and detail-based recognition. Our analyses demonstrate that a formula used in a particular BPS paradigm (Reagh and Yassa 2014) lead to an erroneous conclusion. It is important to realize that this one result, as well as the shortcomings of other BPS formulas, do not ipso facto invalidate all conclusions thus far derived from BPS paradigms. Future BPS paradigms using SDT or other theoretically justified analysis methods will have to bear this out. More generally, researchers should remember that conclusions are contingent upon dependent measures, which are themselves contingent upon theory-laden measurement assumptions that can, and should be tested whenever possible.

Acknowledgments

We thank Dr John Wixted for invaluable advice on Signal Detection Theory; Antonio Vergara for task programming and participant testing; Corbin Cunningham, Dr Howard Egeth, Jorge Menendez, and Dr Daniel Naiman for helpful conversations; Dr Amy Shelton for first introducing me (R.E.L.) to SDT; and our anonymous reviewers for perceptive questions and suggestions. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1232825. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

- Bayley PJ, Wixted JT, Hopkins RO, Squire LR. 2008. Yes/no recognition, forced-choice recognition, and the human hippocampus. *J Cognit Neurosci* **20**(3): 505–512.
- Borota D, Murray E, Keceli G, Chang A, Watabe JM, Ly M, Toscano JP, Yassa MA. 2014. Post-study caffeine administration enhances memory consolidation in humans. *Nat Neurosci* **17**: 201–203.
- Brady TF, Konkle T, Alvarez GA. 2011. A review of visual memory capacity: beyond individual items and toward structured representations. *J Vis* **11**: 4.
- Dougal S, Rotello CM. 2007. “Remembering” emotional words is based on response bias, not recollection. *Psychon Bull Rev* **14**: 423–429.
- Ebbinghaus H. 1885. *Memory: a contribution to experimental psychology*. Von Duckner and Humboldt, Leipzig, Germany.
- Green DM, Swets JA. 1966. *Signal detection theory and psychophysics*. Peninsula Publishing, Los Altos, CA.
- Holden HH, Hoebel C, Loftis K, Gilbert P. 2012. Spatial pattern separation in cognitively normal young and older adults. *Hippocampus* **22**: 1826–1832.
- Kirwan CB, Stark CEL. 2007. Overcoming interference: an fMRI investigation of pattern separation in the medial temporal lobe. *Learn Mem* **14**: 625–633.
- Leal SL, Tighe SK, Yassa MA. 2014. Asymmetric effects of emotion on mnemonic interference. *Neurobiol Learn Mem* **111**: 41–48.
- Leutgeb JK, Leutgeb S, Moser MB, Moser EI. 2007. Pattern separation in the dentate gyrus and CA3 of the hippocampus. *Science* **315**: 961–966.
- Macmillan NA, Creelman CD. 2005. *Detection theory: a users guide*, 2nd ed. Cambridge University Press, New York.
- Migo E, Montaldi D, Norman KA, Quamme J, Mayes A. 2009. The contribution of familiarity to recognition memory is a function of test format when using similar foils. *Q J Exp Psychol (Hove)* **62**: 1198–1215.
- Molitor RJ, Ko PC, Hussey EP, Ally BA. 2014. Memory-related eye movements challenge behavioral measures of pattern completion and pattern separation. *Hippocampus* **24**: 666–672.
- Pollack I, Hsieh R. 1969. Sampling variability of the area under the ROC-curve and of d'. *Psychol Bull* **71**: 161–173.
- Reagh ZM, Yassa MA. 2014. Repetition strengthens target recognition but impairs similar lure discrimination: evidence for trace competition. *Learn Mem* **21**: 342–346.
- Santoro A. 2013. Reassessing pattern separation in the dentate gyrus. *Front Behav Neurosci* **7**: 96.
- Simpson AJ, Fitter MJ. 1973. What is the best index of detectability? *Psychol Bull* **80**: 481–488.
- Stark SM, Yassa MA, Lacy JW, Stark CEL. 2013. A task to assess behavioral pattern separation (BPS) in humans: data from healthy aging and mild cognitive impairment. *Neuropsychologia* **51**: 2442–2449.
- Toner CK, Pirogovsky E, Kirwan B, Gilbert P. 2009. Visual object pattern separation deficits in nondemented older adults. *Learn Mem* **16**: 338–342.
- Vazdarjanova A, Guzowski JF. 2004. Differences in hippocampal neuronal population responses to modifications of an environmental context: evidence for distinct, yet complementary, functions of CA3 and CA1 ensembles. *J Neurosci* **24**: 6489–6496.
- Wixted JT, Stretch V. 2000. The case against a criterion-shift account of false memory. *Psychol Rev* **107**: 368–376.
- Wixted JT, Stretch V. 2004. In defense of the signal detection interpretation of remember/know judgments. *Psychon Bull Rev* **11**: 616–641.
- Yassa MA, Stark CEL. 2011. Pattern separation in the hippocampus. *Trends Neurosci* **34**: 515–525.
- Yassa MA, Lacy JW, Stark SM, Albert MS, Gallagher M, Stark CEL. 2010. Pattern separation deficits associated with increased hippocampal CA3 and dentate gyrus activity in nondemented older adults. *Hippocampus* **21**: 968–979.
- Yassa MA, Mattfeld AT, Stark SM, Stark CE. 2011. Age-related memory deficits linked to circuit-specific disruptions in the hippocampus. *Proc Natl Acad Sci* **108**: 8873–8878.
- Yonelinas AP. 1994. Receiver operating characteristics in recognition memory: evidence for a dual process model. *J Exp Psychol Learn Mem Cogn* **20**: 1341–1354.

Received January 16, 2015; accepted in revised form June 5, 2015.

A signal detection theory analysis of behavioral pattern separation paradigms

Rita E. Loiotile and Susan M. Courtney

Learn. Mem. 2015, **22**:

Access the most recent version at doi:[10.1101/lm.038141.115](https://doi.org/10.1101/lm.038141.115)

References

This article cites 23 articles, 7 of which can be accessed free at:
<http://learnmem.cshlp.org/content/22/8/364.full.html#ref-list-1>

Creative Commons License

This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first 12 months after the full-issue publication date (see <http://learnmem.cshlp.org/site/misc/terms.xhtml>). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).