Journal of Experimental Psychology: Learning, Memory, and Cognition

The Response Dynamics of Recognition Memory: Sensitivity and Bias

Gregory J. Koop and Amy H. Criss

Online First Publication, November 23, 2015. http://dx.doi.org/10.1037/xlm0000202

CITATION

Koop, G. J., & Criss, A. H. (2015, November 23). The Response Dynamics of Recognition Memory: Sensitivity and Bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. http://dx.doi.org/10.1037/xlm0000202

Journal of Experimental Psychology:
Learning, Memory, and Cognition

© 2015 American Psychological Association
0278-7393/15/$12.00 http://dx.doi.org/10.1037/xlm0000202

The Response Dynamics of Recognition Memory: Sensitivity and Bias

Gregory J. Koop
Syracuse University and Eastern Mennonite University

Amy H. Criss
Syracuse University

Advances in theories of memory are hampered by insufficient metrics for measuring memory. The goal of this paper is to further the development of model-independent, sensitive empirical measures of the recognition decision process. We evaluate whether metrics from continuous mouse tracking, or response dynamics, uniquely identify response bias and mnemonic evidence, and demonstrate 1 application of these metrics to the strength-based mirror-effect paradigm. In 4 studies, we show that response dynamics can augment our current analytic repertoire in a way that speaks to the psychological mechanisms underlying recognition memory. We manipulated familiarity and response bias via encoding strength and the proportion of targets at test (Experiment 1) and found that the initial degree of deviation of the mouse movement toward a response is a robust indicator of response bias. In order to better isolate measures of memory strength, we next minimized response bias through the use of 2-alternative forced-choice tests (Experiments 2 and 3). Changes in the direction of movement along the *x*-axis provided an indication of encoding strength. We conclude by applying these metrics to the typical strength-based mirror effect design (Experiment 4) in an attempt to further discriminate between differentiation and criterion-shift accounts.

Keywords: strength-based mirror effect, recognition memory, differentiation, criterion-shifts, mouse-tracking

An umpire in Major League Baseball faces the difficult task of identifying balls and strikes under conditions that offer limited evidence. The location of the pitch in the strike zone can be thought of in terms of quality of evidence. As the location of the pitch nears the boundary of the strike zone the quality of evidence decreases, and the decision becomes more difficult. In order to make a call, the umpire must decide the amount of evidence that is sufficient to call a strike. Some umpires might be predisposed to call borderline pitches strikes (colloquially known as a large strike zone), whereas others will be inclined in the opposite direction (a narrow strike zone). Figure 1 represents a signal detection theory (SDT) framework for this decision. The *x*-axis represents the evidence on which the ball/strike decision is based, and the distributions along this axis represent the frequency with which balls (left distribution) and strikes (right distribution) provide a given amount of evidence. True strikes generally provide more evidence than balls, but due to noise in the perceptual system, this relationship may not always hold as indicated by the overlap in the distributions. The vertical line in Figure 1 represents this *decision criterion*, which provides an indication of the umpire's decision bias. An umpire with a small strike zone will have a conservative criterion (a rightward shift in Figure 1), whereas the opposite would be true of an umpire with a large strike zone (a leftward shift in Figure 1). In this way SDT distinguishes between overall skill (or sensitivity or discriminability) and response bias, and has been productively applied to such fields as recognition memory (e.g., MacMillan & Creelman, 1991; Parks, 1966), psychophysics (e.g., Green & Swets, 1966), medical decision making (e.g., King et al., 1997), and even police officers' decisions about whether or not to fire their weapons (Plant & Peruche, 2005; Correll et al., 2007).

It is difficult to overstate the utility and impact of SDT as a measurement model of aggregate decision tendencies. However, because sensitivity and bias are summary measures calculated on the basis of an individual's aggregate behavior, SDT provides static metrics that do not speak to the dynamic processes underlying individual choices. This is problematic when the goal of research is to discriminate between models whose predictions differ only in the process mechanisms underlying behavior. Examples of this type abound in the field of recognition memory.

In the typical recognition memory experiment, participants study a series of items (words, pictures, etc.) and then complete a test over that material in which they must indicate whether each test probe is a previously studied target or an unstudied foil. Applying the standard SDT representation to this process is straightforward: items vary on the amount mnemonic evidence they provide (also referred to as familiarity, subjective strength, or global match), and the mean of the target distribution is greater than the foil distribution (e.g., replace "strikes" with "targets," and "balls" with "foils" in Figure 1). The decision criterion serves the same role as in the opening example. When mnemonic evidence exceeds the criterion an item is called "old," but when evidence

Gregory J. Koop, Department of Psychology, Syracuse University and Department of Psychology, Eastern Mennonite University; Amy H. Criss, Department of Psychology, Syracuse University.

This material is partially based on work supported by National Science Foundation Grant Number 0951612.

Correspondence concerning this article should be addressed to Gregory J. Koop, Department of Psychology, Eastern Mennonite University, 1200 Park Road, Harrisonburg, VA 22802. E-mail: gregory.koop@emu.edu

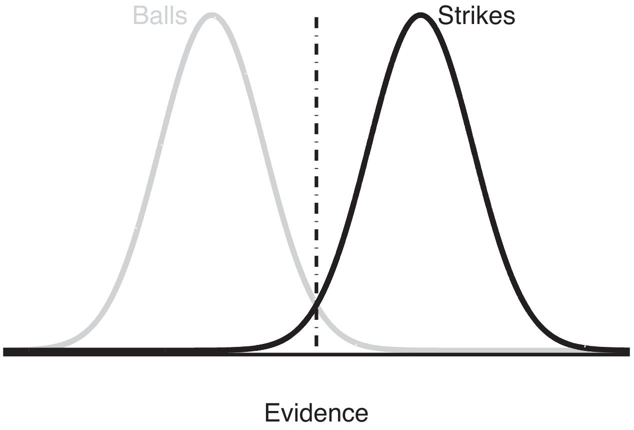

Figure 1. One possible SDT representation of the judgment task faced by baseball umpires. Strikes are typically of higher quality, and therefore have more “evidence” than balls, which is represented on the *x*-axis. The frequency of balls and strikes at a given level of evidence is represented on the *y*-axis. The dashed line represents the decision criterion.

fails to surpass the criterion items are called “new.” Sensitivity thus represents the ease with which foils can be distinguished from targets, and the position of the criterion relative to the distributions represents bias toward an “old” or “new” response.

Although the measures *d*′ and *C* are often used to quantify discriminability and response bias, respectively (e.g., Stanislaw & Todorov, 1999), they rely on the specific assumptions of the SDT measurement model. Problematically, several behavioral findings in the recognition memory literature can be explained equally well by a SDT model that assumes a criterion shift or a SDT model that assumes a shift in the underlying distributions of memory strength (e.g., the strength-based mirror effect to which we turn in Experiment 4). This is often due to the insensitivity of SDT measures to some manipulations of bias (e.g., Banks, 1970; Grider & Malmberg, 2008; Healy & Kubovy, 1978; Kantner & Lindsay, 2010; Rhodes & Jacoby, 2007) and the saturation of SDT in a typical memory paradigm. In addition, SDT measures provide no information about how the decision unfolds. The primary goal of this work is therefore to apply continuous mouse tracking, or response dynamics, to recognition tasks and describe quantitative measurements that map onto response bias and mnemonic evidence.

Response Dynamics

Response dynamics is a method that has advanced theory in phonological processing (Spivey, Dale, Knoblich, & Grosjean, 2010; Spivey, Grosjean, & Knoblich, 2005), face and race perception (Freeman & Ambady, 2011; Freeman, Pauker, Apfelbaum, & Ambady, 2010), categorization (Dale, Kehoe, & Spivey, 2007; Flumini, Barca, Borghi, & Pezzulo, 2014), perceptual decision making (e.g., Lepora & Pezzulo, 2015), and risky decision making (Koop & Johnson, 2013), among others. Critically, unlike typical recognition studies where a response is entered via a single button press, response dynamics provides a more continuous depiction of preference development toward response alternatives. Participants are asked to move a mouse cursor from the bottom center of the screen to either of two spatially disparate choice options, which are located in the upper corners of the screen (e.g., Figure 2a).

Although response dynamics has begun to be widely utilized within cognitive science, to our knowledge there has only been a single application of the method to recognition memory. Papesh and Goldinger (2012) collected continuous response data during a typical single-item recognition task. Following each trial, participants provided confidence ratings on the accuracy of the prior judgment. The data showed that high confidence ratings tended to be preceded by more direct responses than did trials receiving low confidence ratings. Although this application shows that a relationship exists between response dynamics and explicit confidence ratings, the analysis did not extend to how the trajectories of the behavioral response were related to response bias and the quality of mnemonic evidence. Here, we pair recognition memory paradigms with measures of response dynamics to bring new dependent variables to the SDT framework with the ultimate goal of better discriminating between theories of memory.

Experiment 1

We adopted classic experimental manipulations that are known to affect response bias and mnemonic evidence in recognition memory to examine whether measures of response dynamics sys-

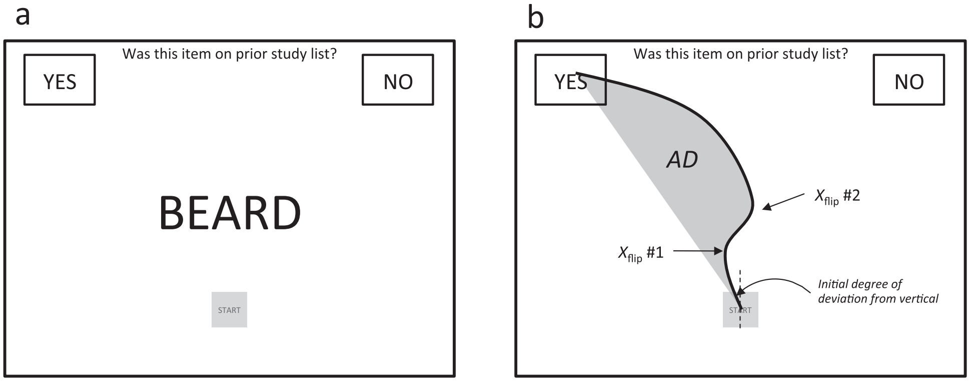
a
Was this item on prior study list?
YES
NO
BEARD
START
b
Was this item on prior study list?
YES
NO
AD
X_{flip} #2
X_{flip} #1
Initial degree of deviation from vertical
START

Figure 2. Panel A provides an example test screen for single item recognition (Experiments 1 and 4). Note that stimulus items did not appear until after participants clicked the *start* box. Panel B provides an example trajectory with associated metrics.

tematically map onto these SDT constructs. We manipulated response bias by varying the proportion of test items that are targets (e.g., Bruhn, Huette, & Spivey, 2014; Criss, 2009; Healy & Kubovy, 1978; Ratcliff, Sheu, & Gronlund, 1992; Rotello, MacMillan, Hicks, & Hautus, 2006). When participants are aware of such base rate manipulations, hits and false alarms increase with increases in the proportion of targets, which indicates greater bias toward the "old" response. We manipulated mnemonic evidence via a depth of encoding manipulation (e.g., Craik & Lockhart, 1972; Duchek & Neely, 1989; Kiliç, 2012). In typical experiments discrimination is higher for items encoded under more challenging, semantic-based tasks compared with shallow, surface-based tasks (cf. Nelson, 1977). This increase in discrimination takes the form of a mirrored-pattern where hits are higher and false alarms are lower for the deeper encoding task, which is referred to as a strength-based mirror effect (SBME).

In terms of behavioral data, we expected to replicate standard findings of an increase in the probability of saying "old" for both targets and foils as the proportion of targets increased. Further, we expected to replicate a SBME for the depth of processing manipulation. In terms of response dynamics, predictions were speculative. Papesh and Goldinger (2012) found more direct trajectories when the decision was followed by a high confidence endorsement (relative to a low confidence judgment of the decision). However, such confidence ratings reflect mnemonic evidence, the criterion, and metacognitive evidence (e.g., Green & Swets, 1966); thus that study does not allow a clear mapping between SDT constructs and response dynamics. The question of interest, then, is what measures of response dynamics are sensitive to changes in bias and changes in memory evidence. A common view is that bias is present prior to evaluating evidence provided by the stimulus and thus should be obvious (nearly) immediately whereas mnemonic evidence accrues over time and is more dominant later in the decision process (e.g., Mulder, Wagenmakers, Ratcliff, Boekel, & Forstmann, 2012; White & Poldrack, 2013). To the extent that the early and late components of the trajectories could be isolated, we expected to see the influence of proportion targets on response bias early and the influence of depth of encoding on memory strength later.

Method

Participants. A total of 161 participants from the Syracuse University research pool took part in Experiment 1. Of the 161 participants in Experiment 1, 87 completed a "weak" encoding task and 74 completed a "strong" encoding task. All participants received partial fulfillment of course requirements in exchange for their participation.

Stimulus materials. Word stimuli were drawn from a pool of 424 medium normative frequency words between 3 and 13 letters in length (median = 6) and ranging between 5.19 and 13.22 log frequency ($M = 8.87$, $SD = 1.12$) in the Hyperspace Analog to Language Corpus (Balota et al., 2007). For each participant, words were randomly selected from this pool and randomly assigned to condition.

Design and procedure. Experiment 1 was a 2 (encoding task: strong or weak) × 5 (percent of test items that are targets: 10, 30, 50, 70, 90) × 2 (test trial type: target or foil) mixed design where encoding task was between-subjects. Upon providing informed consent, participants were informed that their goal for the experiment was to study a series of words and successfully discriminate between studied words and nonstudied words on a subsequent test. Participants were also informed that they would complete five of these study-test cycles. Each study list consisted of 50 words. Participants were randomly assigned to encoding task. In the "strong" encoding condition, participants were instructed to judge whether or not each presented word was pleasant, whereas in the weak condition participants judged whether the word contained the letter *e*. Each study trial started automatically, and participants were unable to enter any response until the word had been onscreen for 1.5 s. Participants then entered their "yes" (pleasant/contains an *e*) or "no" (unpleasant/does not contain an *e*) response via keyboard. Half of the participants indicated a "yes" response with the *z* key, and a "no" response with the */?* key, whereas this order was reversed for the other half of participants. Immediately following each study block, participants completed a 45 s running arithmetic task.

Following a short distractor task that consisted of adding a series of numbers, participants completed a test block of 50 trials. Unlike the study trials where responses were indicated with a single key press, at test participants responded using the mouse. Participants were seated in such a way that the mouse response required movement from the entire arm (i.e., the gain was identical across all computers and set such that individuals could not respond merely by flicking one's wrist). The mouse position could be adjusted at the discretion of the participant. At the start of each test trial, a small *start* box appeared in the bottom center of the screen, along with two response boxes in the upper corners that contained the words *YES* and *NO* (Figure 2a). The left-right order of the response boxes at test matched the left-right order of response options at study. Upon clicking the start box, the test word appeared in the middle of the screen. Participants then indicated whether they recognized that word as one from the study list by clicking on the appropriate response box, at which point the stimulus word disappeared and the start box reappeared. Unbeknownst to participants, we collected the (x,y) mouse coordinates at a rate of 50 Hz[1] from the moment the start box was clicked until a response was made.

The five study-distractor-test blocks differed in their proportion of targets and foils at test (see Figure 3). Every participant completed test blocks containing 5 (10%), 15 (30%), 25 (50%), 35 (70%), or 45 (90%) targets. The remaining items were unstudied foils. The order in which these study-test blocks appeared was randomized for each participant. Participants were told prior to each test block the proportion of trials on which a target would appear, and reminded that their goal on each test was to respond as accurately as possible.

Results

SDT analysis. We first analyzed the data using typical SDT metrics in order to assess whether our manipulations were effective (see Table 1). Mixed analyses of variance (ANOVAs) on the

[1] This experiment was conducted in MATLAB using Psychtoolbox (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007), where we initially sampled mouse-coordinates at a rate of 100 Hz. In order to confront the possibility that such a high sampling rate introduces unnecessary noise in the recording process (we thank an anonymous reviewer for pointing this out), we simply analyzed the data as if we had sampled at 50 Hz. Critically, the results are identical for both the 50- and 100-Hz analyses.

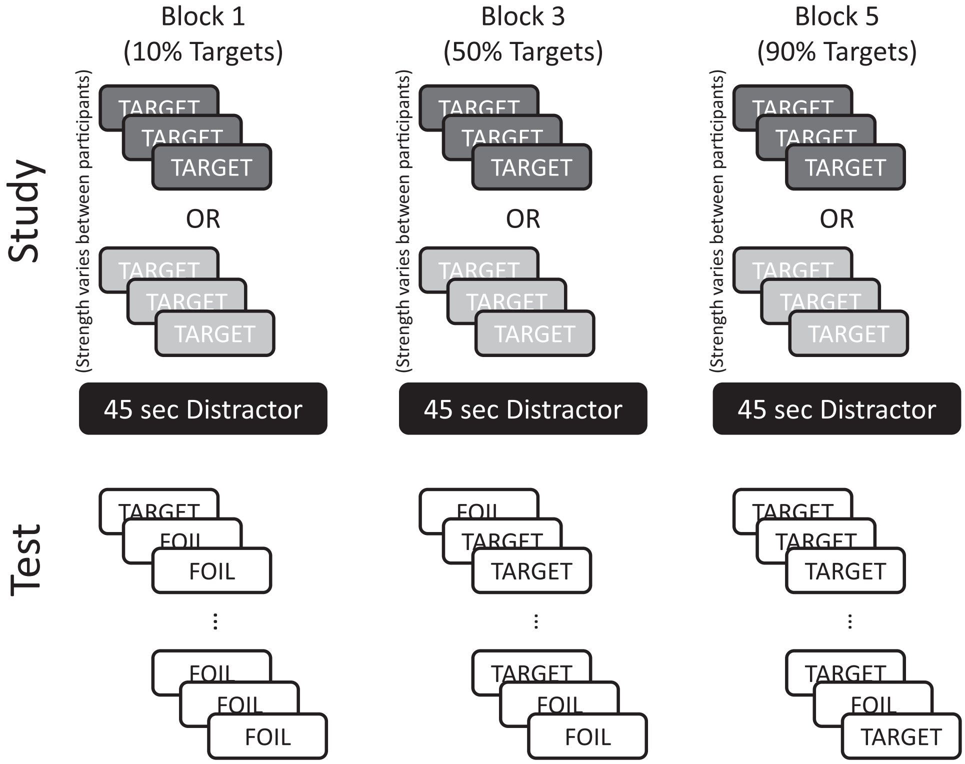

Figure 3. Design and procedure for Experiment 1. Participants completed either "strong" study blocks (indicated by dark boxes), or "weak" study blocks (indicated by light boxes). At test, we varied the percentage of targets (shown only for the 10%, 50%, and 90% blocks for illustrative purposes). Participants were informed about the percentage of targets at test. Order in which participants completed the blocks was randomized.

proportion of test items called "old" confirmed the presence of the SBME, showing an interaction between trial type (target, foil) and encoding strength (strong, weak), $F(1, 159) = 124.94$, $p < .001$, $\eta_p^2 = .44$. We next analyzed bias ($C = -0.5(z(HR) + z(FAR))$ to examine whether our base rate manipulation had the intended effect on recognition bias (see Table 1).[2] Note that this computation of C assumes a fixed foil distribution and is generally interdependent with d' (Grider & Malmberg, 2008). As expected, C was affected by the base rate manipulation, $F(3.36, 534.74) = 44.01$, $\eta_p^2 = .22$, $p < .001$.[3] Participants tended to be more conservative on blocks consisting of few targets and more liberal on blocks with many targets. Finally, these base rate effects were more extreme in strong conditions than in weak conditions, producing a base rate by encoding strength interaction, $F(3.36, 534.74) = 6.32$, $p < .001$, $\eta_p^2 = .04$. We further examined this interaction by looking at the simple main effect of base rate for the strong and weak conditions. Both conditions showed the predicted effect of base rate (Fs > 7.5, ps < .001, $\eta_p^2 = .08$ and .42 for weak and strong, respectively), though this effect was more pronounced in the strong condition than the weak condition.

Trajectory analysis. Figure 4 depicts the time-normalized aggregate response trajectories for each strength by response condition. We calculated a number of derivative measures reported in the literature for each participant's response data (see Table 2). Although recognition studies typically report hit rates and false-alarm rates (i.e., 1-correct rejection rates), it is most reasonable to analyze hits and correct rejections when examining response trajectories to maximize the number of trials contributing to the analysis and thus minimize measurement noise. Throughout the article, statistical analysis is limited to those participants that had at least one trajectory of a given type in each test condition. Three participants did not have a hit in the 10% Targets block, and two participants did not have a correct rejection in the 90% Targets block. These participants were not included in the analyses below. Trials with total response times greater than 3 standard deviations above the mean were excluded from analysis. Although time-normalization was necessary to produce the plots like those in Figure 4, such an adjustment is not necessary for the statistical analyses to follow. Thus, all quantitative analyses throughout the article are performed on the raw trajectory data. We computed all of the same analyses with time-normalization and the results do not change.

Average deviation. One fundamental assumption in response dynamics is that curvature away from one's ultimate choice represents the "competitive pull" of the nonchosen option (cf. McKinstry, Dale,

[2] We report sensitivity (d') for archival purposes but do not consider it further.

[3] We used Greenhouse–Geisser adjusted degrees of freedom where necessary.

Table 1
Accuracy-Based Metrics for Experiment 1

Metric	Percentage targets at test				
	10%	30%	50%	70%	90%
Mean HR (SE)					
Weak	.70 (.03)	.71 (.02)	.76 (.01)	.73 (.02)	.74 (.02)
Strong	.89 (.03)	.86 (.02)	.91 (.02)	.89 (.02)	.91 (.02)
Mean CRR (SE)					
Weak	.79 (.02)	.78 (.01)	.78 (.01)	.74 (.02)	.71 (.03)
Strong	.94 (.02)	.94 (.02)	.93 (.01)	.91 (.02)	.88 (.03)
Mean d' (SE)					
Weak	1.47 (0.09)	1.49 (0.09)	1.67 (0.09)	1.48 (0.09)	1.32 (0.09)
Strong	2.74 (0.10)	2.91 (0.10)	3.06 (0.10)	2.87 (0.10)	2.57 (0.10)
C					
Weak	.19 (.04)	.12 (.05)	.02 (.03)	.03 (.05)	−.11 (.05)
Strong	.36 (.05)	.22 (.05)	.02 (.04)	−.03 (.05)	−.30 (.05)

Note. C is the signal detection parameter of response bias. Zero is unbiased, negative values indicate a liberal bias and positive values indicate a conservative bias. Although hit rate (HR) and false-alarm rate are typically reported, we report correct rejection rate (CRR) to remain consistent with our focus on correct rejection trials during the analysis of response trajectories, as discussed in the text.

& Spivey, 2008). Similar to Papesh and Goldinger (2012), we analyzed the curvature in these response data, and calculated the average deviation (*AD*; Figure 2b) of each response. In order to calculate *AD*, we measured the deviation of each sampled *x,y* pair from a direct path connecting the first and last coordinates of each response. *AD* is simply the average of these deviations across the entirety of the response.[4]

We analyzed *AD* for both hits and correct rejections using 2 (strength) × 5 (base rate) mixed ANOVAs. Both response types showed main effects of base rate condition (hits: $F(2.85, 444.44) = 52.73$, $p < .001$, $\eta_p^2 = .25$; correct rejections: $F(2.60, 407.58) = 99.26$, $p < .001$, $\eta_p^2 = .39$). For hits, linear contrasts demonstrated that responses were most direct in the 90% targets condition, and successively increased in curvature through the 10% targets condition, $F(1, 156) = 120.99$, $p < .001$, $\eta_p^2 = .44$. The opposite pattern was true for correct rejections, $F(1, 157) = 226.59$, $p < .001$, $\eta_p^2 = .59$. Thus, participants showed the most curvature *away* from the correct response on those instances when the correct response was contrary to the most frequent response (e.g., a correct "new" response in the 90% targets condition). Figure 4 indicates that this pattern was more pronounced in the strong encoding condition than in the weak condition. In other words, there was a base rate by strength interaction for hits, $F(2.85, 444.44) = 2.51$, $p = .061$, $\eta_p^2 = .02$, and correct rejections, $F(2.60, 407.58) = 15.52$, $p < .001$, $\eta_p^2 = .09$.

Initial degree. Papesh and Goldinger (2012) suggested that absolute curvature reflects subjective memory strength (e.g., confidence that the item was studied) and indeed this has intuitive appeal. However, analysis of *AD* in the above experiments showed that this measure was also impacted by the test composition, which is a manipulation of bias not memory strength. One possibility is that response bias, induced by the base rate manipulation, reveals itself early in the response process and leads to immediate deflections in the mouse response. This early component has been suggested as an index of bias in simple decision tasks (Buetti & Kerzel, 2009). Following Buetti and Kerzel, we measured *initial degree*, calculated as the degree of deviation from vertical at the point a participant has completed one fifth of the response movement. If participants had not moved horizontally (i.e., $x = 0$), the *initial degree* would be 0°. Moving directly left without any accompanying vertical movement produces an *initial degree* of −90°, whereas moving to the right without vertical movement produces an *initial degree* of 90°. Finally, note we discarded a small number of trials (6.67% of all trials across all participants) in which the initial movement along the *y*-axis was negative.

Figure 5 depicts the distribution of *initial degree of deviation* for every trial of every participant in each experimental condition. The dashed line in Figure 5 represents the average *initial degree* across participants in each condition. It is clear that regardless of whether the test item was a target or a foil and regardless of the ultimate response chosen, participants are more "new" biased when the proportion of targets at test is low, and more "old" biased when the proportion of targets at test is high. A 2 (strength) × 5 (base rate) mixed-factors ANOVA on these data confirmed the pattern obvious in the figure. There was an effect of base rate for hits, $F(2.99, 460.10) = 86.18$, $p < .001$, $\eta_p^2 = .36$, and correct rejections, $F(2.99, 468.69) = 99.72$, $p < .001$, $\eta_p^2 = .39$. Hits and correct rejections also showed an interaction between strength and base rate (hits: $F(2.99, 460.10) = 8.43$, $p < .001$, $\eta_p^2 = .05$; correct rejections: $F(2.99, 468.69) = 10.30$, $p < .001$, $\eta_p^2 = .06$), consistent with the patterns observed for the other measures.[5]

X_{flips} in motion. X_{flips} quantifies the amount of uncertainty over the course of the recognition decision by measuring the number of times a participant's left/right heading changes (Dale, Roche, Snyder, & Mc-

[4] In many mouse-tracking studies, area under the curve is presented as a measure of response curvature. Area under the curve and average deviation are both approximations of the integral. Similar to area under the curve, responses that deviate below the direct path are included as negative deviations (cf. Freeman & Ambady, 2011).

[5] We also evaluated whether the relationship between bias and *initial degree* changes according to the time taken to initiate movement (cf. Buetti & Kerzel, 2009). We divided the response by a median split on RT and included this factor in the analysis. The results are identical and there was no interaction between fast/slow RTs and any other variable. We conducted a similar analysis for Experiment 4 with the same results—no dependency of initial degree on RT.

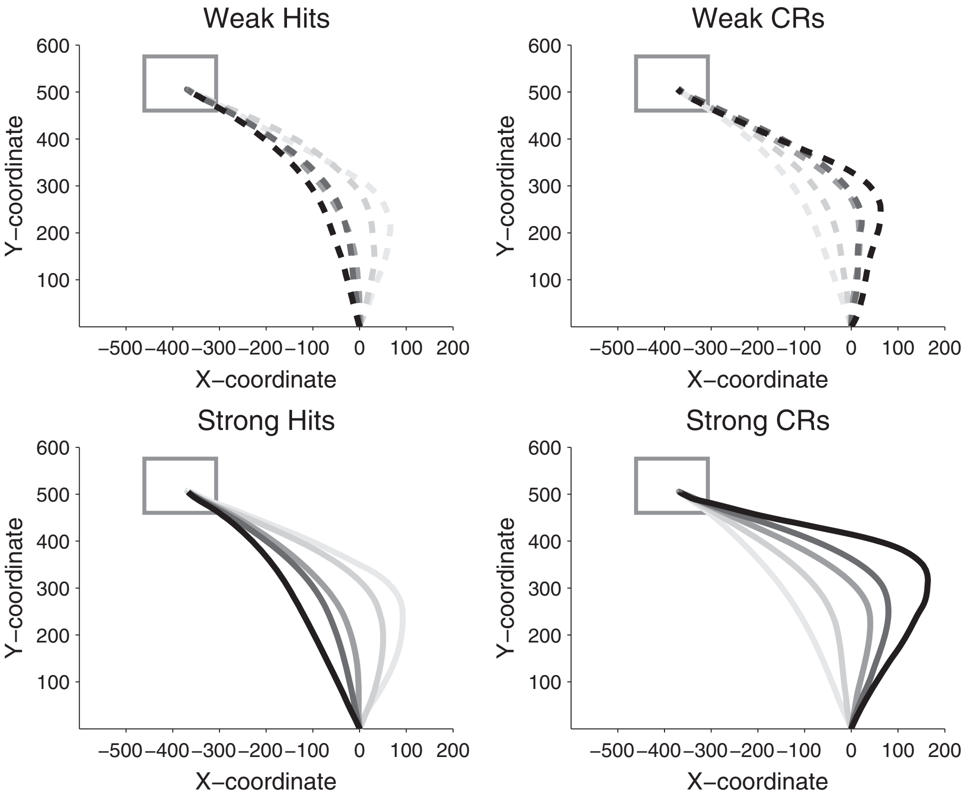

Figure 4. Aggregate response trajectories for Experiment 1. Line color, moving from light to dark, reflects blocks consisting of 10%, 30%, 50%, 70%, and 90% targets. CR = correct rejection.

Call, 2008; Duran, Dale, & McNamara, 2010; Freeman & Ambady, 2011; Hehman, Stolier, & Freeman, 2015; Koop, 2013; Koop & Johnson, 2013).[6] To evaluate changes in the decision process outside of the early bias-sensitive component (captured by *initial degree*) and presumably to increase sensitivity to mnemonic evidence we divided each response trajectory into early and late movement windows. The suggested point at which to segment the response trajectory is after one fifth of the response movement has been completed (cf. Buetti & Kerzel, 2009). Thus, any indecision on the part of participants occurring after this point (the latter four fifths of the response, referred to as X_{flips} *in motion*) is more likely due to the quality of evidence rather than some combination of response bias and mnemonic quality. Recent work in intertemporal choice utilizing primary component analysis identified X_{flips} as being particularly sensitive to indecision caused by choice alternatives eliciting roughly equal preference (Cheng & González-Vallejo, 2015). In other words, X_{flips} were most sensitive to trials where individuals had a difficult time discriminating between the two options. This unique characteristic of X_{flips} makes it a good candidate to indicate memory strength.

Table 2 provides the number of X_{flips} *in motion* for hits and correct rejections in Experiment 1. For hits, X_{flips} *in motion* do not show an interaction between base rate and strength ($F < 1$). For hits there was also an effect of strength independent of base rate, $F(1, 156) = 4.70$, $p = .032$, $\eta_p^2 = .03$. Inspection of the pattern across blocks indicates that strong targets provoked fewer reversals than did weak targets. In addition, the base rate manipulation affected the number of X_{flips} *in motion*, and participants showed the least uncertainty in the 90% targets block and the most uncertainty in the 10% targets block, $F(2.96, 462.42) = 28.33$, $p < .001$, $\eta_p^2 = .15$. For correct rejections there was a base rate by strength interaction, $F(3.08, 483.10) = 6.62$, $p < .001$, $\eta_p^2 = .04$. Just as with the other metrics, participants in the strong encoding condition were more influenced by the base rate manipulation than were individuals in the weak encoding condition. There was also a main effect of base rate, with participants showing increasing X_{flips} *in motion* moving from the 10% targets condition to the 90% targets condition, $F(3.08, 483.10) = 42.92$, $p < .001$, $\eta_p^2 = .22$.[7]

Discussion

Our goal for Experiment 1 was to examine whether response dynamics could provide unique indicators of sensitivity and re-

[6] Some measures of the reversal in direction, such as the zero-crossing of the velocity profile, have been discussed as proxies for the onset of new motor plans. Note, however, that X_{flips} necessarily measure the same data. Indeed, using this alternative formulation does not change the results.

[7] One possible critique of the X_{fips} metric is that it is susceptible to random "twitches" or coordinate parsing errors by the experimental program. In order to address these claims, we recalculated X_{flips} using more conservative methods. First, we counted only those X_{flips} that persisted at least 100 ms before switching back. Second, we calculated a measure of X_{flips} where flips had to be initially greater than a single pixel to eliminate parsing errors. Under these more restrictive calculations, the pattern of results did not change. The single exception is that in Experiment 4 a nonsignificant effect of strength became statistically significant when using the 100-ms criterion.

Table 2
Derivative Measures From Hit and Correct Rejection Response Trajectories in Experiment 1

Metric	Percentage targets at test				
	10%	30%	50%	70%	90%
			Hits		
Mean average deviation (SE)					
Weak	83.49 (8.54)	69.92 (6.73)	45.26 (4.08)	47.87 (4.26)	36.87 (3.63)
Strong	94.51 (9.09)	76.10 (7.17)	46.53 (4.35)	38.85 (4.54)	22.48 (3.87)
Mean X_{flips} *in motion* (SE)					
Weak	1.54 (0.08)	1.38 (0.06)	1.21 (0.05)	1.23 (0.05)	1.11 (0.05)
Strong	1.40 (0.09)	1.29 (0.06)	1.15 (0.06)	1.09 (0.06)	0.87 (0.05)
			Correct rejections		
Mean average deviation (SE)					
Weak	31.13 (2.81)	46.12 (4.07)	65.34 (4.97)	71.06 (7.34)	86.47 (8.89)
Strong	23.55 (3.01)	44.25 (4.36)	73.45 (5.33)	100.16 (7.87)	146.25 (9.53)
Mean X_{flips} *in motion* (SE)					
Weak	1.11 (0.05)	1.23 (0.05)	1.42 (0.05)	1.50 (0.07)	1.38 (0.08)
Strong	0.89 (0.05)	1.18 (0.05)	1.42 (0.06)	1.51 (0.07)	1.64 (0.08)

sponse bias in a typical single-item recognition experiment. To accomplish this aim, we manipulated strength between-participants via a depth of encoding task, and manipulated bias within-participants by varying the proportion of targets presented at test. The experimental manipulations replicated typical findings in terms of accuracy: participants showed higher hit rates and lower false-alarm rates in strong encoding conditions relative to weak encoding conditions. Furthermore, participants' response bias changed to reflect the most common stimulus type for that block (i.e., most "old" biased when the majority of test trials were targets). In terms of response dynamics, the first notable finding from these experiments was that the pattern of mouse movement clearly reflected the test composition manipulation. As shown in *AD*, responses were most direct when they matched the most frequent correct response for each block. This pattern appeared to be driven by early deflections, as measured by the *initial degree* of deviation from the start button. Thus, the effect of test composition was present in even the earliest mouse movements. Another goal was to examine whether aspects of the response trajectory varied with encoding strength, which would ostensibly reflect the locations of evidence distributions in SDT independent of response bias. Neither of our proposed strength measures (*AD* and X_{flips} *in motion*) showed a consistent ability to depict differences in encoding strength without being overwhelmed by the test composition manipulation. Given that the effects of the base rate manipulation were seen early in the response trajectories, it was hoped that $X_{\text{flip}s}$ *in motion* could provide a unique measure of strength by ignoring those X_{flips} occurring early in the trial (within the first fifth of movement) and focusing on those occurring during the heart of each trial (*in motion*). For hits, individuals in the strong condition showed fewer X_{flips} *in motion* in their response than did individuals in the weak condition, though this trend did not hold for correct rejections.

Papesh and Goldinger (2012) suggested that the curvature in response trajectories might correspond to memory strength as indicated by self-reported confidence. However, many other applications of SDT consider confidence as a measure of subjective response bias (e.g., receiver operating characteristic analyses) rather than a 'pure' measure of mnemonic evidence. From that perspective, all findings thus far confirm that measurements of mouse tracking reflect bias. The ability to clearly measure changes in response bias has significant theoretical implications. Banks (1970) noted that "While most researchers will find *Cj* a useful measure in spite of its limitations, the need for a perfectly general psychological index of criterion still exists" (p. 86). We suggest that the initial degree of mouse movements might provide the solution. In support of this claim, we replicated this study *without* informing participants about the base rates for each test list. The choice data and *C* indicated a smaller change in response bias than Experiment 1, but *initial degree* continued to be a robust indicator of changes to response bias.[8] Next, we continue to pursue the goal of identifying a metric of mnemonic evidence when response bias for an "old" or "new" response is eliminated by using a forced choice design.

Experiments 2 and 3

In Experiments 2 and 3 we sought to limit the effects of bias on response trajectories by altering the experimental design. In a two-alternative forced-choice (2AFC) task, participants must choose which of two test items (one studied target and one unstudied foil) was presented on the study list. Because this only requires a relative judgment (which option has more evidence) rather than evaluating a single option against a threshold, there is no decision criterion in the SDT-sense to adjust (cf. Starns, Staub, & Chen, 2015). We therefore utilized two-alternative forced-choice designs with mixed encoding strength (Experiment 2) and pure encoding strength (Experiment 3) manipulations. Thus, the aim of these two experiments was to examine the ability of response dynamics to depict memory strength under conditions presumably absent the influence of response bias.

[8] Data available upon request and on Amy H. Criss's website.

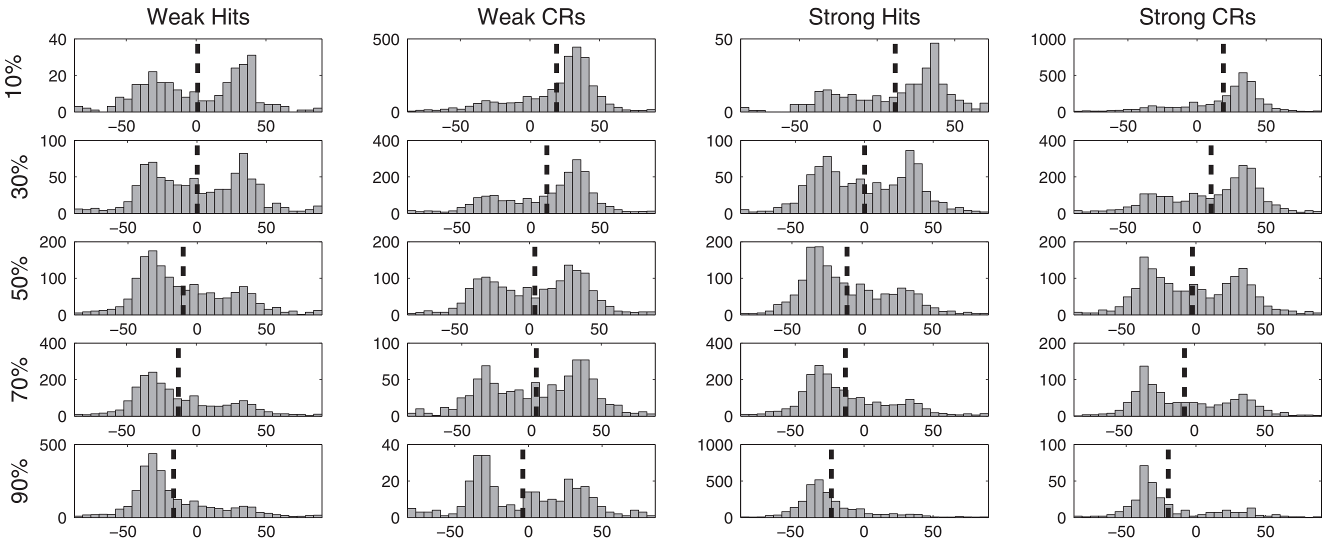

Figure 5. Histograms of *initial degree of deviation* for every trial in each of the base rate conditions in Experiment 1. The *x*-axis represents degree of deviation from vertical and the *y*-axis represents frequency. Negative degrees represent movement toward the "OLD" response, positive degrees represent movement toward the "NEW" response, and 0 degrees represents a perfectly vertical response. Note that because of the manipulation of base rate, the number of trials in each histogram differs. Dashed black line represents the mean of each distribution. CR = correct rejection.

Method

Participants. Participants were recruited from the Syracuse University research participation pool and received credits toward the fulfillment of course requirements. There were 58 participants in Experiment 2 and 55 in Experiment 3.

Stimulus materials. The word pool was identical to that used in Experiment 1.

Design and procedure. In both experiments, participants completed four study-test cycles of 50 words each. The encoding tasks used in both studies were identical to those used in Experiment 1. In Experiment 2 we utilized completely mixed-strength study lists, where encoding strength randomly varied across item within each study block (see Figure 6). The order in which strong and weak encoding tasks appeared was fully randomized trial-by-trial with the lone constraint that the number of strong and weak trials was equivalent over the course of the experiment. Encoding instructions appeared 500 ms prior to presentation of the study word. Participants were allowed to enter a response 1.5 s after stimulus presentation (via keyboard, exactly as in Experiment 1). In Experiment 3 we utilized a pure-strength design where encoding task varied between—but not within—blocks (see Figure 6). Prior to each study block, participants saw a prompt informing them of the encoding task they would be performing for the current block. Each participant in Experiment 3 saw two strong blocks and two weak blocks, the order of which was randomized. Participants were again required to wait 1.5 s after stimulus presentation to enter a response.

Following each study list, participants engaged in a short distractor task before being tested over the study list. The test phase in Experiments 2 and 3 was two-alternative forced-choice (see Figure 6) and included 50 test trials. The response boxes were populated with a studied target and an unstudied foil after participants clicked the "start" box. Participants then simply had to click on what they believed to be the previously studied word. The left/right presentation order of targets and foils was randomly selected for each trial.

Results

Accuracy data. Table 3 contains the proportion of correct responses for Experiments 2 and 3. As expected, participants were

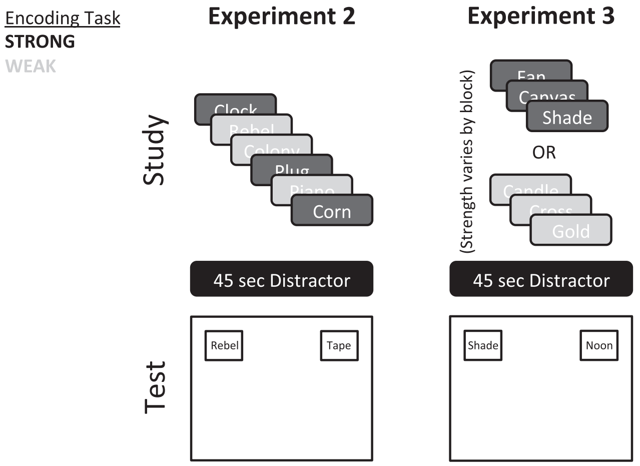

Figure 6. Design for Experiments 2 and 3. In the study phase, trial color represents encoding task, where dark is a deep encoding task and light is a shallow encoding task. Note that for Experiment 2 encoding strength was manipulated within study block whereas in Experiment 3 encoding strength was manipulated entirely between blocks.

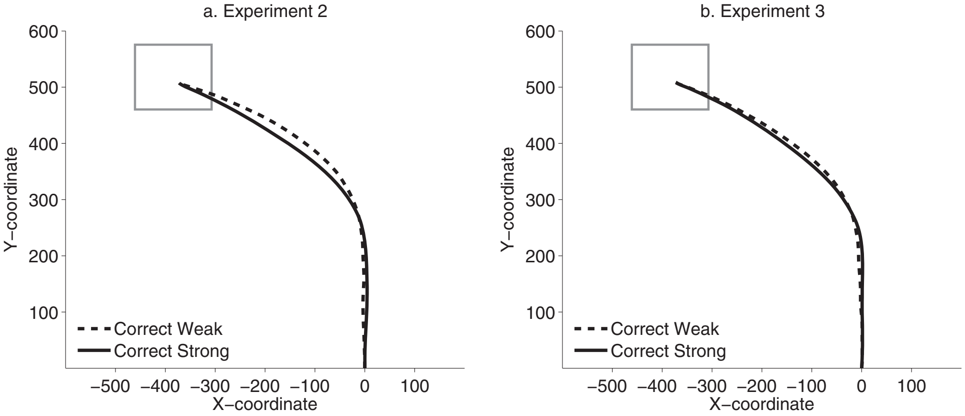

Figure 7. Aggregate response trajectories for Experiments 2 (a) and 3 (b). Solid lines represent selection of the strong (correct) response, whereas dashed lines represent selection of the weak (correct) response.

better at selecting the target when targets were from the strong encoding condition relative to when targets were from the weak encoding condition. This held true in both the within-list design of Experiment 2, $t(57) = 9.77$, $p < .001$, $d = 1.28$, and the between-list design of Experiment 3, $t(54) = 11.59$, $p < .001$, $d = 1.56$.

Trajectory analysis. Figure 7 shows the aggregate mouse trajectories divided by encoding strength (weak/strong) for mixed list-strength (a) and pure list-strength (b) designs. Across both studies, strong targets appeared to be selected more directly than did weak targets. In Experiment 2, strong targets elicited more direct responses (i.e., lower *AD*) than weak targets, $t(57) = 4.61$, $p < .001$, $d = 0.61$, whereas X_{flips} *in motion* similarly showed that participants were less indecisive for strong targets than for weak targets, $t(57) = 5.30$, $p < .001$, $d = 0.70$. The same pattern was found in Experiment 3 (though it is difficult to visualize in the aggregate trajectories). *AD* indicates that participants did, in fact, show more direct paths when selecting strong targets than they did when selecting weak targets, $t(54) = 2.25$, $p = .029$, $d = 0.30$. X_{flips} *in motion* indicated that participants showed less uncertainty when selecting strong targets than when selecting weak targets, $t(54) = 2.85$, $p = .006$, $d = 0.38$.[9]

Discussion

Our goal in Experiments 2 and 3 was to examine whether curvature (*AD*) and uncertainty (X_{flips} *in motion*) in the mouse response were indicative of mnemonic evidence as manipulated by the classic experimental variable encoding strength. In Experiment 1, strength most often interacted with the bias manipulation, and neither *AD* or X_{flips} *in motion* was consistently a unique signature of encoding strength. In order to better clarify the effects of strength on curvature, we attempted to remove the possibility of bias contaminating the measure of strength by switching to a two-alternative forced choice design. Both experiments showed that correct selections of strong targets were more direct (as measured by *AD*) and subject to less uncertainty (as measured by X_{flips} *in motion*) than selections of weak targets. In other words, these measures can provide an indication of memory evidence, but these effects were perhaps dwarfed by the bias manipulation in the first experiment. Collectively, these data suggest that memory strength can be seen both in *AD* and X_{flips} *in motion*, and we hesitantly suggest that X_{flips} *in motion* may be a slightly more robust indicator of strength because it can be more easily disentangled from effects of *initial degree.*

Table 3
Accuracy and Response Metrics for Experiments 2 and 3

Metric	Experiment 2 *M* (*SE*)	Experiment 3 *M* (*SE*)
Pr(Correct)		
Weak	.84 (.01)	.81 (.02)
Strong	.94 (.01)	.95 (.01)
Average deviation		
Weak	81.98 (5.30)	72.73 (5.47)
Strong	71.60 (5.59)	66.54 (4.95)
X_{flips} *in motion*		
Weak	1.52 (0.06)	1.40 (0.06)
Strong	1.36 (0.05)	1.31 (0.06)

Note. All reported trajectory data are exclusively from correct response trials.

Experiment 4

The studies presented above demonstrate that importing response dynamic metrics from other domains into the SDT framework for recognition memory can convey information about response bias (*AD*, *initial degree*) and memory strength (*AD*, X_{flips} *in motion*). Such information may prove critical in theoretical debates regarding the contributions of memory evidence and response bias to performance. For example, we can apply the method to a debate surrounding the strength-based mirror effect (SBME; Stretch & Wixted, 1998; Hilford, Glanzer, & Kim, 1997). The SBME is the finding that strengthening items at study, whether through repetition, duration, or depth of encoding, leads to improved recognition performance by way of a higher hit rate and a lower false-alarm rate for the strongly encoded list compared with the weakly encoded list. The decrease in false-

[9] In 2AFC designs *initial degree* is not related to bias for an old or new response (i.e., it would indicate a bias to pick the item on the left or right side of the screen). In the interest of thoroughness, we examined whether *initial degree* differed as a function of target strength and it did not, Experiment 2: $t(57) = 0.84$, $p = .402$ Experiment 3: $t(54) = 0.21$, $p = .833$.

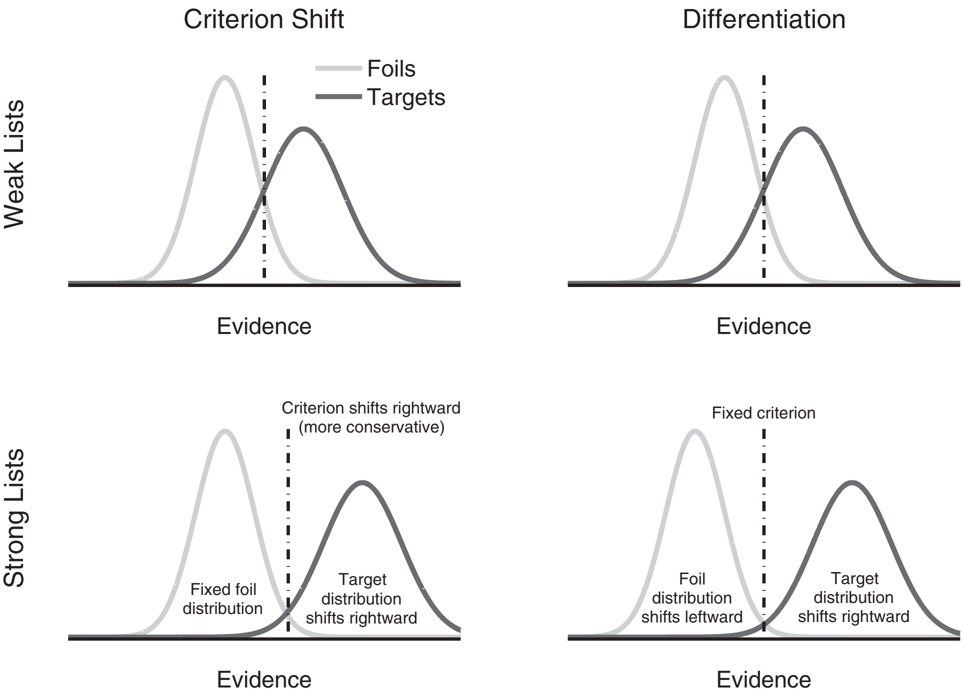

Figure 8. Stylized distributions of mnemonic evidence for targets (dark) and foils (light), including the decision criterion (dashed line). Comparing across list-strength rows demonstrates both models predict a rightward shift in target distributions. Critically, the criterion-shift account (left column) predicts a fixed foil distribution and a criterion shift, whereas a differentiation account (right column) predicts a fixed criterion and shift in the foil distribution.

alarm rates is particularly interesting because foil items seem to benefit from a stronger encoding task for which they, by definition, were not present. One theoretical account suggests that this reliable effect is the product of changes to both the target and foil distributions (the differentiation account; e.g., Criss, 2006; Criss & Koop, 2015), whereas a second account assumes it is produced by a fixed foil distribution and a shift in the criterion (the criterion-shift account; e.g., Stretch & Wixted, 1998).

The differentiation account is that as episodic memory traces become more complete and more accurate, the probability of correctly matching a test item to its appropriate memory trace increases while the probability of accidentally confusing a foil with a previously encoded target decreases. The target distribution increases in evidence and the foil distribution decreases (see Figure 8). Differentiation models require no change in the criterion to account for the SBME, though they do have a criterion that is free to vary (Criss, 2006; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997).

Opposing the differentiation account is the criterion-shift account that arose from applications of SDT to recognition memory. The assumption was that the foil distribution reflects a baseline familiarity of items that is unaffected by experimental manipulations. Under this assumption, changes to the FAR can only be produced by shifting the decision criterion (see Hirshman, 1995; Stretch & Wixted, 1998; Cary & Reder, 2003; Starns, White, & Ratcliff, 2010, for different implementations of this same idea). This assumption highlights a critical difference in how the criterion-shift and differentiation accounts view the source of the foil distribution and its functional mobility (see Figure 8). In the differentiation account, when the contents of memory are well stored (e.g., a strongly encoded list), any randomly selected foil will produce a relatively poor match. When the contents of memory are poorly stored (e.g., a weakly encoded list), the foil is more likely to match by chance. Thus the distributions reflect evidence generated by the match between the test items (targets or foils) and the content of episodic memory. The amount of evidence therefore changes as a function of that content (note that differentiation models are not versions of SDT but instead are process models that specify the representation of information in memory, the processes of encoding, and the decision rule).

Critically, the differentiation - criterion-shift debate has persisted despite the application of a number of different methodologies, including analysis of discrete recognition behavior (Criss, 2006; Starns et al. 2010), subjective familiarity ratings (Criss, 2009; Starns, Ratcliff, & White, 2012), response time distributions (Criss, 2009; Starns, White, & Ratcliff, 2012) and neural measures (Criss, Wheeler, & McClelland, 2013; Hemmer, Criss, & Wyble, 2011) highlighting the need for novel measures.

In a final experiment, we bring the metrics and methods established in the first three experiments to bear on the typical SBME design. Experiment 1 indicated that response dynamics could clearly depict changes in response bias by way of the *initial degree* metric. Therefore, if the SBME is due to decision bias, responses following strong lists should show immediate deflections that are more "new" biased than responses following weak lists. Alternatively, if the SBME is due to changes in the foil distribution, we should see effects largely in our measure of strength (X_{flips} *in motion*) but definitely not in *initial degree*. Because *AD* reflects both response bias and mnemonic evidence, it is not diagnostic here.

Method

Participants. 38 participants from the Syracuse University research participation pool took part in this experiment in exchange for partial fulfillment of course requirements.

Stimulus materials. Word stimuli and selection procedures were identical to all prior experiments.

Design and procedure. Participants completed four study-test cycles of 50 words each. Participants performed the shallow encoding task on two study blocks, and the deep encoding task on two other study blocks followed by a single item *yes/no* recognition test containing 50 trials, half targets and half foils randomly intermixed. The order in which the four blocks appeared was randomized for each participant. Participants were instructed prior to each block about the encoding task. Encoding tasks, timing, and all other details were identical to those used in all previous studies.

Results

As expected, participants' choice data showed the SBME. Relative to weak lists, strong lists elicited higher hit rates, $t(37) = 7.50$, $p < .001$, $d = 1.22$ and lower false-alarm rates, $t(37) = -5.19$, $p < .001$, $d = 0.84$ (see Table 4).

Observation of Figure 9 suggests that the most apparent difference between trajectories is that hits (dark lines) are substantially more direct than are correct rejections (light lines). There is no obvious pattern of differences between strong and weak conditions for either hit or CR trials. A 2 (trial type: hits or correct rejections) × 2 (strength: strong or weak) repeated-measures ANOVA on *AD* confirmed that correct rejections showed greater curvature than hits, $F(1, 37) = 5.98$, $p = .019$, $\eta^2_p = .14$, but that there was no main effect of the strength manipulation, nor was there an interaction between the two factors (*F*s < 1).

Moving to X_{flips} *in motion*, a 2 (trial type) × 2 (strength) repeated-measures ANOVA indicated that correct rejections showed more X_{flips} *in motion* than did hits, $F(1, 37) = 4.85$, $p = .034$, $\eta^2_p = .12$ (see Table 4). Furthermore, strong trials elicited marginally less uncertainty than weak trials but that difference was not significant, $F(1, 37) = 3.04$, $p = .090$, $\eta^2_p = .08$. There was not a significant interaction between these factors ($F < 1$). X_{flips} *in motion* provides a measure of memory evidence and provides at least an indication, though not compelling, that strong targets and foils provide more evidence than do weak targets and foils.

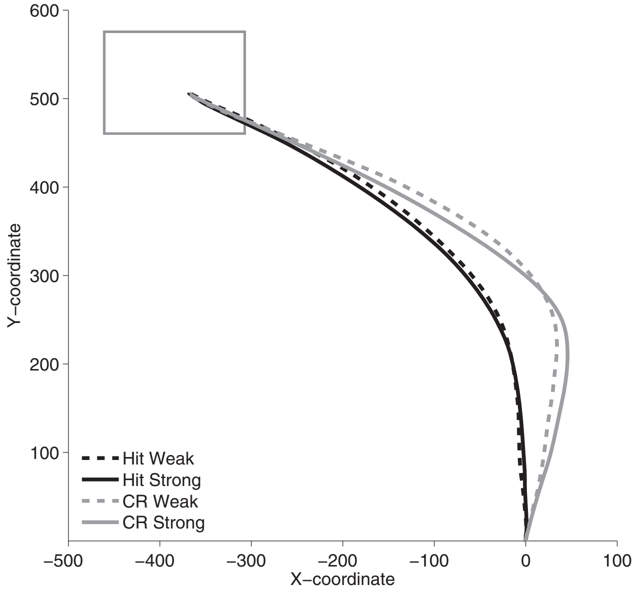

Figure 9. Aggregate response trajectories from Experiment 4 for hits (dark lines) and correct rejections (CR; light lines) following strong (solid lines) and weak (dashed lines) lists.

Figure 10 shows the distribution for *initial degree* in each condition. Recall that the criterion-shift account predicts a conservative shift in the criterion for strong lists relative to weak lists. A 2 (trial type) × 2 (strength) repeated-measures ANOVA failed to reveal any main effects or interactions involving strength, (*F*s < 1).

Table 4
Accuracy and Response Metrics for Experiment 4

Metric	Trial type	
	Targets (*SE*)	Foils (*SE*)
Pr(Correct)		
Weak	.71 (.03)	.79 (.02)
Strong	.88 (.02)	.89 (.02)
Average deviation		
Weak	54.09 (6.26)	75.76 (7.17)
Strong	50.05 (7.10)	75.18 (8.35)
X_{flips} *in motion*		
Weak	1.18 (0.08)	1.37 (0.08)
Strong	1.13 (0.07)	1.28 (0.08)

Note. Pr(Correct) for target trials represents the hit rate, whereas Pr(Correct) for foil trials represents the correct rejection rate. Trajectory data (average deviation, X_{flips} *in motion*) for target and foil trials are calculated from hits and correct rejections, respectively.

Discussion

In this final experiment, the aim was to use response dynamics to provide a test of competing accounts of the strength-based mirror effect. Specifically, if the SBME were produced by a criterion shift, the *initial degree* for strong lists should be more "new" biased than that for weak lists. Alternatively, if the SBME were due to changes in the quality of evidence provided by the test items, strong lists should produce less indecision as indicated by fewer X_{flips} *in motion*. Metrics calculated from the continuous response data showed no evidence to support the conclusion that participants adopted a more conservative criterion following strong lists than weak lists despite being robust indices of response bias in Experiment 1. There is a slight indication that X_{flips} *in motion* were more likely for weak than strong lists. In other words, we are very confident that *initial degree* reflects response bias and we find no evidence for differences in response bias in the SBME paradigm. We are less confident that X_{flips} *in motion* index sub-

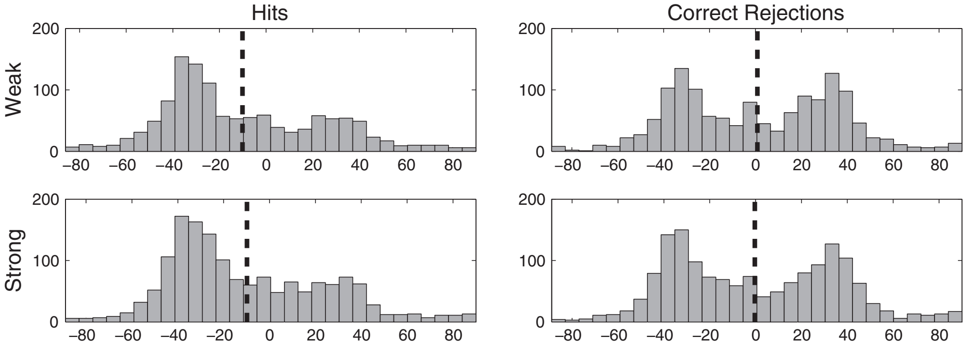

Figure 10. Histograms of *initial degree of deviation* for every trial in Experiment 4. The *x*-axis represents degree of deviation from vertical and the *y*-axis represents frequency. Negative degrees represent movement toward the "OLD" response, positive degrees represent movement toward the "NEW" response, and 0 degrees represents a perfectly vertical response. Dashed black line represents the mean of each distribution.

jective memory strength and we find weak evidence for differences in memory evidence in the SBME paradigm. This in combination with many previous studies showing evidence for differentiation (see Criss & Koop, 2015, for a review; Criss, 2006; Criss, 2009; Criss, 2010; Criss, Aue, & Kiliç, 2014; Criss & Howard, 2015; Criss, Malmberg, & Shiffrin, 2011; Criss et al., 2013; McClelland & Chappell, 1998; Murnane & Shiffrin, 1991; Ratcliff, Clark, & Shiffrin, 1990; Shiffrin, Ratcliff, & Clark, 1990; Shiffrin & Steyvers, 1997) lead us to interpret the data provided by Experiment 4 as being more in accordance with the differentiation account of the SBME than the criterion-shift account.

General Discussion

The primary aim of this work was to examine whether response dynamics could provide an index of sensitivity and response bias in recognition memory with the long term goal of using these metrics to advance theory, especially in cases where it is impossible to evaluate whether changes in memory performance arise from a change in the decision criterion or the evidence distributions using known SDT metrics. The secondary aim was to apply response dynamics metrics to one such empirical situation, the strength-based mirror-effect paradigm. In Experiment 1 (and in an unpublished replication) we manipulated depth of encoding and response bias in tandem. The primary finding from this study was that even participants' first response movements measured via *initial degree* indicated base-rate induced bias and this bias carried over into measurements of *AD*. In a bias-free two-alternative forced-choice design (Experiments 2 and 3), both *AD* and X_{flips} *in motion* reflected encoding strength. Selections of strong targets tended to be more direct and showed less uncertainty than selections of weak targets. Experiment 4 used the typical SBME design, where a single-item test phase follows a pure-strength study phase. In accordance with the predictions of the differentiation account, participants did not become more conservative following strong lists relative to weak lists as measured by *initial degree*. Furthermore, the descriptive pattern in X_{flips} *in motion* was such that correct responses following strong lists showed less uncertainty than those following weak lists. Although not definitive, these data provide converging evidence in favor of a differentiation account of the SBME (see Criss & Koop, 2015, for a review)

Advantages of Response Dynamic Metrics

A common critique of response dynamics is that the method does not offer anything that traditional analyses of response times cannot. The collection of mouse-tracking data in no way precludes collection of response time data. However, use of simple response time analyses fails to match the resolution afforded by response dynamics about the nature of the decision process. For example, in Experiment 1 mean response times would be blind to any initial movements *away* from the ultimately selected choice as is very clear when using response dynamic metrics. Such reversals are of significant theoretical interest (Resulaj, Kiani, Wolpert, & Shadlen, 2009; Koop & Johnson, 2013; Koop, 2013; Barca & Pezzulo, 2015; Walsh & Anderson, 2009; Freeman & Dale, 2013) and indicate continued evidence processing following response initiation. In the present context, the reversals seen in Experiment 1 indicate that base rate induced bias has an immediate impact on the recognition process that must subsequently be overcome with additional evidence accumulation. Depicting such a reversal would be impossible with simple reaction time (RT) analyses.

Analysis of response time distributions address some of these concerns in that they provide more information than accuracy or discrete choices alone. However, such analyses are uncommon, in part due to serious practical limitations including the large number of observations (e.g., requiring several sessions from participants in most applications to memory) and the computational skills required to conduct such analyses. Further, like SDT measures, analysis of response time distributions depends on the theoretical assumptions of the model, whereas response dynamics are theory-free. Indeed for the experimental manipulations reported here, analysis of response time distribution was conducted with the diffusion model (Criss, 2010; Starns et al., 2012) and the results are inconclusive in that the critical parameter (drift rate) could be interpreted in terms of a response bias or memory evidence.

Response dynamics provide a new way to evaluate the constructs underlying memory. One of the main contributions of this article is the identification of *initial degree* as a robust indicator of response bias. Of the other two metrics evaluated here, X_{flips} *in motion* was perhaps a slightly better indicator of memory strength. Response dynamics are a relatively new measurement tool in cognitive psychology (Spivey et al., 2005) and as such the complete space of possible analyses is not yet fully established. Indeed throughout the peer-review process, we became aware of the high variability in preferred measurements and the disparate views about which measures are ideal. This suggests a clear avenue for further investigation. We focused on established metrics *AD*, *initial degree*, and X_{flips} *in motion*, however many other measurements exist and others are in development that better preserve the temporal dynamics of the response (e.g., Cox, Kachergis, & Shiffrin, 2012). The application of response dynamics to measure theoretical mechanisms underlying memory is novel. In developing this connection, care must be taken to integrate multiple dependent measures including discrete responses, response time, and response dynamics into a cohesive theoretical framework. The work here is a step toward that end. A promising avenue for further research is to evaluate other aspects of the trajectory that better reflect memory evidence and directly integrate with models of memory.

Strength and Criterion Shifts

The secondary goal of this research was to use response dynamics to address the theoretical debate as to whether differentiation or a criterion shift best describes the data when encoding strength is manipulated between lists. Although a review of the literature indicates that differentiation is the best theoretical approach, much research has focused on how the criterion changes in response to encoding strength. For example, many papers seek to determine whether the criterion is set based on expectations developed about the study list or the test items and whether or not the criterion changes on a trial-by-trial basis to match these expectations (e.g., Cary & Reder, 2003; Hirshman, 1995; Hockley & Niewiadomski, 2007; Morrell, Gaitan, & Wixted, 2002; Starns et al., 2010, 2012; Stretch & Wixted, 1998; Verde & Rotello, 2007).

We previously examined the fundamentals of this relationship by playing it out to its logical extreme: test lists consisting of entirely studied or unstudied items. If the criterion changes in response to the strength of the study or test list, then these shifts should be most noticeable when the test list consists entirely of targets or entirely of foils. Work involving such test lists (Cox & Dobbins, 2011; Koop, Criss, & Malmberg, 2015; Wallace, 1982) has shown a remarkable similarity between these test lists and standard test lists that include both targets and foils. These findings suggest that individuals are poor at using (even dramatic) unsignaled differences in strength to shift the decision criterion in recognition testing. Criterion shifts of the size one might expect if individuals can adjust the criterion on the basis of strength differences were only seen when participants were also provided with corrective response feedback for tests of distractor-free or target-free lists (Koop et al., 2015).

Different Types of Criteria

Though manipulations of base rate have a long history of use as a bias manipulation (e.g., Criss, 2009, 2010; Estes & Maddox, 1995; Healy & Kubovy, 1978, 1981; Rhodes & Jacoby, 2007), the drift diffusion measurement model (Ratcliff, 1978) incorporates additional bias parameters that are not affected by such a manipulation. In short, the claim is that manipulations of base rate affect the *amount* of information required to make a recognition decision, whereas other manipulations (including the SBME design) affect the *quality* of information required (Starns et al., 2012; White & Poldrack, 2013). Critically, the latter, called the drift criterion, cannot be discriminated from the quality of evidence present in the stimulus. In other words, this type of criterion is bound in the process of evidence evaluation. We have not yet sought to identify a component of the trajectory that corresponds to the drift criterion and the current data provide no evidence about the validity of the drift criterion. It is possible that this construct reflects a parameter specific to the diffusion model but not a unique signature of cognitive processing that can be evaluated independently. Although it is possible that there are multiple types of criteria, this debate awaits further evidence.

Conclusion

The series of experiments presented here demonstrates the ability of response dynamics to convey information about common constructs in the memory literature like sensitivity and response bias. This method provides dependent variables that can be easily applied to debates in the broader literature without being tied to the theoretical baggage associated with measurement models like SDT or the diffusion model. Our specific application of response dynamics to the SBME provided data more in accordance with the differentiation account. More broadly, response dynamics can bring novel tests of process-based predictions in domains as diverse as dual-process models of recognition, recall-to-reject strategies in plurality discrimination, and other debates surrounding interference versus criterion shifts. The development of such process-oriented predictions certainly reflects cognitive systems that operate dynamically, and thus the use of methods that can at least partially preserve the continuous nature of the response process is an important step in developing increasingly accurate models of cognition.

References

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39,* 445–459. http://dx.doi.org/10.3758/BF03193014

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin, 74,* 81–99. http://dx.doi.org/10.1037/h0029531

Barca, L., & Pezzulo, G. (2015). Tracking second thoughts: Continuous and discrete revision processes during visual lexical decision. *PLoS ONE, 10,* e0116193. http://dx.doi.org/10.1371/journal.pone.0116193

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436. http://dx.doi.org/10.1163/156856897X00357

Bruhn, P., Huette, S., & Spivey, M. (2014). Degree of certainty modulates anticipatory processes in real time. *Journal of Experimental Psychology: Human Perception and Performance, 40,* 525–538. http://dx.doi.org/10.1037/a0034365

Buetti, S., & Kerzel, D. (2009). Conflicts during response selection affect response programming: Reactions toward the source of stimulation. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 816–834. http://dx.doi.org/10.1037/a0011092

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language, 49,* 231–248. http://dx.doi.org/10.1016/S0749-596X(03)00061-5

Cheng, J., & González-Vallejo, C. (2015, April). *Action dynamics in intertemporal choice reveal different facets of indecisiveness.* Poster session presented at the Annual Meeting of the Midwestern Psychological Association, Chicago, IL.

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology, 92,* 1006–1023. http://dx.doi.org/10.1037/0022-3514.92.6.1006

Cox, G. E., Kachergis, G., & Shiffrin, R. M. (2012). Gaussian process regression for trajectory analysis. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1668–1673). Austin, TX: Cognitive Science Society.

Cox, J. C., & Dobbins, I. G. (2011). The striking similarities between standard, distractor-free, and target-free recognition. *Memory & Cognition, 39,* 925–940. http://dx.doi.org/10.3758/s13421-011-0090-3

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 671–684. http://dx.doi.org/10.1016/S0022-5371(72)80001-X

Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory and Language, 55,* 461–478. http://dx.doi.org/10.1016/j.jml.2006.08.003

Criss, A. H. (2009). The distribution of subjective memory strength: List strength and response bias. *Cognitive Psychology, 59,* 297–319. http://dx.doi.org/10.1016/j.cogpsych.2009.07.003

Criss, A. H. (2010). Differentiation and response bias in episodic memory: Evidence from reaction time distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 484–499. http://dx.doi.org/10.1037/a0018435

Criss, A. H., Aue, W., & Kiliç, A. (2014). Age and response bias: Evidence from the strength based mirror effect. *Quarterly Journal of Experimental Psychology, 67,* 1910–1924.

Criss, A. H., & Howard, M. (2015). Models of episodic memory. In J. Busemeyer, J. T. Townsend, Z. Wang, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 165–183). New York, NY: Oxford University Press.

Criss, A. H., & Koop, G. J. (2015). Differentiation in episodic memory. In J. Raaijmakers, A. H. Criss, R. Goldstone, R. Nosofsky, & M. Steyvers (Eds.), *Cognitive modeling in perception and memory: A Festschrift for Richard M. Shiffrin* (pp. 112–125). New York, NY: Psychology Press.

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language, 64,* 119–132.

Criss, A. H., Wheeler, M. E., & McClelland, J. L. (2013). A differentiation account of recognition memory: Evidence from fMRI. *Journal of Cognitive Neuroscience, 25,* 421–435.

Dale, R., Kehoe, C., & Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory & Cognition, 35,* 15–28. http://dx.doi.org/10.3758/BF03195938

Dale, R., Roche, J., Snyder, K., & McCall, R. (2008). Exploring action dynamics as an index of paired-associate learning. *PLoS ONE, 3,* e1728.

Duchek, J. M., & Neely, J. H. (1989). A dissociative Word-Frequency × Levels-of-Processing interaction in episodic recognition and lexical decision tasks. *Memory & Cognition, 17,* 148–162. http://dx.doi.org/10.3758/BF03197065

Duran, N. D., Dale, R., & McNamara, D. S. (2010). The action dynamics of overcoming the truth. *Psychonomic Bulletin & Review, 17,* 486–491. http://dx.doi.org/10.3758/PBR.17.4.486

Estes, W. K., & Maddox, W. T. (1995). Interactions of stimulus attributes, base rates, and feedback in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1075–1095. http://dx.doi.org/10.1037/0278-7393.21.5.1075

Flumini, A., Barca, L., Borghi, A. M., & Pezzulo, G. (2014). How do you hold your mouse? Tracking the compatibility effect between hand posture and stimulus size. *Psychological Research*. Advance online publication. http://dx.doi.org/10.1007/s00426-014-0622-0

Freeman, J. B., & Ambady, N. (2011). Hand movements reveal the time-course of shape and pigmentation processing in face categorization. *Psychonomic Bulletin & Review, 18,* 705–712. http://dx.doi.org/10.3758/s13423-011-0097-6

Freeman, J. B., & Dale, R. (2013). Assessing bimodality to detect the presence of a dual cognitive process. *Behavior Research Methods, 45,* 83–97. http://dx.doi.org/10.3758/s13428-012-0225-x

Freeman, J. B., Pauker, K., Apfelbaum, E. P., & Ambady, N. (2010). Continuous dynamics in the real-time perception of race. *Journal of Experimental Social Psychology, 46,* 179–185. http://dx.doi.org/10.1016/j.jesp.2009.10.002

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Grider, R. C., & Malmberg, K. J. (2008). Discriminating between changes in bias and changes in accuracy for recognition memory of emotional stimuli. *Memory & Cognition, 36,* 933–946. http://dx.doi.org/10.3758/MC.36.5.933

Healy, A. F., & Kubovy, M. (1978). The effects of payoffs and prior probabilities on indices of performance and cutoff location in recognition memory. *Memory & Cognition, 6,* 544–553. http://dx.doi.org/10.3758/BF03198243

Healy, A. F., & Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 344–354. http://dx.doi.org/10.1037/0278-7393.7.5.344

Hehman, E., Stolier, R. M., & Freeman, J. B. (2015). Advanced mouse-tracking analytic techniques for enhancing psychological science. *Group Processes & Intergroup Relations, 18,* 384–401. http://dx.doi.org/10.1177/1368430214538325

Hemmer, P., Criss, A. H., & Wyble, B. (2011, November). *Assessing a neural basis for differentiation accounts of recognition memory*. Poster session presented at the Psychonomic Society meeting. Seattle, WA.

Hilford, A., Glanzer, M., & Kim, K. (1997). Encoding, repetition, and the mirror effect in recognition memory: Symmetry in motion. *Memory & Cognition, 25,* 593–605. http://dx.doi.org/10.3758/BF03211302

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 302–313. http://dx.doi.org/10.1037/0278-7393.21.2.302

Hockley, W. E., & Niewiadomski, M. W. (2007). Strength-based mirror effects in item and associative recognition: Evidence for within-list criterion changes. *Memory & Cognition, 35,* 679–688. http://dx.doi.org/10.3758/BF03193306

Kantner, J., & Lindsay, D. S. (2010). Can corrective feedback improve recognition memory? *Memory & Cognition, 38,* 389–406. http://dx.doi.org/10.3758/MC.38.4.389

Kiliç, A. (2012). *Output interference and strength based mirror effect in recognition memory (Doctoral dissertation)*. Retrieved from SURFACE Psychology-Dissertations.

King, A. C., Kiernan, M., Oman, R. F., Kraemer, H. C., Hull, M., & Ahn, D. (1997). Can we identify who will adhere to long-term physical activity? Signal detection methodology as a potential aid to clinical decision making. *Health Psychology, 16,* 380–389. http://dx.doi.org/10.1037/0278-6133.16.4.380

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? [ECVP abstract]. *Perception, 36,* 14.

Koop, G. J. (2013). An assessment of the temporal dynamics of moral decisions. *Judgment and Decision Making, 8,* 527–539.

Koop, G. J., Criss, A. H., & Malmberg, K. J. (2015). The role of mnemonic processes in pure-target and pure-foil recognition memory. *Psychonomic Bulletin & Review, 22,* 509–516. http://dx.doi.org/10.3758/s13423-014-0703-5

Koop, G. J., & Johnson, J. G. (2013). The response dynamics of preferential choice. *Cognitive Psychology, 67,* 151–185. http://dx.doi.org/10.1016/j.cogpsych.2013.09.001

Lepora, N. F., & Pezzulo, G. (2015). Embodied choice: How action influences perceptual decision making. *PLoS Computational Biology, 11*(4), e1004110. http://dx.doi.org/10.1371/journal.pcbi.1004110

MacMillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. New York, NY: Cambridge University Press.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105,* 724–760.

McKinstry, C., Dale, R., & Spivey, M. J. (2008). Action dynamics reveal parallel competition in decision making. *Psychological Science, 19,* 22–24. http://dx.doi.org/10.1111/j.1467-9280.2008.02041.x

Morrell, H. E., Gaitan, S., & Wixted, J. T. (2002). On the nature of the decision axis in signal-detection-based models of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 1095–1110. http://dx.doi.org/10.1037/0278-7393.28.6.1095

Mulder, M. J., Wagenmakers, E. J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2012). Bias in the brain: A diffusion model analysis of prior probability and potential payoff. *The Journal of Neuroscience, 32,* 2335–2343. http://dx.doi.org/10.1523/JNEUROSCI.4156-11.2012

Murnane, K., & Shiffrin, R. M. (1991). Word repetitions in sentence recognition. *Memory & Cognition, 19,* 119–130. http://dx.doi.org/10.3758/BF03197109

Nelson, T. O. (1977). Repetition and depth of processing. *Journal of Verbal Learning and Verbal Behavior, 16,* 151–171. http://dx.doi.org/10.1016/S0022-5371(77)80044-3

Papesh, M. H., & Goldinger, S. D. (2012). Memory in motion: Movement dynamics reveal memory strength. *Psychonomic Bulletin & Review, 19,* 906–913. http://dx.doi.org/10.3758/s13423-012-0281-3

Parks, T. E. (1966). Signal-detectability theory of recognition-memory performance. *Psychological Review, 73,* 44–58. http://dx.doi.org/10.1037/h0022662

Plant, E. A., & Peruche, B. M. (2005). The consequences of race for police officers' responses to criminal suspects. *Psychological Science, 16,* 180–183. http://dx.doi.org/10.1111/j.0956-7976.2005.00800.x

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108. http://dx.doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 163–178. http://dx.doi.org/10.1037/0278-7393.16.2.163

Ratcliff, R., Sheu, C. F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99,* 518–535.

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009, September 10). Changes of mind in decision-making. *Nature, 461,* 263–266. http://dx.doi.org/10.1038/nature08275

Rhodes, M. G., & Jacoby, L. L. (2007). On the dynamic nature of response criterion in recognition memory: Effects of base rate, awareness, and feedback. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 305–320. http://dx.doi.org/10.1037/0278-7393.33.2.305

Rotello, C. M., Macmillan, N. A., Hicks, J. L., & Hautus, M. J. (2006). Interpreting the effects of response bias on remember-know judgments using signal detection and threshold models. *Memory & Cognition, 34,* 1598–1614.

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 179–195. http://dx.doi.org/10.1037/0278-7393.16.2.179

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM-retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145–166. http://dx.doi.org/10.3758/BF03209391

Spivey, M. J., Dale, R., Knoblich, G., & Grosjean, M. (2010). Do curved reaching movements emerge from competing perceptions? A reply to van der Wel et al. *Journal of Experimental Psychology: Human Perception and Performance,* 36:251–254, 2009. http://dx.doi.org/10.1037/a0017170

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences of the United States of America, 102,* 10393–10398. http://dx.doi.org/10.1073/pnas.0503903102

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31,* 137–149. http://dx.doi.org/10.3758/BF03207704

Starns, J. J., Ratcliff, R., & White, C. N. (2012). Diffusion model drift rates can be influenced by decision processes: An analysis of the strength-based mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 1137–1151. http://dx.doi.org/10.1037/a0028151

Starns, J. J., Staub, A., & Chen, T. (2015, July). *Implications of response time and eye movement data for models of forced choice recognition*. Paper presented at the 14thAnnual Summer Interdisciplinary Conference, Mammoth Lakes, CA.

Starns, J. J., White, C. N., & Ratcliff, R. (2010). A direct test of the differentiation mechanism: REM, BCDMEM, and the strength-based mirror effect in recognition memory. *Journal of Memory and Language, 63,* 18–34. http://dx.doi.org/10.1016/j.jml.2010.03.004

Starns, J. J., White, C. N., & Ratcliff, R. (2012). The strength-based mirror effect in subjective strength ratings: The evidence for differentiation can be produced without differentiation. *Memory & Cognition, 40,* 1189–1199. http://dx.doi.org/10.3758/s13421-012-0225-1

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1379–1396. http://dx.doi.org/10.1037/0278-7393.24.6.1379

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition, 35,* 254–262. http://dx.doi.org/10.3758/BF03193446

Wallace, W. P. (1982). Distractor-free recognition tests of memory. *The American Journal of Psychology, 95,* 421–440. http://dx.doi.org/10.2307/1422134

Walsh, M. M., & Anderson, J. R. (2009). The strategic nature of changing your mind. *Cognitive Psychology, 58,* 416–440. http://dx.doi.org/10.1016/j.cogpsych.2008.09.003

White, C. N., & Poldrack, R. A. (2013). Using fMRI to constrain theories of cognition. *Perspectives on Psychological Science, 8,* 79–83. http://dx.doi.org/10.1177/1745691612469029

Received June 24, 2014
Revision received August 25, 2015
Accepted August 26, 2015 ■