

# The reliability of criterion shifting in recognition memory is task dependent

Bryan A. Franks<sup>1</sup> · Jason L. Hicks<sup>1</sup>

Published online: 11 July 2016  
© Psychonomic Society, Inc. 2016

**Abstract** We explored 2 research questions: whether criterion shifts within different recognition tasks are reliable across testing sessions and whether such shifts correlate across the different tasks. In Experiment 1, we established consistent group-level criterion shifting based on expected strength of target items in the test. False-alarm rates were higher when expected strength was weak as opposed to strong, even when expected strength cues were unblocked at test. Test–retest reliabilities in this strength-cuing environment were good. In Experiment 2, we manipulated either the probability (high or low) or expected memory strength (strong or weak) of target items and had people perform both tasks on each of 2 days of testing. Varying the probability of target items consistently produced criterion shifts, as did manipulating target memory strength. Regarding individual differences, shifting in a given test context predicted shifting on a second day in that same test context. However, one’s tendency to shift a decision criterion on 1 type of test context did not predict one’s tendency to do so in the other. The extent to which people shifted their decision criterion in recognition memory testing was largely dependent on the type of test.

**Keywords** Recognition memory · Criterion shifts · Criterion setting · Individual differences

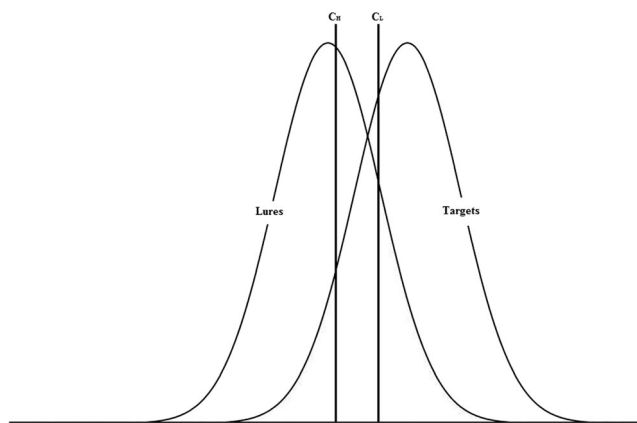
Recognition memory is the ability to identify previously experienced stimuli and to distinguish these familiar stimuli from novel ones. In a typical recognition memory experiment,

people study a set of items (e.g., words, pictures) under various encoding conditions and, after a brief delay, are given a memory test in which they respond “old” or “new” or give confidence ratings of their memory for a mixture of studied and nonstudied items. One prominent framework used to characterize recognition memory performance is signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 2005). Note that SDT is not a model of memory, per se, but is a measurement model that has been used successfully to characterize detection paradigms based on stimulus strength (e.g., Wickelgren, 1968), including memory. Operationalizing recognition memory in the context of SDT, people set a standard of evidence along a continuum of memory strength, which is then used as the basis for memory decisions. Figure 1 depicts this framework visually. The target distribution (i.e., studied items) sits farther to the right because of that distribution’s higher overall strength. The distance between the means of the distributions ( $d'$ ) represents one common measure of memory discriminability. A criterion must be established to serve as a decision threshold to distinguish old from new items, with items being called “old” if their strength is perceived to be greater than the criterion. Lures that pass the criterion are false alarms, and therefore false-alarm rates (FARs) can serve as a basic measure of criterion placement if the lure distribution is fixed. Two possible criteria are shown in the figure, the left-most vertical line representing a greater willingness to call test items “old” as opposed to the right-most line. Using one criterion versus another for different items within a given recognition test is considered a within-list criterion shift.

Some research on criterion shifts has attempted to induce them by directly manipulating memory strength via study repetitions (e.g., Hicks & Starns, 2014; Stretch & Wixted, 1998), study duration (Bruno, Higham, & Perfect, 2009), study-test delay (Singer & Wixted, 2006), or by manipulating the base

✉ Bryan A. Franks  
bfran19@lsu.edu

<sup>1</sup> Department of Psychology, Louisiana State University, Baton Rouge, LA 70803-5501, USA



**Fig. 1** Theoretical distributions for targets and lures. The vertical lines represent different conceptual criterion placements for high probability targets ( $C_H$ ) and for low probability ( $C_L$ ) targets

rates of test items (e.g., Healy & Kubovy, 1978; Rhodes & Jacoby, 2007; Selmecky & Dobbins, 2013). However, even when these manipulations are accompanied by feedback (Verde & Rotello, 2007) or by strength-predictive test cues (Hicks & Starns, 2014), within-list-based criterion shifts are not always observed. It is not uncommon for researchers to conduct multiple experiments examining criterion shifts only to find null effects, small effects, or both (e.g., Bruno et al., 2009; Hicks & Starns, 2014; Verde & Rotello, 2007). One reason for these mixed outcomes could be large individual differences in the amount that people are willing to shift their criterion (Aminoff et al., 2012). For example, suppose that slightly less than half of participants in an experiment show a small to moderate amount of shifting, whereas the majority of participants show absolutely no criterion shift. This would lead to a small, negligible average shift as a group, even though many individuals exhibited shifting.

Research has used SDT measures to investigate recognition memory at the level of the individual to explain these individual differences and what they mean for memory performance in general. For example, Kantner and Lindsay (2012) found that where people set their criterion (relatively conservative, neutral, or liberal) was highly consistent across tests, both within the same testing session and 1 week later. The authors interpreted this finding as criterion setting reflecting a cognitive trait such that individuals naturally differ in the amount of memorial evidence they need to classify items as old (see also Kantner & Lindsay, 2014). Aminoff et al. (2012) explored criterion shifts using a target probability, or base rate, manipulation at test and found considerable variability in the magnitude at which people were willing to shift their criterion within a single test. Some people shifted a great deal whereas others did not. Nonetheless, the amount of shifting was correlated across different tests and stimuli within the same experimental session. According to Aminoff et al. (2012), this finding “likely reflects a general tendency in the flexibility of criterion shifting” (p. 1025). The overarching goal of our

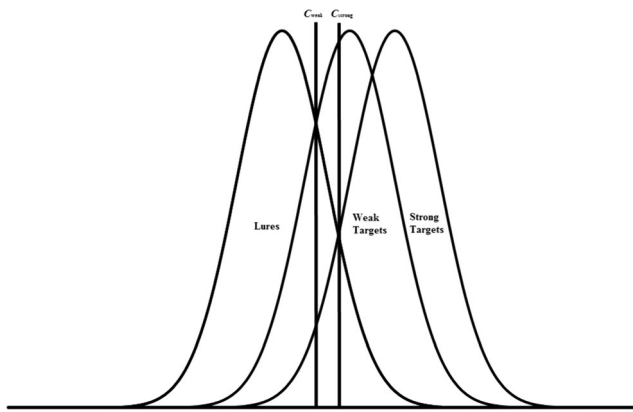
study was to evaluate this assertion within the context of recognition memory.

We explored Aminoff et al.’s (2012) claim by investigating two primary research questions. First, are criterion shifts reliable across study-test cycles when participants attempt to shift their criterion in response to cues of expected memory strength rather than to probability cues? Establishing that shifts within a given test context are reliable is a precondition of examining further generalization. Second, is the magnitude of criterion shifts reliable across different test formats (i.e., probability based vs. strength based)? Correlated criterion shifts across different test formats would support Aminoff et al.’s (2012) claims that criterion shifting may be trait-like, whereas a failure to find this relationship would suggest that amount a person is able to shift their criterion is task dependent.

### Empirical shifts in target probability versus strength-based manipulations

Aminoff et al. (2012) tested people with different target probability test blocks. In high-probability blocks, the probability of targets was 70 % and lures was 30 %. In low-target probability blocks, these percentages were reversed. The different blocks contained between six and nine test items each and alternated across the test. People were told about these differences and given instructions that high-probability blocks would be shown in a certain font color and low-probability blocks in a different color. FARs changed systematically with this manipulation—they were higher in high-probability blocks than in low-probability blocks, with a FAR difference of about 20 %. This difference represents a criterion shift that people were able to apply during the test, with criterion  $C_H$  in Fig. 1 used during high-probability blocks and criterion  $C_L$  used during low-probability blocks. Probability-based criterion shifts, sometimes referred to as base rate shifts, rely on people’s awareness of the probability differences within blocks or on test-based performance feedback (Han & Dobbins, 2008; Rhodes & Jacoby, 2007), even in cases of extreme base-rate manipulations (e.g., Cox & Dobbins, 2011; Koop, Criss, & Malmberg, 2015). In our experiments we applied a probability manipulation similar to Aminoff et al. (2012), with awareness to produce a shift.

Our second method of producing a shift relied on a strength-based manipulation (e.g., Hicks & Starns, 2014; Starns & Olchowski 2015). This scenario is depicted in Fig. 2, with theoretical strength distributions for lures, items studied once (weak targets), and items studied four times (strong targets). Notice that the strong targets have greater memory strength than the weak targets and that the weak targets are stronger in memory than the lures. The vertical lines  $C_{Strong}$  and  $C_{Weak}$  represent the potential criteria for the



**Fig. 2** Theoretical distributions for lures, weak targets, and strong targets. The two vertical lines represent the different criteria. The line labeled ( $C_{\text{weak}}$ ) represents a conceptual criterion placement for weak targets and the line labeled ( $C_{\text{strong}}$ ) represents a conceptual criterion placement for strong targets

strong and weak targets, respectively. Hicks and Starns (2014) showed that one way to motivate strength-based criterion shifts was to alternate blocks of test items that include lures with only strong targets or with only weak targets. Font color of the test stimuli was used to cue the expected strength of target items for each block, implicitly suggesting that people might establish different thresholds for calling items “old” when shown in one color versus the other. When people expected to encounter only strong targets with lures, their FARs were lower than when they expected only weak targets with lures. This criterion shift was largest when the like-strength blocks were either 20 or 40 items in length. The shift was weakest (statistically null) when test items were not presented in like-strength test blocks. Null shifts in within-subjects, strength-based manipulations are more the norm when strength cues in tests are unblocked (e.g., Hicks & Starns, 2014; Starns & Olchowski 2015; Stretch & Wixted, 1998).

One notable exception to this common finding comes from Starns and Olchowski (2015), who found large strength-based criterion shifts in an unblocked test condition. They explored whether null group shifts in prior work resulted from peoples’ unwillingness to make an effortful shift or because people failed to process the strength cues. To this end, they used a study-test procedure very similar to Hicks and Starns (2014; Experiments 3A–3C) and manipulated the test format by varying participants’ response options. In the two-key condition, participants simply responded “old” or “new” for each test item. In the three-key condition, participants used one key to respond “old” to words with a strong cue, a separate key to respond “old” to words with a weak cue, and a third key to identify items of either strength cue “new.” Starns and Olchowski (2015) reasoned that by using separate keys to respond “old” to items of different strength classes, participants were forced to process the strength cues to make a response, whereas in the two-key condition, participants may

disregard the strength cues (Benjamin, Diaz, & Wee, 2009). The authors found a 13 % (Experiment 1A) and 11 % (Experiment 2) difference between the weak and strong FARs in the three-key conditions and no significant FAR differences in the two-key conditions. Starns and Olchowski (2015) also found that these significant group-level criterion shifts were observed over multiple study-test cycles (Experiment 3).

### Comparison of strength-based and probability-based criterion shifts

Prior work has been successful in producing criterion shifts when using a manipulation of target probability (e.g., Aminoff et al., 2012) versus a manipulation of expected memory strength (e.g., Hicks & Starns, 2014; Stretch & Wixted, 1998) at test. However, prior work has not elicited probability-based criterion shifts solely from varying the base rates or probability of targets versus lures but rather in concert with other manipulations. The same experimental procedures that promote probability-based shifts also promote strength-based shifts. For example, manipulations of feedback (e.g., Estes & Maddox, 1995; Verde & Rotello, 2007), response-key mappings (e.g., Rhodes & Jacoby; Starns & Olchowski, 2015), and some implicit or explicit information at test that a subject can use to judge a test item’s potential probability (e.g., Heit, Brockdorff, & Lamberts, 2003; Selmecky & Dobbins, 2013), or expected strength (Hicks & Starns, 2014), are often necessary. Additionally, participants’ natural awareness of base rates (Titus, 1973; Rhodes & Jacoby, 2007) or of expected strength information (Starns & Olchowski 2015) seems to influence shifting. In our own laboratory we found that manipulating the base rates of targets versus lures does not motivate a criterion shift without explicit cues that accompany test items as being either likely old or likely new (see Koop & Criss, 2015, for an analogous finding).<sup>1</sup> Thus, among these types of shifts, participants need to be made aware of the contingencies, and experimental procedures must be put into place to ensure that probability or expected strength cues are fully processed—then it is up to the subject to apply the processed cue information in the context of the recognition judgments.

Despite these behavioral similarities between shifts produced by base rates and by expected strength, there are also some apparent differences. One difference is that these manipulations may impact different aspects of the recognition

<sup>1</sup> In our study, one condition was a replication of the Aminoff et al. (2012) work wherein participants were given color cues that signaled the probability of a test item being old. In a second condition, we kept the test composition exactly the same but removed the color cues. Participants shifted their criterion only when given cues. We are happy to share these data with anyone who is interested.

decision process that unfolds over time (e.g., Criss, 2009, 2010; White & Poldrack, 2014). Drift diffusion modeling, for example, posits two important parameters that might influence a basic SDT measure of criterion setting (Criss, 2010; Starns, Ratcliff, & White, 2012). A starting point parameter reflects how conservative or liberal one's criterion might be at the start of a trial. Base-rate manipulations have been shown to impact this particular parameter by influencing an initial expectation about what type of response is most likely to be correct, described by White and Poldrack (2014) as *response expectancy*. In contrast, the drift-criterion parameter in the diffusion model influences the amount of evidence needed to make a particular choice decision, which depends on the accumulation of evidence over repeated sampling after the test item has been perceived. Starns et al. (2012) showed that expected strength had an influence on the drift criterion, rather than on the starting point. Drift rates for lures differed when they were in strong-target versus weak-target test blocks, suggesting that people changed their thresholds based on expected strength. Similarly, White and Poldrack (2014) showed the same influence on drift criterion (and drift rates) when asking people before each trial to be either conservative (i.e., expect a strong memory signal before responding) or liberal (i.e., expect a weak memory signal before responding). The dissociation between base rates influencing the starting point, but evidence or strength expectancy influencing the drift criterion, could be one reason effects of these different manipulations might not correlate at the coarse level we are examining.<sup>2</sup>

To summarize, although modeling evidence suggests a dissociation between aspects of criterion setting affected by base rates versus expected strength, many of the same factors that influence expected strength-based shifts also influence base-rate shifts, and arguably do so for the same reasons (e.g., enhancing awareness, learning about the test composition, being fully aware of test cues). For both types of shifts, people must fully process cues and apply that information on their own. Whether participants adjust decision rules throughout test from feedback or by being given explicit strength or probability cues up front, it is still up to participants to decide when and if to use these cues. To the extent that criterion shifting is

related to a general proclivity or ability to apply cues to shift a criterion even in the context of different decision rules (e.g., maximize accuracy vs. maximize long-run reward; Bohil & Maddox, 2003), we reasoned that there should be some consistency in shifting behavior, regardless of whether those retrieval cues pertain to probability information or expected memory strength. Indeed, in Aminoff et al.'s (2012) proposed model of criterion shifting, procedural differences and memory ability are only part of what influences whether a person will shift. In addition to these components, their model also includes factors such as participants' "willingness" to shift, the extent to which they use cue information, inherent characteristics (e.g., personality traits, affect), and participants' actual ability to shift their criterion. Hence, we acknowledge that although comparing shifting ability in different types of recognition tasks has an inherent limitation, there are also shared factors that influence both strength-based and probability-based criterion shifts, and so we might expect to observe some consistency in shifting behavior regardless of the task.

To address the predictive relationship between criterion shifting in two different recognition tasks, we asked participants to complete recognition tests in which we manipulated either the base rate (probability based) or expected memory strength (strength based) of the test items. Before addressing this question, we first needed to establish that strength-based criterion shifts were correlated across tests. As noted by Hicks and Starns (2014), the reliability of strength-based shifts has yet to be established. In a pilot study, we initially explored this question using the 20-item block condition from the Hicks and Starns (2014) two-key paradigm and by testing participants in two sessions spaced 2 days apart. Unfortunately, we found no overall group shift on either day and did not find reliability of criterion shifts among individuals across tests.<sup>3</sup> We explored this question again in Experiment 1, but this time using different test formats and having participants complete two study-test cycles within the same session. More specifically, we adopted the three-key test condition from Starns and Olchowski (2015) because this methodology produced larger criterion shifts than the Hicks and Starns work did and because Starns and Olchowski found that the group-level criterion shifts were reliable across study-test cycles. Additionally, in Experiment 1 we were interested in exploring whether the effect of blocking test items (Hicks & Starns, 2014) would replicate using the three-key test procedure. At test we employed the three-key procedure in 40-item and 10-item blocks as well as in an unblocked test condition. To anticipate our results, we found reliable shifts to varying degrees in all conditions.

In Experiment 2, participants completed a probability-based and strength-based recognition test in an initial testing

<sup>2</sup> There are also studies in the neuroscience literature isolating neural signatures associated with criterion setting in a base-rate manipulation (e.g., Aminoff et al. 2012; Jaeger, Konkel, & Dobbins, 2013), but not others investigating an expected-strength shift of the type that we investigated here. However, Criss, Wheeler, and McClelland (2013) demonstrated a distinction between criterion-related versus memory-discriminability-related neural signatures in a pure-list strength examination of the strength-based mirror effect. Koop and Criss (2015) generally reviewed work in modeling and in neuroscience examining a differentiation account of how people treat lures in the context of strong targets versus in the context of weak targets in pure lists. It is unknown whether these types of findings generalize to a mixed-list encoding of strength differences and a mixed-list retrieval examining expected strength, where differences in false alarm rates are attributed to criterion shifts (e.g., Hicks & Starns, 2014) as opposed to differentiation.

<sup>3</sup> We are willing to make our pilot data available to anyone who is interested.



session and again 2 days later. If a person's willingness to shift a decision criterion is an inherent cognitive trait (cf. Aminoff et al., 2012), then the amount a person is willing to shift should be consistent across days and across testing formats. In contrast, if criterion shifting in recognition memory does not correlate across tasks, then the potential to establish a general criterion-shifting ability or trait is limited and likely to be task dependent.

## Experiment 1

### Method

#### *Participants*

Participants were 161 psychology students from Louisiana State University who participated in this experiment to fulfill a partial course requirement or for extra credit.

#### *Materials and procedure*

We took 360 words with the following characteristics from the MRC Psycholinguistic Database ([http://www.psy.uwa.edu.au/mrcdatabase/uwa\\_mrc.htm](http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm)): concreteness, familiarity, and imagability ratings all between 200 and 600 on scales ranging from 100 to 700; Kučera-Francis written frequency between 5 and 888; and word length between five and nine letters. The stimuli used for each software program were equivalent on these stimulus dimensions. All stimuli were presented using E-Prime software (Psychology Software Tools, Pittsburgh, PA).

### Strength-based programs

Two strength-based recognition memory software programs were created, each comprised of 180 randomly assigned words. The software randomly selected 80 words from this set to serve as encoded targets and another 80 to serve as lures at test. Ten untested filler items were presented at the beginning of the encoding phase and 10 others at the end of the encoding phase. Equal numbers of the targets were randomly assigned to a weak or strong encoding condition, which reflected the number of study presentations during the encoding procedure (1 or 4, respectively). Thus, the encoding procedure consisted of 220 presentations, which were randomized for each participant. Study items were presented individually for 700 ms and were followed by a blank 100 ms interstimulus interval (ISI). The test consisted of 80 targets and 80 lures in which color cues were given to aid participants in their decisions. Strong targets were presented in red font color on the right side of the screen and weak targets in green

font color on the left side of the screen. Half of the lures were randomly assigned to appear in either red and on the right or green and on the left (i.e., expected strength cues). Participants were randomly assigned to one of three test conditions: 40-item blocks, 10-item blocks, or unblocked test. In the first two conditions, test items were grouped into strength blocks in which there were always an equal number of like-strength targets and lures. Hence, the 40-item block test was comprised of four test blocks of 40 items and the 10-item block test was comprised of 16 blocks of 10 items. In the unblocked condition, test items were not blocked by strength and were randomly presented.

At encoding, participants were told they would study a list of words for a later, unspecified memory test. We also told them they would study some words only once and that some words would be presented multiple times. At test, our instructions closely matched that of Starns and Olchowski (2015). Participants were told that for items presented in red on the right of the screen, they would decide whether that word was studied four times or not studied at all using the keys labeled "4×" and "New," respectively. For items presented in green on the left side of the screen, participants were told they would decide whether the word was studied one time or not studied at all using the keys labeled "1×" and "New," respectively.

Participants completed the two different computer programs (i.e., study-test lists), each of which contained different stimuli for the study-test procedure. These two programs were completed back to back, and we counterbalanced the order of which study-test procedure was completed first. Testing sessions lasted approximately 30 minutes.

## Results and discussion

We first present information regarding performance on the recognition tasks, especially the degree to which criterion shifts were produced at a condition level. Then we focus on individual differences in shifting and whether they correlate across and within each of the three test conditions. Two participants from the unblocked test condition were excluded, one because of a criterion shift measure that was more than 3 standard deviations from the group mean and another whose data were not stored correctly by the computer software. We also examined the data for potential bivariate outliers for the correlational analyses (Tabachnick & Fidell, 2013). No further data points were excluded based on this screening.

Table 1 presents data for each condition, including hit rates (HR), false alarm rates (FAR), a measure of discrimination ( $d'$ ), and a measure of criterion shifting. Regarding the HR and FAR measures, it is obvious that the expected strength factor had opposite effects on HR versus FAR, producing a two-way interaction between (expected) strength and stimulus class,  $F(1, 158) = 668.01$ ,  $MSE = .006$ ,  $p < .001$ ,  $\eta_p^2 = .81$ ,

**Table 1** Group recognition performance in Experiment 1

Condition	Hit rate		False alarm rate		$d'$		FAR shift
	Weak	Strong	Weak	Strong	Weak	Strong	
40-item blocks ( $n = 55$ )							
Test 1	.59 (.02)	.71 (.02)	.35 (.02)	.24 (.02)	.67 (.06)	1.42 (.09)	.11
Test 2	.59 (.02)	.73 (.01)	.34 (.02)	.26 (.03)	.70 (.06)	1.43 (.07)	.08
10-item blocks ( $n = 54$ )							
Test 1	.61 (.02)	.74 (.02)	.37 (.02)	.28 (.02)	.69 (.06)	1.41 (.09)	.09
Test 2	.60 (.02)	.76 (.02)	.36 (.02)	.26 (.02)	.68 (.02)	1.44 (.08)	.10
Unblocked ( $n = 50$ )							
Test 1	.61 (.02)	.80 (.02)	.37 (.03)	.30 (.03)	.68 (.07)	1.49 (.09)	.06
Test 2	.65 (.02)	.79 (.02)	.39 (.02)	.31 (.03)	.72 (.08)	1.50 (.10)	.08

Note. Standard errors of the mean in parentheses.

within a mixed-model ANOVA with factors of Condition (40-item, 10-item, unblocked), Test (1 vs. 2), Expected Strength, and Stimulus Class (targets vs. lures). Therefore, we analyzed the HR and FAR measure separately with 3 (Condition: 40-item, 10-item, or unblocked)  $\times$  2 (Test: Test 1 or Test 2)  $\times$  2 (Strength: Strong or Weak) mixed-factorial ANOVAs. For the HR, there was only a main effect of strength,  $F(1, 156) = 304.76$ ,  $MSE = .011$ ,  $p < .001$ ,  $\eta_p^2 = .66$ . There was no effect of condition,  $F(1, 156) = 2.27$ ,  $p = .11$ , or of test,  $F(1, 156) = 2.42$ ,  $p = .12$ . None of the two-way interactions was significant, all  $F_s < 1$ , nor was the three-way interaction,  $F(1, 156) = 2.57$ ,  $p = .08$ . Similarly, for the FAR there was only a main effect of expected strength,  $F(1, 156) = 126.88$ ,  $MSE = .01$ ,  $p < .001$ ,  $\eta_p^2 = .445$ , demonstrating that people produced fewer false alarms in the context of an expected greater strength. The remaining effects and interactions were not significant, all  $F_s < 1.04$ .

Discriminability ( $d'$ ) was examined using a 3 (Condition: 40-item, 10-item, or unblocked)  $\times$  2 (Test: Test 1 or Test 2)  $\times$  2 (Strength: expected strong or weak) mixed-factorial ANOVA. As anticipated, there was a main effect of strength such that memory was better for the strong test items,  $F(1, 156) = 641.71$ ,  $MSE = .125$ ,  $p < .001$ ,  $\eta_p^2 = .80$ . The remaining main effects and interactions were not significant, all  $F_s < 1$ .

### Criterion shifting at the group level and at the individual level

We created an index of criterion shifting based on the FAR for expected-weak versus expected-strong blocks. We note that the FAR as a measure of criterion placement is appropriate under an assumption that the lure item distribution is fixed and does not react to any of our experimental manipulations (see Fig. 2). For example, one might argue that the difference in FAR across these expected-strength blocks is due to a

differentiation process rather than a criterion shift (see Criss, 2006, for a discussion). However, our mixed-list strength encoding and test procedure makes differentiation very unlikely as an explanation for the FAR differences, favoring a criterion-shift explanation instead (see Hicks & Starns, 2014; Starns & Olchowski 2015, for further discussion).<sup>4</sup> Moreover, a common alternative metric of criterion placement in SDT—the  $C$  index—is problematic for our purposes. This index measures the standardized distance of the criterion relative to the intersection of the target and lure distributions. Negative values of  $C$  occur when the criterion is to the left of that intersection and positive values when it is to the right. For our expected strength conditions, the  $C$  index of criterion shifting is inappropriate because the HR and FAR measures react in opposite directions with the repetition manipulation. The resulting measure of  $C$  for strong or weak items is therefore affected just as much, if not more, by the movement in target item distributions as compared to the changes in FAR across expected-weak versus expected-strong lures. For example, one might produce a relatively more negative value of  $C$  for strong than for weak items only because the HR for strong targets is so much greater than for weak targets, even if the absolute criterion placement for the weak items moves to the left on the memory strength axis. In this sense, focusing on the FAR alone as a measure of criterion placement, and of shifting, eliminates contamination from changes in the HR. This argument also applies to the  $\ln(\beta)$  criterion index in an SDT framework, which is the log likelihood ratio of the height of the target distribution relative to the lure distribution at the point where a criterion lies. Other researchers have strongly

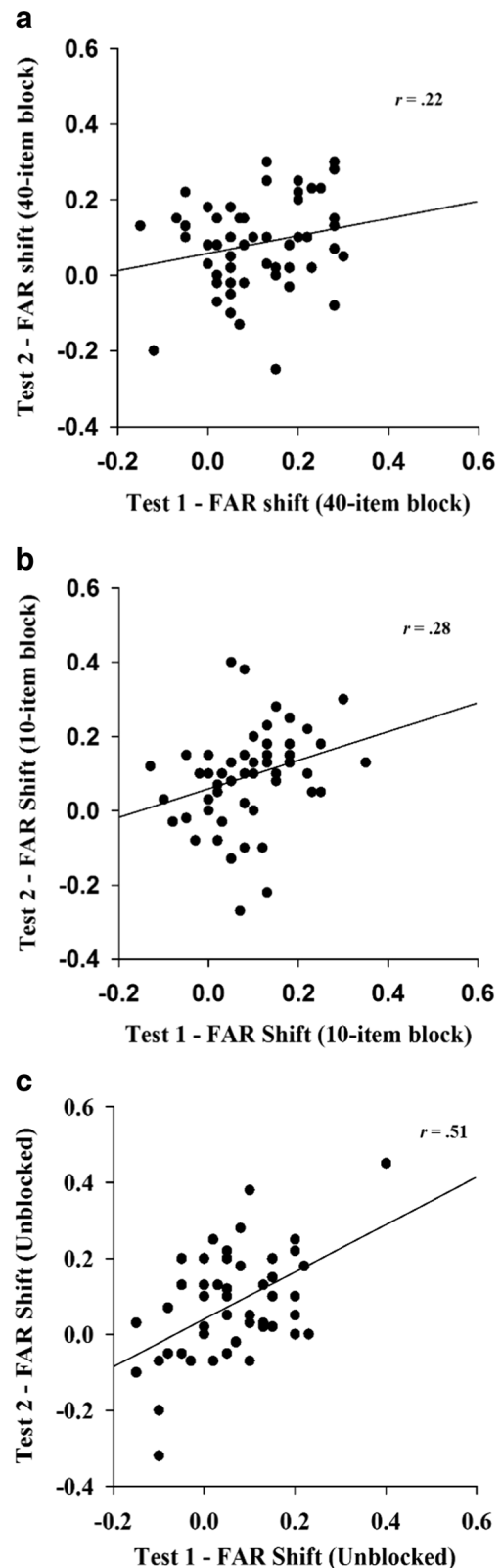
<sup>4</sup> Briefly, differentiation as a candidate explanation for changes in FAR is most likely when the strength of studied items is manipulated either between subjects or between lists for a given subject (see Criss, 2006, for an example of the differentiation-based interpretation), not in cases like ours, where the strength of encoded items was mixed (Hicks & Starns, 2014; Starns & Olchowski 2015).

advocated or at least used FAR differences to quantify criterion shifts for related reasons (Bruno et al., 2009; Dobbins & Kroll, 2005; Hicks & Starns, 2014; Hockley & Niewiadomski, 2007; Singer, 2009; Singer & Wixted, 2006; Starns & Olchowski 2015; Verde & Rotello, 2007). Despite our reservations about the appropriateness of the SDT-based measures of criterion or bias as compared to the FAR measure for measuring shifting behavior, we report our results using these metrics as well so that readers can evaluate the full complement of data. To foreshadow, our outcomes and interpretations do not change across these various measures of criterion shifting.

This shifting measure we created is the FAR of strong blocks subtracted from the FAR of weak blocks, henceforth called the “FAR shift.” As a group, the 40-item condition exhibited large criterion shifts on Test 1,  $t(54) = 7.33$ ,  $p < .001$ , Cohen’s  $d = .99$ , and also on Test 2,  $t(54) = 5.19$ ,  $p < .001$ , Cohen’s  $d = .70$ . Likewise, the 10-item condition also displayed large criterion shifts on Test 1,  $t(53) = 6.76$ ,  $p < .001$ , Cohen’s  $d = .92$ , and on Test 2,  $t(53) = 5.30$ ,  $p < .001$ , Cohen’s  $d = .72$ . We observed moderate criterion shifts in the unblocked condition on both tests: Test 1,  $t(49) = 4.02$ ,  $p < .001$ , Cohen’s  $d = .56$ ; Test 2,  $t(49) = 3.89$ ,  $p < .001$ , Cohen’s  $d = .55$ . As considered earlier in our analysis of the FAR, the size of these differences in FAR for expected-strong versus expected-weak blocks did not differ.

Next, we examined whether the degree of shifting on each test was consistent across individuals. We first explored this reliability across and then within our three experimental conditions. There was noticeable variation in shifting within each task. Using our FAR shift measure and collapsing across the test blocking groups, the tendency to shift one’s criterion was consistent across tests,  $r(157) = .34$ ,  $p < .001$ . Regarding the reliability of criterion shifts within each condition, Fig. 3 displays the relevant scatter plots. Each panel displays the correlations between shifting on Tests 1 and 2 for each group. For the 40-item block condition (Fig. 3a) there was no significant reliability in shifting from Test 1 to Test 2,  $r(53) = .22$ ,  $p = .10$ . For the 10-item block condition (Fig. 3b), participants displayed reliable criterion shifts,  $r(52) = .28$ ,  $p = .039$ . We also found reliable criterion shifts in the unblocked condition (Fig. 3c),  $r(48) = .51$ ,  $p < .001$ .<sup>5</sup> We also calculated the

<sup>5</sup> We also examined the relationship between the FAR measures in both experiments using Spearman’s rho. In Experiment 1, the correlation for the full data set was  $r_s(157) = .31$ ,  $p < .001$ . The analysis by block indicated significantly correlated shifts for the unblocked,  $r_s(48) = .39$ ,  $p = .005$ , and 10-item block conditions,  $r_s(52) = .37$ ,  $p = .006$ . Again, the correlation for the 40-item block condition was not significant,  $r_s(53) = .22$ ,  $p = .11$ . Hence, we obtained the same overall pattern of results whereby the reliability is numerically largest for the unblocked condition and then drops in the 10-item and 40-item blocks. In Experiment 2, the correlation between the strength-based criterion shifts was significant,  $r_s(109) = .33$ ,  $p < .001$ , as was the correlation for the probability-based shifts,  $r_s(109) = .30$ ,  $p = .001$ .



**Fig. 3** Correlations between criterion shifts on Test 1 and Test 2 for the (a) 40-item block, (b) 10-item block, and (c) unblocked test conditions in Experiment 1

reliability of criterion shifts in each condition using the measures  $C$  and  $\ln(\beta)$  derived from SDT. These results are displayed in the top portion of Table 2 and converge on our interpretations based on the FAR shift measure.

To explore whether the reliability of criterion shifts was significantly different among the three test conditions, we performed a Fisher's  $r$  to  $Z$  transformation to compare each correlation. The correlation coefficients between the 40-item and 10-item block conditions were not significantly different from each other,  $Z = .33, p = .75$ . The unblocked and 10-item block conditions were not different,  $Z = 1.31, p = .19$ , nor were the unblocked and 40-item block conditions,  $Z = 1.70, p = .09$ .

To summarize, participants completed two study-test cycles of a strength-based recognition procedure in which we manipulated the test format. On average, there was a 9 % difference in the FAR, indicative of large criterion shifts. Each test condition produced significant criterion shifts, and the size of those shifts was unaffected by blocking test items. The lack of a blocking effect indicates that having people use the three-key response procedure is sufficient for them to process the expected-strength cues appropriately, rendering the blocking manipulation much less important. Regarding the reliability of shifts between tests, we found a significant correlation indicating that participants demonstrated similar shifting abilities across tests. Furthermore, the extent to which participants demonstrated reliable shifts depended somewhat on the test format, with the most reliable shifts occurring in the unblocked condition and then the 10-item and 40-item blocks, respectively.

Blocking test items had no significant effect on the size of the criterion shifts within each group, but blocking nominally

**Table 2** Reliability of criterion shift measures in Experiments 1 and 2 as zero-order correlations

Condition	FAR Shift	$\ln(\beta)$ Shift	$C$ Shift
Experiment 1 (Test 1 & 2)			
40-item blocks	.22*	.45	.33
10-item blocks	.28	.30	.36
Unblocked	.51	.35	.43
Overall	.34	.38	.37
Experiment 2 (Day 1 & 2)			
Probability based	.38	.42	.51
Strength based	.34	.35	.44

*Note.* For each of the three criterion measures, *shift* refers to the difference on that measure [FAR,  $\ln(\beta)$ , or  $C$ ] between strength-based (strong vs. weak) or probability-based (high vs. low) test conditions. On strength-based tests, shifts were calculated by subtracting the criterion measure for weak items from the same criterion measure for strong items. On probability-based tests, shifts were calculated by subtracting the criterion measure of low-probability items from the same criterion measure of high-probability items. Correlations of shifts were calculated for Test 1 and Test 2 (Experiment 1) or for Day 1 and Day 2 (Experiment 2). For all correlations reported here,  $p < .05$ , except where noted by an asterisk.

affected the reliability of shifts. Although the unblocked test condition exhibited the smallest overall group shift, it also produced the largest amount of shifting reliability across tests ( $r = .51$ ). The size of this correlation is very similar to that found by Aminoff et al. (2012) using a probability-based test across two study-test blocks. Our 40-item block condition produced the largest numerical group shifts but also demonstrated the lowest reliability in criterion shifting from test to test. The 10-item block condition was in between the other two conditions both in terms of the size of the group shift and shifting reliability.

## Experiment 2

To address the predictive relationship between criterion shifting in two different recognition tasks, in Experiment 2 we asked participants to complete recognition tests in which we manipulated either the base rate (probability based) or expected memory strength (strength based) of the test items. The probability-based tests were very similar to those used by Aminoff et al. (2012). For the strength-based test, we used the unblocked test condition from Experiment 1 because it produced the highest correlation across study-test cycles. Participants completed each type of test in their first testing session and again 2 days later.

## Method

### Participants

Participants were 116 undergraduate psychology students from Louisiana State University who participated in this experiment to fulfill a partial course requirement or for extra credit.

### Materials and procedure

We selected 360 additional words with the same stimuli characteristics as those in Experiment 1 from the MRC Psycholinguistic Database. These words were randomly selected to be stimuli in one of two probability-based recognition tasks created using E-Prime software.

Two probability-based recognition memory programs were created, each comprised of 180 words, similar to Aminoff et al.'s (2012) study. For each program, 90 words were presented individually at encoding for 2,000 ms followed by a 500 ms ISI. The test stimuli included these targets and the remaining 90 unstudied items served as lures. Different versions of the software counterbalanced which items served as targets versus lures. These test items were split into multiple blocks of probability conditions, which reflected the likelihood of a test item being previously studied. The high-



probability blocks contained, on average, 70 % targets and 30 % lures. The low-probability blocks contained 30 % targets and 70 % lures. The blocks were between six and nine items of a given target probability. High-probability blocks were displayed in blue font color and low-probability blocks were displayed in orange. We varied whether participants started with a high or low-probability test block. The test consisted of 24 blocks, of which there were an equal number of high-probability and low-probability blocks. Each program had a different and fixed testing sequence, whereby each participant completed the same randomized sequence of test blocks of alternating probabilities for each program. The software randomly selected test items from the appropriate set of either high-probability or low-probability items for each block. If participants took longer than 2 seconds to respond to any word, a message preceded the next test item, prompting them to balance speed and accuracy. A short practice phase was included to familiarize participants with the study-test procedure.

The software program informed participants they would study words for a later test. After the encoding phase, participants were informed they would receive color cues at test to help them make their memory decisions. Specifically, they were told that words presented in blue had a 70 % chance of being studied, and words shown in orange had a 30 % chance of being studied. Additionally, participants were instructed to press the ‘/’ key for “old” decisions and to press the ‘z’ key for “new” decisions.

For the strength-based test context, we used the same two programs from Experiment 1, and all participants completed the unblocked test condition. Participants completed two experimental sessions, which were separated by a 48-hour delay. During the first session, participants completed one probability-based and one strength-based program, and we counterbalanced which one was completed first. In the second session, participants completed these tasks again (with different stimuli) and started the session with a different type of test than they did in the first session. Each testing session lasted about 30 minutes.

## Results and discussion

Three participants were excluded because they failed to complete at least one of the four recognition tests. The data were screened for outliers as in Experiment 1, which led to the exclusion of two additional participants for having a FAR shift measure more than 3 standard deviations from the mean, leaving a sample size of 111. The relevant group-level recognition data are presented in Table 3. First, we present analyses of HR, FAR, and recognition discriminability. The strength-based programs were analyzed using a 2 (Test session: Day 1 or Day 2)  $\times$  2 (Strength: strong or weak) repeated-measures ANOVA.

For the HR, there was a main effect of strength,  $F(1, 110) = 139.28$ ,  $MSE = .012$ ,  $p < .001$ ,  $\eta_p^2 = .559$ , and a significant effect of test session,  $F(1, 110) = 6.56$ ,  $MSE = .016$ ,  $p = .012$ ,  $\eta_p^2 = .056$ . Both of these effects were qualified by a significant interaction,  $F(1, 110) = 15.22$ ,  $MSE = .006$ ,  $p < .001$ ,  $\eta_p^2 = .122$ . Post hoc tests revealed this interaction was due to a slightly higher HR on Day 1 versus Day 2 for strong items,  $t(110) = 4.51$ ,  $p < .001$ , Cohen’s  $d = .43$ , and equivalent HRs on both sessions for weak items,  $t(110) = 0.09$ ,  $p = .93$ . For the FAR, there was a main effect of strength,  $F(1, 110) = 97.26$ ,  $MSE = .017$ ,  $p < .001$ ,  $\eta_p^2 = .469$ , indicating there were more false alarms to weak than strong items. There was also a main effect of test session,  $F(1, 110) = 16.96$ ,  $MSE = .030$ ,  $p < .001$ ,  $\eta_p^2 = .134$ , indicating there were more false alarms in the second session. The interaction was not significant,  $F < 1$ .

Discriminability was again measured using  $d'$  and was better for the strong test items,  $F(1, 110) = 431.96$ ,  $MSE = .156$ ,  $p < .001$ ,  $\eta_p^2 = .797$ . A main effect of test session indicated that memory was better on Day 1,  $F(1, 110) = 51.15$ ,  $MSE = .186$ ,  $p < .001$ ,  $\eta_p^2 = .317$ . Additionally, the interaction between strength and test session was significant,  $F(1, 110) = 8.19$ ,  $MSE = .153$ ,  $p = .005$ ,  $\eta_p^2 = .069$ . Post hoc tests revealed this interaction was due to a larger drop in discriminability from Day 1 to Day 2 for strong items,  $t(110) = 6.45$ ,  $p < .001$ , Cohen’s  $d = .61$ , than for weak items,  $t(110) = 3.91$ ,  $p < .001$ , Cohen’s  $d = .37$ .

For the probability-based programs, hits and false alarms were analyzed using a 2 (Test session: Day 1 or Day 2)  $\times$  2 (Probability: high or low) repeated-measures ANOVA. A main effect of probability indicated that the HR was higher for high probability items,  $F(1, 110) = 232.25$ ,  $MSE = .020$ ,  $p < .001$ ,  $\eta_p^2 = .679$ . Additionally, a main effect of test session indicated that the HR was slightly higher on Day 1 versus Day 2,  $F(1, 110) = 21.30$ ,  $MSE = .016$ ,  $p < .001$ ,  $\eta_p^2 = .162$ . The interaction was not significant,  $F(1, 110) = 2.29$ ,  $p = .13$ . For the FAR, there was no effect of test session,  $F(1, 110) = 1.50$ ,  $p = .22$ , but there was an effect of probability,  $F(1, 110) = 134.52$ ,  $MSE = .021$ ,  $p < .001$ ,  $\eta_p^2 = .55$ . The interaction was also significant,  $F(1, 110) = 7.89$ ,  $MSE = .009$ ,  $p = .006$ ,  $\eta_p^2 = .067$ . Post hoc tests indicated that this was due to the difference between high and low probability false alarms being smaller on Day 1,  $t(110) = 8.13$ ,  $p < .001$ , Cohen’s  $d = .77$ , than on Day 2,  $t(110) = 11.14$ ,  $p < .001$ , Cohen’s  $d = 1.06$ .

Discriminability for the probability-based tests was overall better on Day 1 than on Day 2,  $F(1, 110) = 13.80$ ,  $MSE = .326$ ,  $p < .001$ ,  $\eta_p^2 = .111$ . There was also a main effect of probability,  $F(1, 110) = 15.28$ ,  $MSE = .119$ ,  $p < .001$ ,  $\eta_p^2 = .122$ . Both of these effects were qualified by a significant interaction,  $F(1, 110) = 4.63$ ,  $MSE = .089$ ,  $p = .034$ ,  $\eta_p^2 = .040$ . Post hoc test showed that discriminability was better for high- versus low-probability items on Day 1,  $t(110) = 4.64$ ,  $p < .001$ , Cohen’s  $d = .44$ , and equivalent for high- and low-probability items on Day 2,  $t(110) = 1.46$ ,  $p = .15$ .

**Table 3** Group recognition performance in Experiment 2

	Probability-based tests				Strength-based tests			
	Day 1		Day 2		Day 1		Day 2	
	High	Low	High	Low	Weak	Strong	Weak	Strong
HR	.78 (.01)	.59 (.02)	.74 (.01)	.52 (.02)	.65 (.01)	.80 (.01)	.65 (.02)	.74 (.02)
FAR	.39 (.02)	.26 (.02)	.44 (.02)	.25 (.01)	.43 (.02)	.30 (.02)	.50 (.02)	.37 (.02)
$d'$	1.17 (.06)	.98 (.06)	.91 (.06)	.84 (.06)	.63 (.04)	1.52 (.05)	.45 (.04)	1.12 (.07)
Shift	.13 (.02)		.18 (.02)		.12 (.01)		.12 (.02)	

Note. Standard errors of the mean in parentheses

### Criterion shifting at the group level and at the individual level

Criterion shifts were calculated in the same manner as in Experiment 1. Participants exhibited moderate-to-large criterion shifts on the strength-based tests on both Day 1,  $t(110) = 8.47$ ,  $p < .001$ , Cohen's  $d = .80$ , and Day 2,  $t(110) = 7.72$ ,  $p < .001$ , Cohen's  $d = .73$ . Large and significant group shifts were also observed on both days for the probability-based tests: Day 1,  $t(110) = 7.15$ ,  $p < .001$ , Cohen's  $d = .77$ , and Day 2,  $t(110) = 11.06$ ,  $p < .001$ , Cohen's  $d = 1.06$ .

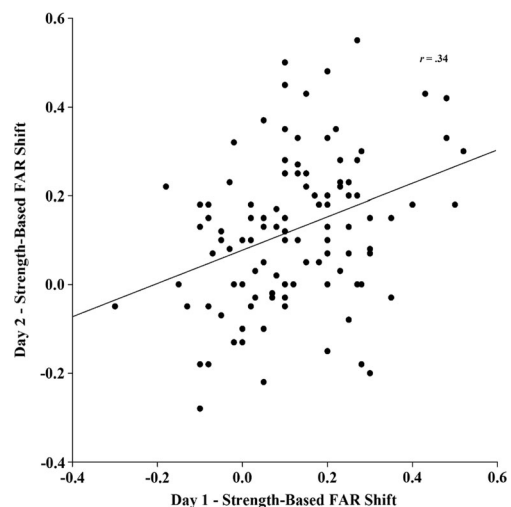
We next examined whether the degree of shifting on each type of test was consistent across individuals using our FAR shift measure. Participants' tendency to shift their criterion on the strength-based tests was consistent from Day 1 to Day 2,  $r(109) = .34$ ,  $p < .001$ . For the probability-based tests, we also observed reliable shifts from Day 1 to Day 2,  $r(109) = .38$ ,  $p < .001$ . Figures 4 and 5, respectively, display the scatterplots for these relationships. A similar conclusion is drawn when examining the SDT measures of shifting in the bottom portion of Table 2.

To address our primary research question, we examined whether criterion shifts were correlated on the same day for the two different recognition tests. Using our FAR shift measure, we found no correlation in the amount of shifting on the strength-based versus the probability-based tests on either Day 1,  $r(109) = .08$ ,  $p = .43$ , or on Day 2,  $r(109) = -.15$ ,  $p = .12$ . These conclusions did not change when similar correlations between strength-based and probability-based shifts were computed for the criterion measures  $C$ ,  $r = .08$  for Day 1 and  $r = -.06$  for Day 2, or  $\ln(\beta)$ ,  $r = -.06$  for Day 1 and  $r = .09$  on Day 2. To ensure that our interpretation of the correlations between our shifting measures was not limited by the observed differences in discriminability (Macmillan & Creelman, 2005), we calculated semipartial correlations between our shifting measures that statistically controlled for the relevant differences in discriminability measures.<sup>6</sup>

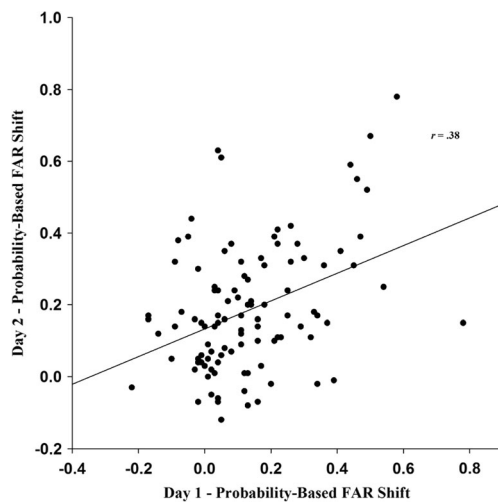
<sup>6</sup> We thank Amy Criss for bringing this to our attention during the review process and for suggesting additional analyses.

For the strength-based tests, the semipartial correlation between the FAR shift measures on Day 1 and Day 2, while controlling for the strong and weak  $d'$  measures on both days, was actually larger,  $r(105) = .43$ ,  $p < .001$ , than the zero-order correlation ( $r = .34$ ). We performed a similar analysis for the probability-based data. The semipartial correlation between the FAR shift measures, while controlling for the high- and low-probability  $d'$  measures on both days, was slightly smaller,  $r(105) = .32$ ,  $p < .001$ , but still comparable the zero-order correlation ( $r = .38$ ). Additionally, we also examined the shifts across tasks on Day 1 and Day 2 by examining the semipartial correlations that control for the four discriminability measures on a given day. As with the zero-order correlations, there was no relationship in shifting across tasks on Day 1,  $r(105) = .10$ ,  $p = .25$ , or Day 2,  $r(105) = .001$ ,  $p = .99$ . Thus, even when taking into account discriminability differences, we still find the same pattern of results as when examining the zero-order correlations.

In summary, participants exhibited large criterion shifts on the probability-based and strength-based tests on both days. Additionally, there was a large amount of variation in the



**Fig. 4** Correlation between strength-based criterion shift measures on Day 1 and Day 2 in Experiment 2



**Fig. 5** Correlation between probability-based criterion shift measures on Day 1 and Day 2 in Experiment 2

amount an individual was willing to shift his or her criterion. Participants' criterion shifts on the strength-based tests were correlated across days, as were the shifts on the probability-based tests. However, the size of criterion shifts on the strength-based test did not predict the size of shifts on the probability-based test.

## General discussion

The purpose of this study was to investigate whether criterion shifting in recognition memory might be characterized as a general ability or cognitive trait, as suggested by Aminoff et al. (2012). To this end, in Experiment 1 we first explored whether participants would exhibit reliable strength-based criterion shifts within a single testing session using a three-key test procedure we adapted from Starns and Olchowski (2015). We varied whether test items were presented in like-strength blocks of varying sizes and found large group-level criterion shifts in blocked and unblocked test conditions. The test–retest correlations of criterion shifts in each condition were not significantly different from each other, but were numerically largest in the unblocked condition. We replicated the reliability of strength-based criterion shifts in Experiment 2 and also found reliable shifts on the probability-based tests. Critically, however, these two types of criterion shifts were not correlated with each other.

To our knowledge, this is the first study to examine the reliability of within-list strength-based criterion shifts over testing sessions. We did not find that blocking (or lack thereof) significantly affected the reliability of criterion shifts, however this is likely because of a lack of statistical power, particularly in comparing the two blocked conditions to the unblocked condition. Indeed, in Experiment 1 the correlation between criterion shifts was about twice as large in the unblocked condition ( $r = .51$ ) compared to the 40-item ( $r = .22$ ) or 10-item

( $r = .28$ ) conditions. One possible reason for these differences in reliability might reflect the differences in the number of potential shifts that can be observed in each condition. In the 40-item and 10-item conditions, there are three and 15 chances, respectively, for participants to shift their criterion. In the unblocked condition, participants (potentially) have a chance to shift their criterion on every trial, depending on what item the software program selects. It could be the case that participants do not shift their criterion at every possible opportunity, but when they do shift, they do so with some regularity. Hence, giving participants more potential opportunities to shift may afford a better chance at gauging that behavior and detecting reliable shifts.

In contrast to previous work using a two-key procedure in strength-based recognition memory manipulations, we found that blocking test items with a three-key procedure did not affect the size of the group-level criterion shifts. In the Hicks and Starns (2014) two-key experiments, larger shifts were produced and more likely to be replicated when test items were presented in larger (i.e. 40-item or 20-item) blocks of like-strength test items. Hicks and Starns also tried to induce criterion shifts using a 10-item block and unblocked test condition on three separate occasions each. They found only one significant shift using the 10-item condition and one significant shift using the unblocked test. In our first experiment using the three-key methodology, we found significant shifts in the 10-item and unblocked condition on both attempts. Additionally, the shifts we found in 10-item and unblocked test conditions were about 2 and 4 times as large, respectively, as what Hicks and Starns found in those conditions. Thus, we replicated the Starns and Olchowski (2015) findings and extended them to blocked-test conditions. It appears that the benefit of blocking test items only facilitates strength-based criterion shifts when participants are not forced to process the strength cues otherwise (e.g., Hicks & Starns, 2014; Starns & Olchowski 2015).

A common finding across studies is that a sole manipulation of memory strength at encoding is insufficient for producing within-list criterion shifts at test. Verde and Rotello (2007) made a similar statement, saying, “One has to doubt the ability of memory strength itself to cue bias shifts under reasonable circumstances” (p. 260). Indeed, strength-based shifts depend on explicit cues at test coupled with either a test composition (Hicks & Starns, 2014) or response format (Starns & Olchowski 2015) that aids participants in appreciating the difference between strong and weak items. Verde and Rotello were able to induce shifts without explicit strength cues by using feedback. However, as the authors note, it is unclear in their study what the feedback manipulation is doing, exactly. Feedback might enhance participants' awareness of the strength differences, then allowing them to use expected memory strength as a cue, or, alternatively, the feedback may lead participants to adjust their metacognitive approach to

subsequent recognition decisions. Regardless, feedback-induced shifts in within-list, strength-based paradigms are not a consistent finding (see Hicks & Starns, 2014).

Other researchers have found within-list shifts using manipulations of memory strength at encoding (Bruno et al., 2009; Morrell, Gaitan, & Wixted, 2002; Singer & Wixted, 2006), though some of these shifts may not be aptly characterized as strength-based shifts per se. For example, in the Singer and Wixted (2006) study, categorized lists were studied at different delays; hence, memory strength was technically confounded by category membership (see Hicks & Starns, 2014, for a discussion of this issue). Thus, it is ambiguous whether shifts here result from using category membership as a cue for expected memory strength or whether participants are making more of a heuristic decision based on the category itself (i.e., the category was a cue to strength, but also was what helped to define strong versus weak). Analogously, the Bruno et al. (2009) study confounded memory strength with semantic (Experiment 1) and perceptual (Experiment 2) categories at encoding, hence the within-list shifts they observed are also not necessarily strength based. The larger point is that within-list criterion shifts (strength based or otherwise) do not occur without some sort of external cue—those cues can directly signal expected memory strength like the color cues used by Hicks and Starns (2014) or like the perceptual and semantic cues given by Bruno et al. (2009).

Our finding that probability-based criterion shifts were correlated across 2 days of testing replicates the Aminoff et al. (2012) study and extends their findings by demonstrating that, even across days, individuals show a consistency in their willingness to shift their decision criterion in response to probability information. This finding suggests that shifting behavior in this task is quite reliable across stimulus types (Aminoff et al., 2012) and across time. The correlation we found in Experiment 2 ( $r = .38$ ) is roughly two thirds of that reported by Aminoff et al. (2012), although they tested participants in a single testing session rather than across days, as we did. Similarly, the reliability of our strength-based shifts dropped when tests were given on different days.

Although we found evidence in both experiments that people can reliably shift their criterion within a given type of recognition test, our data do not support Aminoff et al.'s (2012) claim that people's shifting generalizes. Some participants exhibited large and moderate shifts in one type of task, but not another. Our finding that criterion shifting does not appear to be trait-like provides an interesting contrast to the work showing that criterion setting is highly consistent across different stimuli and test contexts. Kantner and Lindsay (2012, 2014) have shown in many different test contexts that where a recognition criterion is set it is highly consistent across individuals. This was true even when the criterion measurement on two tasks were days apart and presented in very different settings and circumstances (Kantner & Lindsay, 2014). Our

own data support this conclusion. For example, in Experiment 2, the FAR in weak conditions of the strength-based tests on Day 1 correlated well with the FAR in high- and low-probability conditions on Day 2 ( $r = .48$  and  $r = .57$ , respectively). The correlations among all of our FAR measures in Experiment 2 ranged from .38 at the lowest to .74 at the highest, and similar patterns were true of our data with other potential criterion measures, such as  $C$  or  $\ln(\beta)$  from signal-detection theory. These patterns also held for Experiment 1. One potential avenue for future research is to explore whether an individual's relatively consistent criterion *setting* relates to their criterion *shifting*, notably, under what circumstances these separate behaviors might or might not be related. It could be that a person's willingness to shift his or her criterion is related to where it is initially set. As noted by Aminoff et al. (2012), the need for a person to shift his or her criterion becomes less necessary when that person's criterion setting starts off relatively optimal or neutral. Perhaps the individuals most likely to exhibit large criterion shifts are those who start off with a sufficiently extreme liberal or conservative setting. In any event, the relationship between criterion setting and criterion shifting remains an open question.

**Author Note** Our pilot data were presented at the 53rd Annual Meeting of the Psychonomic Society, Toronto, ON, Canada, November 2013. We wish to thank Katherine Cantwell, Karyn Warner, Andrea Robinson, Benjamin Perrodin, Jessica West, Beth Turansky, and Kierstin Marshall for their indispensable help with data collection.

## References

- Aminoff, E. M., Clewett, D., Freeman, S., Frithsen, A., Tipper, C., Johnson, A., & Miller, M. B. (2012). Individual differences in shifting decision criterion: A recognition memory study. *Memory & Cognition*, *40*(7), 1016–1030. doi:10.3758/s13421-012-0204-6
- Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review*, *116*(1), 84–115. doi:10.1037/a0014351
- Bohil, C. J., & Maddox, W. T. (2003). A test of the optimal classifier's independence assumption in perceptual categorization. *Perception & Psychophysics*, *65*(3), 478–493. doi:10.3758/BF03194577
- Bruno, D., Higham, P. A., & Perfect, T. J. (2009). Global subjective memorability and the strength-based mirror effect in recognition memory. *Memory & Cognition*, *37*(6), 807–818. doi:10.3758/MC.37.6.807
- Cox, J. C., & Dobbins, I. G. (2011). The striking similarities between standard, distractor-free, and target-free recognition. *Memory & Cognition*, *39*(6), 925–940. doi:10.3758/s13421-011-0090-3
- Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory and Language*, *55*(4), 461–478. doi: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*(4), 297–319. doi: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(2), 484–499. doi:10.1037/a0018435
- Dobbins, I. G., & Kroll, N. A. (2005). Distinctiveness and the recognition mirror effect: Evidence for an item-based criterion placement heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(6), 1186–1198. doi:10.1037/0278-7393.31.6.1186
- 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(5), 1075–1095. doi:10.1037/0278-7393.21.5.1075
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Oxford: Wiley.
- Han, S., & Dobbins, I. G. (2008). Examining recognition criterion rigidity during testing using a biased-feedback technique: Evidence for adaptive criterion learning. *Memory & Cognition*, 36(4), 703–715. doi:10.3758/MC.36.4.703
- 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(5), 544–553.
- Heit, E., Brockdorff, N., & Lamberts, K. (2003). Adaptive changes of response criterion in recognition memory. *Psychonomic Bulletin & Review*, 10(3), 718–723. doi:10.3758/BF03196537
- Hicks, J. L., & Starns, J. J. (2014). Strength cues and blocking at test promote reliable within-list criterion shifts in recognition memory. *Memory & Cognition*, 42(5), 742–754. doi:10.3758/s13421-014-0397-y
- 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(4), 679–688. doi:10.3758/BF03193306
- Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition*, 40(8), 1163–1177. doi:10.3758/s13421-012-0226-0
- Kantner, J., & Lindsay, D. S. (2014). Cross-situational consistency in recognition memory response bias. *Psychonomic Bulletin & Review*, 21(5), 1272–1280. doi:10.3758/s13423-014-0608-3
- Koop, G. J., & Criss, A. H. (2015). The response dynamics of recognition memory: Sensitivity and bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. doi:10.1037/xlm0000202. Advance online publication.
- 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(2), 509–516. doi:10.3758/s13423-014-0703-5
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah: Erlbaum.
- Morrell, H. R., 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(6), 1095–1110. doi:10.1037/0278-7393.28.6.1095
- 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(2), 305–320. doi:10.1037/0278-7393.33.2.305
- Selmeczy, D., & Dobbins, I. G. (2013). Metacognitive awareness and adaptive recognition biases. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(3), 678–690. doi:10.1037/a0029469
- Singer, M. (2009). Strength-based criterion shifts in recognition memory. *Memory & Cognition*, 37(7), 976–984. doi:10.3758/MC.37.7.976
- Singer, M., & Wixted, J. T. (2006). Effect of delay on recognition decisions: Evidence for a criterion shift. *Memory & Cognition*, 34(1), 125–137. doi:10.3758/BF03193392
- Starns, J. J., & Olchowski, J. E. (2015). Shifting the criterion is not the difficult part of trial-by-trial criterion shifts in recognition memory. *Memory & Cognition*, 43(1), 49–59. doi:10.3758/s13421-014-0433-y
- 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(5), 1137–1151. doi:10.1037/a0028151
- 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(6), 1379–1396. doi:10.1037/0278-7393.24.6.1379
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson.
- Titus, T. G. (1973). Continuous feedback in recognition memory. *Perceptual and Motor Skills*, 37(3), 771–776. doi:10.2466/pms.1973.37.3.771
- Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition*, 35(2), 254–262. doi:10.3758/BF03193446
- White, C. N., & Poldrack, R. A. (2014). Decomposing bias in different types of simple decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(2), 385–398.
- Wickelgren, W. A. (1968). Unidimensional strength theory and component analysis of noise in absolute and comparative judgments. *Journal of Mathematical Psychology*, 5, 102–122.