The striking similarities between standard, distractor-free, and target-free recognition

Justin C. Cox · Ian G. Dobbins

Published online: 8 April 2011
© Psychonomic Society, Inc. 2011

Abstract It is often assumed that observers seek to maximize correct responding during recognition testing by actively adjusting a decision criterion. However, early research by Wallace (Journal of Experimental Psychology: Human Learning and Memory 4:441–452, 1978) suggested that recognition rates for studied items remained similar, regardless of whether or not the tests contained distractor items. We extended these findings across three experiments, addressing whether detection rates or observer confidence changed when participants were presented standard tests (targets and distractors) versus “pure-list” tests (lists composed entirely of targets or distractors). Even when observers were made aware of the composition of the pure-list test, the endorsement rates and confidence patterns remained largely similar to those observed during standard testing, suggesting that observers are typically not striving to maximize the likelihood of success across the test. We discuss the implications for decision models that assume a likelihood ratio versus a strength decision axis, as well as the implications for prior findings demonstrating large criterion shifts using target probability manipulations.

Keywords Recognition memory · Signal detection · Criterion · Distractor-free recognition · Target-free recognition

Electronic supplementary material The online version of this article (doi:10.3758/s13421-011-0090-3) contains supplementary material, which is available to authorized users.

J. C. Cox (✉) · I. G. Dobbins
Department of Psychology, Washington University,
Campus Box 1125, 1 Brookings Drive,
St. Louis, MO 63130, USA
e-mail: j.cox@wustl.edu

The standards by which observers judge the veracity of memories are referred to as *decision criteria*, and it is often assumed that observers dynamically regulate these criteria in order to maximize the likelihood of favored relative to less favored outcomes. For example, in the ideal case, observers would use a much more stringent criterion for judging an individual in an eyewitness lineup as recognized, as compared to judging a fellow conference attendee as recognized, given the relative costs of errors under the two circumstances. Given the obvious advantage that adaptive memory criteria would confer, recent research has refocused on examining the degree of control that observers are actually able to exercise over recognition memory criteria.

Recent research, usually conducted under a basic signal detection theory framework (see Fig. 1), has presented a mixed picture regarding the degree of control that participants exercise over criterion placement.

For example, observers appear to adjust the criterion in a strategic fashion in response to manipulations of the nature of the lure materials (semantically similar vs. dissimilar to probes), the a priori familiarity of the materials, and the presence of feedback (Benjamin & Bawa, 2004; Curran, DeBuse, & Leynes, 2007; Han & Dobbins, 2008, 2009; Rhodes & Jacoby, 2007; Verde & Rotello, 2007), although it remains unclear whether these shifts are achieved through a common mechanism. In contrast, there are several recent reports of surprising criterion rigidity in the face of manipulations of test materials that should induce an opportunistic criterion shift if observers are seeking to maximize correct responding (Estes & Maddox, 1995; Morrell, Gaitan, & Wixted, 2002; Stretch & Wixted, 1998b). For example, Stretch and Wixted (1998b) were unable to shift the decision criterion, even when participants were made aware that one class of targets would be differentially strengthened. Specifically, during study, words

Fig. 1 Example of the equal-variance one-dimensional signal detection model. Test items (targets and lures) form distributions of mnemonic evidence. In order to discriminate between old and new items, an observer must set a value (criterion) by which to parse this evidence as coming from an old or a new item. The “c” indicates the location of the old/new criterion; items with less mnemonic evidence than this value are deemed “new,” while items with greater evidence are deemed “old.” The dashed lines denote confidence criteria. Items that fall above or below these criteria are judged as “old” or “new” with varying degrees of confidence. L, M, and H refer to low, medium, and high confidence, respectively

were presented in either red or green font, and participants were told that words in red font would be repeated several times, whereas words in green font would be presented only once each. They were further instructed that both old and new items on an upcoming recognition test would be presented in either red or green font, with old items appearing in the same font as they had been presented in at study (Stretch & Wixted, 1998b). If participants dynamically regulated criteria as a function of expected evidence strengths, they should have used different criteria for red and green words at test. The authors found that while the strength manipulation affected hit rates, the false alarm rates for red and green items were equivalent, suggesting that participants did not dynamically update the decision criterion, even when provided with an unambiguous cue that would be useful. Stretch and Wixted concluded that perhaps participants are unwilling to shift the decision criterion on an item-by-item basis, due to the cognitive effort required to do so (see also Morrell et al., 2002).

Whereas Stretch and Wixted (1998b) intermixed differentially strengthened classes of items within a test list, Verde and Rotello (2007) manipulated the strengths of targets across two halves of a test list. If observers could utilize memory strength itself as a cue to update the decision criterion, they should adopt different criteria for each half of the test when the average item strength changed across test halves. In the first four experiments reported, Verde and

Rotello manipulated item strength by either the duration of presentation (Exp. 1) or the frequency of presentation (Exps. 2, 3, 4). Following the study phase, participants in each experiment completed two seamless recognition blocks: One test block examined memory for strong targets, while the other examined memory for weak targets. Although Verde and Rotello found that they could manipulate initial criterion placement, they were unable to encourage a criterion shift by simply manipulating the distributions of target strength within the tests (i.e., only strong targets vs. only weak targets). The researchers concluded that observers do not typically utilize memory strength cues to modulate and update decision criterion placement, and instead rely upon cues not directly tied to memory representations, such as performance feedback.

Overall, the work reviewed above suggested that observers are unable or unwilling to use either changes in the distribution of strong versus weak memoranda within the test or reliable associations between average strength levels and nonepisodic characteristics of probes (e.g., strength–color pairings) as a direct basis for shifting decision criteria to maximize correct responding. Remarkably, a perhaps even stronger test of this hypothesis was illustrated in several studies that focused on whether the use of distractors during recognition testing was necessary, or instead was dictated more by convention (Strong & Strong, 1916; Underwood, 1972; Wallace, 1978, 1982; Wallace, Sawyer, & Robertson, 1978). Wallace (1980) argued that recognition memory simply referred to the knowledge that “a given old object was previously experienced” (p. 696) and that the inclusion of distractors on a memory test was instead “appropriate for examining recognition-discrimination” (p. 696) rather than the processes underlying recognition memory. While he noted that recognition discrimination was an interesting process in its own right, valid tests of recognition memory, he argued, should not include novel alongside familiar items. More specifically, he argued that the intuition that participants would strategically report all or most items as remembered if they knew that the test list contained no lures was inherently false.

Supporting his claims, Wallace (1982; Wallace et al., 1978) demonstrated remarkable similarities in the endorsement rates of studied items when comparing standard and distractor-free recognition tests (a form of recognition test in which participants are only tested on previously seen items). This similarity was present regardless of whether participants were aware of the manipulation (Wallace et al., 1978) or not (Wallace, 1982). Wallace (1982) argued that if observers are driven to maximize correct responding, they should claim to recognize every item on a distractor-free test—essentially, adopting an extremely liberal response criterion and drastically improving detection. Instead, his results, like those of Verde and Rotello (2007), imply that

observers do not or cannot directly use the distribution of evidence within the test list as a signal to shift their response criteria in an optimal direction—even when they are informed about how the evidence will be distributed.

Despite the surprising nature of these findings, distractor-free recognition has not garnered much attention since Wallace's (1978) experiments over 30 years ago. McKelvie (1993) utilized a distractor-free test to investigate memory for faces. The results replicated Wallace's (1978) findings, showing that recognition rates for distractor-free testing were approximately equal to standard recognition test hit rates. Ley and Long (1987, 1988) also employed the distractor-free recognition method in a series of studies, and they once again demonstrated the surprising rigidity that participants adopted with regard to updating their response criteria, leading to the fairly strong argument that "studies of *pure* recognition memory . . . should be limited to distractor-free tests" (Ley & Long, 1988, p. 409).

One plausible explanation for Wallace's (1978) findings is that observers may enter a recognition test with the expectation that they should recognize some amount of studied items; it is not entirely unreasonable to believe that participants enter a memory experiment motivated to detect studied items. That is, when participating in tests described as "memory" tests, individuals may focus predominantly on their ability to detect a sizable portion of items as old. Under the distractor-free manipulation, this goal orientation is not threatened (and in fact is easily achieved), so they may not feel the need to adjust their approach to the classifications; thus, they do not update their decision criteria. Such an account begs the question of whether participants would be similarly unaffected if targets were wholly removed from the test list (a "target-free" test), since in this situation they would be subjectively recognizing very few of the items. Thus, under a "motivated-to-recognize" account, when one is only being tested with studied materials, there will be frequent instances satisfying the subjective goal of remembering. In contrast, if observers find themselves rarely endorsing items as studied, which would be the case in a target-free testing situation, this might serve as a signal that their memory is somehow failing them, which in turn may lead to a liberal shift in the criterion. In contrast, if target-free endorsement rates looked equivalent to those under standard testing, this would lend further credence to the idea that observers rely on relatively stable evidence-to-decision mappings that are not linked to the relative densities of the old or new items present at test.

Here we revisit the distractor-free and target-free recognition paradigms, focusing on the participants' awareness of the manipulation, the confidence of their reports, and whether the tests are free of distractors or free of targets. Critically, although the prior work demonstrated comparable endorsement rates of studied materials in isolation or when they were

intermixed with novel distractors, it is not clear whether participants would express the same level of confidence in their reports in these two situations. More specifically, since confidence reports are also held to reflect the placement of separate and perhaps independently manipulable decision criteria under current signal detection models (Hautus, Macmillan, & Rotello, 2008; Macmillan & Creelman, 2005; Morrell et al., 2002; Stretch & Wixted, 1998a), it could well be the case that these criteria shift even though the criterion governing the categorical judgment between old and new items remained largely unaffected by the manipulation. Furthermore, research from the metacognitive and recognition literatures suggests that memory confidence is not always tightly intertwined with memory evidence (Brewer, Sampaio, & Barlow, 2005; Busey, Tunnicliff, Loftus, & Loftus, 2000; Chandler, 1994; Cleary, Winfield, & Kostic, 2007; Dobbins, Kroll, & Liu, 1998; Sporer, Penrod, Read, & Cutler, 1995), suggesting that confidence reports may be sensitive to metacognitive strategies or factors that operate at some level separate from the underlying mnemonic evidence. Wallace's (1978, 1982; Wallace et al., 1978) early findings do not speak to whether or not confidence remains unaltered during standard versus distractor-free tests. Additionally, evidence of shifting confidence in the face of an insensitive old/new decision criterion could inform recognition memory models that assume dependence between the decision criterion and the confidence criteria.

In **Experiment 1**, we compared distractor-free versus standard recognition under conditions in which the participants were forewarned about the different natures of the two types of tests. **Experiment 2** was similar to **Experiment 1**, but instead we used equivalent and more traditional instructions for both the standard and distractor-free recognition tests. Finally, in **Experiment 3** we contrasted standard, distractor-free, and target-free recognition tests. In all three experiments, both endorsement rates and observer confidence ratings were collected. Overall, the data demonstrate that observers do not use the distribution of recognition evidence encountered within a test as an impetus to shift the recognition criterion in an optimal manner. Additionally, the mappings between evidence and confidence also appeared largely similar under standard and modified tests. The implications of these findings for recognition decision models and prior demonstrations of criterion shifts under base rate manipulations are considered further in the General Discussion.

Experiment 1

The purpose of **Experiment 1** was to replicate and extend the findings of Wallace's (1978) original experiment. Participants indicated both old/new decisions and their confidence in these judgments. Additionally, the inclusion of confidence

ratings potentially enabled a more sensitive comparison of the two testing conditions than in the earlier work of Wallace et al. (1978), since, as noted earlier, although gross hit rates may not change across the two testing conditions, the confidence of judgments may. Finally, because observers were made explicitly aware of the manipulation, this tested the observers' willingness to shift the criterion in order to maximize correct responding under conditions that required minimal inference. That is, under these conditions, there was presumably no doubt on the part of the individual that increasing the tendency to label items as "old" would increase detection rates in the distractor-free version of the test.

Method

Participants A total of 38 Washington University undergraduate students participated in return for course credit. One participant was removed due to poor performance ($d' = 0.13$) during the standard recognition condition, leaving 37 participants for the analysis. Informed consent was obtained for all participants, as required by the university's institutional review board.

Materials and procedure The experiment was generated via the Psychophysics Toolbox (Version 3.0.8) (Brainard, 1997; Pelli, 1997) implemented in MATLAB using a standard keyboard and PC. The stimuli were drawn from a pool of 1,216 words selected from the MRC Psycholinguistic Database (Wilson, 1988). The words in the pool had an average of 7.09 letters, 2.35 syllables, and a mean Kučera–Francis frequency of 8.84. This word pool was used for all experiments. Target and lure lists were formed by randomly sampling from this word pool for each participant.

The experiment consisted of two study/test cycles with each study and test phase separated by a 5-min filler task involving simple addition problems. During study, participants indicated the number of syllables for 60 serially presented items in a self-paced manner. Each word was presented individually on screen, and participants indicated whether the presented word contained "1/2/3 or more" syllables using the corresponding number key on the number pad. During the filler task, the computer would randomly generate two 1- to 3-digit numbers that participants would be required to add together. Participants typed their responses using the keypad, and the next filler trial began after they pressed the "Enter" key.

A self-paced recognition test was administered immediately following the filler task. The following instructions were presented prior to the standard recognition test:

In the following list you will be required to distinguish between words that you just performed the syllable-rating task upon and intermixed new words that you have not seen before in this study.

We are interested in whether you experience a feeling of having previously encountered each word from the prior list (feels old) or whether the word strikes you as new (feels new). For items that feel old, this may resemble the sense of familiarity for the item itself, and it may or may not be accompanied by specific thoughts that you had when you previously saw the word—for example, thinking it was a difficult word to pronounce. Alternatively, if a word feels new, you may experience a lack of familiarity with the item. For each presented word please press '1' if the word feels 'old', or '2' if it instead feels 'new'. Following each report you will also be asked to rate the strength of the feeling (1–mild, 2–moderate, or 3–strong). There will be prompts on the screen to help you remember the key assignments.

The following instructions were presented prior to the distractor-free recognition test:

In the following list, you will be shown the same list of words that you just performed the syllable-rating task upon, although they will be shown in a different order. We are interested in whether you experience a feeling of having previously encountered each word from the prior list (feels old) or whether the word strikes you as new (feels new). For items that feel old, this may resemble the sense of familiarity for the item itself, and it may or may not be accompanied by specific thoughts that you had when you previously saw the word—for example, thinking it was a difficult word to pronounce. Alternatively, although every word was previously shown, some may strike you as feeling new, as though they were not shown to you before. There are no correct or incorrect answers. We are simply interested in how the words 'strike you' in terms of feelings of 'oldness' or 'newness' for each word.

For each presented word please press '1' if the word feels 'old', or '2' if it instead feels 'new'. Following each report you will also be asked to rate the strength of the feeling (1–mild, 2–moderate, or 3–strong). There will be prompts on the screen to help you remember the key assignments.

Both sets of test instructions correctly informed participants about the presence or absence of new items in an upcoming recognition test. Both sets of instructions encouraged participants to respond how each item "struck them" in terms of perceived oldness or newness. Instructions for the distractor-free test emphasized the subjective nature of the judgment by informing participants that there were "no correct or incorrect answers."

During the recognition tests, items were presented individually on screen, and participants first indicated

whether they felt that a given item was old or new by pressing either “1” or “2” on the number keypad. Immediately following the old/new decision, participants were asked to rate the subjective strength of that feeling as “mild,” “moderate,” or “strong” by using the “1,” “2,” or “3” key, respectively, on the number pad. The use of “mild,” “moderate,” or “strong” descriptors allowed us to use the same ratings across the standard and distractor-free groups, which again were asked to rate their subjective impressions of the recognition probes. The standard recognition test list was composed of all 60 studied items and an equal number of novel distractors. The distractor-free test list was composed of all 60 studied items.

Following completion of the first test, each participant was administered a second study/test cycle, this time completing whichever recognition test condition they had not yet completed. Test orders were counterbalanced across participants.

Results and discussion

Our analyses examined how old item endorsements and the confidence of those endorsements were affected by the presence or absence of distractors.

Hit rate and hit confidence analyses Mean hit rates and the mean confidence assigned to hits are shown in Table 1, and the distributions of endorsements across levels of confidence are shown in Fig. 2. A 2×2 mixed ANOVA examining the effects of test type (standard vs. distractor free) and test order on hit rates demonstrated a main effect of test type [$F(1, 35) = 4.16$, $p < .05$, $\eta^2 = .11$], with a higher mean hit rate in the standard than in the distractor-free conditions. There was no main effect of test order [$F(1, 35) = 0.06$, n.s.] nor an order $\times$ test type interaction [$F(1, 35) = 0.07$, n.s.]. The latter two extremely low F values suggest no hint of any effects or interactions with test order in the data, and hence we do not consider this factor further.

A similar 2×2 (test type $\times$ test order) mixed ANOVA was used to examine the mean confidence assigned to hits. For

mean confidence, a main effect of test type [$F(1, 35) = 5.94$, $p < .05$, $\eta^2 = .14$] indicated that participants were actually less confident during hits in the distractor-free test than in the standard test. Again, there was no main effect of test order [$F(1, 36) = 0.42$], nor an interaction between test type and test order [$F(1, 35) = 3.02$, $p = .09$, $\eta^2 = .08$].

Contrary to the null findings of Wallace et al. (1978), we found that the two test types did differ in hit rates and hit confidence. However, this difference was in a counterintuitive direction—in the distractor-free condition, when participants were aware that they would only be tested on old items, they correctly endorsed fewer of them, and did so with less confidence. Thus, in contrast to an opportunistic or adaptive criterion, observers actually became slightly more cautious and uncertain under conditions in which they knew the actual study status of all of the test probes. There are at least two possible explanations for this outcome—one evidence-based and the other linked to demand characteristics.

Prior theorists have suggested that observers place the criterion so as to halve the range of the experienced evidence (Estes & Maddox, 1995), or somewhat similarly, they may place it on the basis of an estimate of the modal evidence value. Psychologically, this would be linked to the belief that the test is balanced with respect to targets and distractors. Either mechanism would lead observers to become more conservative when lures are removed, since the lower range of evidence values would presumably be greatly reduced. Critically, however, for observers to take this approach during the distractor-free version of the test, they would have to completely discount the instructions that inform them that only old items will be tested. We know of no strong reason for them to do so, particularly since the first few items of the test should confirm that the instructions are valid. Given this argument, we do not consider this account further.

Turning to potential demand characteristics, the differences between the tests may reflect the observers’ response to the instruction that there were “no correct or incorrect answers,” and that we were only interested in how each item “struck them” in the distractor-free condition. This may have led them to introspect somewhat more rigorously in the distractor-free condition.

Table 1 Mean hit rates and confidence assigned to hits for the two test types and two order conditions (Exp. 1), with standard deviations presented in parentheses

Test Order	Hit Rate			Hit Confidence		
	Standard	Distractor-Free	Mean	Standard	Distractor-Free	Mean
SR-DF	.76 (.09)	.71 (.10)	.74 (.09)	2.52 (0.30)	2.50 (0.32)	2.53 (0.32)
DF-SR	.76 (.09)	.72 (.13)	.74 (.11)	2.51 (0.29)	2.38 (0.30)	2.44 (0.30)
Mean	.76 (.09)	.72 (.11)		2.52 (0.29)	2.45 (0.32)	

SR stands for “standard recognition,” and DF for “distractor-free recognition.”

Fig. 2 Proportions of each response type in both standard and distractor-free recognition conditions (Exp. 1). Confidence ratings range from a high-confidence correct response to a high-confidence incorrect response

To control for these possibilities, we adopted a similar protocol in [Experiment 2](#). However, the test instructions were equated between both conditions (standard and distractor-free), with both adopting instructions in which the proportions of old and new items in the test list were unstated. If observers were insensitive to aggregate differences in evidence across the tests, and instead were exercising differential rigor due to the different instructions preceding the tests in [Experiment 1](#), then the patterns of responding would be predicted to be even more similar in [Experiment 2](#) than in [Experiment 1](#).

Experiment 2

Method

Participants [Experiment 2](#) consisted of 36 Washington University undergraduate students enrolled in general psychology. All participants received course credit for their participation, and informed consent was obtained as required by the university's institutional review board. Two participants were excluded from the analysis because of poor performance ($d' = -0.46$ and -0.05) on the standard recognition test. All statistics reflect the remaining 34 participants.

The experimental protocol was identical to that in [Experiment 1](#), save for the test instructions, which were equated for both the standard and distractor-free conditions.

In both conditions, participants were given the following instructions:

In the following list, you will be shown another list of words. It will contain the same words you saw previously, while it may also contain new words. We are interested in whether you experience a feeling of having previously encountered each word from the prior list (feels old) or whether the word strikes you as new (feels new).

For items that feel old, this may resemble a sense of familiarity for the item itself, similar to the feeling you get when you bump into an acquaintance even if you don't know specifically why they strike you as familiar. Alternatively, a word may trigger remembrance of specific thoughts or events that you experienced when you previously saw the word—for example, thinking to yourself that the word was difficult to pronounce, or seemed misspelled, et cetera. For items that feel new, you may experience a lack of familiarity with the item or it may stand out as novel in the context of this experiment.

There are no correct or incorrect answers. We are simply interested in how the words 'strike you' in terms of feelings of 'oldness' or 'newness' for each word.

For each presented word please press '1' if the word feels 'old', or '2' if it instead feels 'new'.

Following each report you will also be asked to rate the strength of the feeling (1—mild, 2—moderate, or 3—strong). There will be prompts on the screen to help you remember the key assignments. If you have any questions, feel free to ask the experimenter.

In this way, participants were given no explicit indication of whether a recognition test would or would not contain distractor items.

Results and discussion

Hit rate and hit confidence analyses Overall hit rates and the confidence assigned to hits are shown in [Table 2](#). As in [Experiment 1](#), we analyzed the difference in hit rates across test formats using a 2×2 (test type $\times$ test order) mixed ANOVA. In this experiment, however, there was no main effect of test type [$F(1, 32) = 0.09$, n.s.] or test order [$F(1, 32) = 0.10$, n.s.], nor an interaction [$F(1, 32) = 0.52$, n.s.]. The mean confidence for hits was analyzed using a similar 2×2 (test type $\times$ test order) mixed ANOVA. There were no main effects of test type [$F(1, 32) = 2.69$, $p = .11$, $\eta^2 = .08$] or test order [$F(1, 33) = 0.24$, n.s.]. The interaction between test type and test order on mean confidence ratings for hits approached significance [$F(1, 33) = 3.91$, $p = .06$, $\eta^2 = .11$],

Table 2 Mean hit rates and confidence assigned to hits for the two test types and two order conditions (Exp. 2), with standard deviations presented in parentheses

Test Order	Hit Rate			Hit Confidence		
	Standard	Distractor-Free	Mean	Standard	Distractor-Free	Mean
SR-DF	.80 (.10)	.78 (.11)	.79 (.11)	2.56 (0.26)	2.66 (0.31)	2.61 (0.28)
DF-SR	.78 (.08)	.78 (.11)	.78 (.09)	2.66 (0.25)	2.65 (0.30)	2.66 (0.27)
Mean	.78 (.09)	.78 (.11)		2.61 (0.26)	2.66 (0.30)	

SR stands for “standard recognition,” and DF for “distractor-free recognition.”

which suggested that, during the standard test, participants who initially took a standard recognition test were less confident about hits than were participants who initially took a distractor-free test (2.56 vs. 2.66; see Table 2). This difference in hit confidence during the standard test did not survive post-hoc comparisons ($p < .05$, Tukey’s HSD). Given the high similarity of confidence in three of the four cell means and the similar distributions of endorsement confidence (Fig. 3), we will not interpret this marginally significant order-related effect further.

Overall, the use of matched instructions in the two types of tests appears to have eliminated the main effect of test type on accuracy and confidence observed in Experiment 1. This lends support to the idea that participants in the first experiment may have engaged in a more rigorous or cautious examination of recognition evidence because they were made explicitly aware that the test list would not contain any distractors.

As Wallace (1980) argued, it seems that observers do not capitalize on the distribution of item evidence in order to maximize correct responding. In Experiment 2, when

instructions were fully matched, there was no effect of test type on accuracy or confidence, and indeed, when they were explicitly aware that all of the test items were studied in Experiment 1, observers appeared to scrutinize their memories more rigorously or cautiously, or the opposite of an opportunistic criterion shift.

The above findings indicate that the assessment of positive recognition evidence remains very similar across standard and distractor-free tests. Perhaps this result is not so counterintuitive. As suggested in the introduction, participants may enter a recognition memory experiment expecting, unless otherwise instructed, to actually recognize some sizable portion of items during the test. As long as this coarse goal is met (i.e., an item is being recognized every few trials) and no external information suggests that performance is somehow deficient, participants may assume that the current criterion is adequate. In short, a motivated-to-recognize account assumes that individuals are not neutral with respect to outcomes in situations described as tests of memory and that they strive to successfully demonstrate the ability to recognize studied targets. If this general goal is achieved, there is no clear need to adjust the approach to the task. To test this account, Experiment 3 includes a test with no targets. If participants maintained typical false alarm rates on such a test, this would clearly conflict with a motivation to demonstrate successful recognition, because participants should be responding “new” to almost all probes. Hence, in contrast to distractor-free tests, target-free tests might instead be expected to motivate a criterion shift even when instructions are matched across standard and pure-list tests. In fact, this account makes the somewhat counterintuitive prediction that participants would adopt more liberal criteria during a target-free test because of their perceived lack of positive recognition responses and, hence, actually reduce the number of correct responses made.

Ley and Long (1987, 1988) appear to be the only other researchers to have used a target-free recognition paradigm. In each of their experiments, the authors compared performance between standard recognition and pure-list tests. In the standard test, targets and lures were randomly intermixed across the test. In contrast, during the pure-list

Fig. 3 Proportions of each response type in both standard and distractor-free recognition conditions (Exp. 2). Confidence ratings range from a high-confidence correct response to a high-confidence incorrect response

test, only targets were encountered during the first half of the test, and only distractors during the second half. Participants were not explicitly aware of the structure of the test lists, and response confidence was not collected. In their first experiment, Ley and Long (1987) did not indicate whether the difference in the overall false alarm rates was significant (.14 for target-free and .21 for standard recognition¹). In their next experiment, Ley and Long (1988) reported a significant difference in the false alarm rates between target-free and standard recognition (.11 for target-free and .17 for standard recognition; see note 1). Thus, the latter result suggests that observers potentially employed an opportunistic criterion shift, increasing correct rejection rates in the target-free test. These results potentially conflict with those of Wallace (1978), although there was a critical difference in the paradigm. As noted above, the Ley and Long pure lists were in fact segregated lists, such that the first half of the list contained only targets and the last half only distractors. Thus, aside from the fact that Wallace (1982) did not consider target-free situations, the blocked testing format Ley and Long used may have yielded a more salient cue to observers due to the abrupt change in responding “old” as they transitioned from the first to the second half of the test. In this third experiment, we reexamined the target-free paradigm while also including a measure of response confidence, in order to better capture how observers parsed their own recognition experiences during testing. Furthermore, we relied on a pure-list method in which all of the items were either targets or distractors, allowing us to compare the results to our prior experiments with those of Wallace and colleagues (1978).

Experiment 3

In Experiment 3, we contrasted response rates and confidence across standard, target-free, and distractor-free recognition tests. With this novel procedure, we directly compared whether participants assigned old/new and confidence judgments differently when old and new items were presented jointly or singly in a given test.

Participants The sample for Experiment 3 consisted of 36 Washington University undergraduates enrolled in general

psychology. One participant was removed due to poor performance ($d' = -1.32$) during the standard recognition test. Such performance was likely due to the participant misunderstanding the key assignments, because performance for the other two tests was above average. Regardless, this participant's data were removed from the analyses, and all statistics represent the remaining 35 participants. All participants received course credit for their participation, and informed consent was obtained from all, as required by the university's institutional review board.

Procedure During this experiment, participants took each of three types of recognition memory test: standard, distractor-free, and target-free. The study and filler tasks for each test were identical to those in the previous experiments. Upon completion of the filler task, participants received instructions about the upcoming memory test. The instructions for all three memory tests were identical to the those used during Experiment 2, except that participants were instructed that a given test would contain “some combination” of words they had seen previously and words they had not previously encountered in the context of the experiment. The standard recognition test list was composed of all 60 studied items and an equal number of novel distractors. The distractor-free test list was composed of all 60 studied items. The target-free test list was composed of 60 novel distractors. Orders of the study–test cycles (standard, distractor-free, and target-free) were fully counterbalanced, yielding six distinct study–test orders.

Results

Standard vs. distractor-free recognition Overall hit rates and the confidence assigned to hits are shown in Table 3 and illustrated in Fig. 4. Our initial analyses focused on hit rates and hit confidence across the standard and distractor-free test versions using 2×6 (test type $\times$ test order) mixed ANOVAs. For hits, there was no main effect of test type [$F(1, 29) = 1.89, p = .18, \eta^2 = .06$], no main effect of test order [$F(5, 29) = 0.53, n.s.$], and no test type $\times$ test order interaction [$F(5, 29) = 0.88, n.s.$]. For hit confidence, there was no main effect of test type [$F(1, 29) = 2.35, p = .13, \eta^2 = .07$] and no main effect of test order [$F(5, 29) = 2.33, p = .07, \eta^2 = .29$], although the latter effect did approach significance. Furthermore, the test type $\times$ test order interaction was not significant [$F(5, 29) = 1.03, p = .41, \eta^2 = .15$]. Examining the confidence ratings in Table 3 provides a potential explanation for the trending test order effect: The “standard–target-free–distractor-free” (S-T-D) order group was generally more confident in their responses than the remaining five order groups. We will further consider

¹ The original articles reported the mean number of false recognitions for each test. These were converted to proportions for the purposes of this article. The original numbers Ley and Long reported in their 1987 study were 4.94 false alarms for the target-free test and 7.42 false alarms for the standard recognition test. For their 1988 study, Ley and Long reported 5.74 false alarms for the target-free test and 8.87 false alarms for the standard recognition test.

Table 3 Mean hit rates and confidence assigned to hits for the three test types and six order conditions (Exp. 3), with standard deviations presented in parentheses

Test Order	Hit Rate			Hit Confidence		
	Standard	Distractor-Free	Mean	Standard	Distractor-Free	Mean
S-D-T	.75 (.07)	.72 (.13)	.74 (.10)	2.40 (0.16)	2.45 (0.42)	2.42 (0.30)
S-T-D	.74 (.08)	.74 (.14)	.74 (.11)	2.83 (0.15)	2.75 (0.16)	2.79 (0.16)
D-S-T	.75 (.11)	.68 (.19)	.72 (.15)	2.44 (0.34)	2.17 (0.39)	2.30 (0.36)
D-T-S	.69 (.15)	.71 (.15)	.70 (.15)	2.42 (0.48)	2.41 (0.30)	2.42 (0.39)
T-S-D	.79 (.13)	.72 (.15)	.76 (.14)	2.65 (0.17)	2.65 (0.19)	2.65 (0.18)
T-D-S	.64 (.19)	.65 (.21)	.64 (.20)	2.42 (0.45)	2.32 (0.42)	2.37 (0.44)
Mean	.73 (.13)	.70 (.16)		2.53 (0.34)	2.46 (0.36)	

S stands for “standard recognition,” D for “distractor-free recognition,” and T for “target-free recognition.”

whether this group was generally more confident during all reports and conditions at the end of the Results section.

Standard vs. target-free recognition To compare standard and target-free recognition tests, we examined correct rejection rates and confidence across the two test formats (see Table 4). A 2×6 (test type $\times$ test order) mixed ANOVA on correct rejection rates demonstrated a trending main effect of test type [$F(1, 29) = 3.52, p = .07, \eta^2 = .11$], no main effect of test order [$F(5, 29) = 1.66, p = .17, \eta^2 = .22$], and no test type $\times$ test order interaction [$F(5, 29) = 1.85, p = .13, \eta^2 = .24$]. Further examination of the trending main effect of test type indicated that participants made numerically fewer rejections during the target-free test; however, this difference did not survive post-hoc comparison (Tukey’s HSD, $p > .05$). Critically, this numerical difference

was in the direction opposite the one that would be expected under an opportunistic criterion shift.

A 2×6 (test type $\times$ test order) mixed ANOVA on correct rejection confidence demonstrated no main effect of test type [$F(1, 30) = 2.18, p = .15, \eta^2 = .07$]; however, there was a main effect of test order [$F(5, 30) = 3.83, p < .01, \eta^2 = .40$]. The interaction between test type and test order was not significant [$F(5, 30) = 0.70, n.s.$] (see Fig. 5). Examination of Table 4 shows that once again the S-T-D order group demonstrated much higher confidence than the other groups, likely driving the significant order effect and suggesting that this group was generally higher in the confidence of all reports, regardless of whether old or new. We consider this possibility next.

As mentioned above, one of the between-subjects order conditions seemed to demonstrate aberrantly high levels of response confidence across all tests; that is, participants in this group tended to use the extreme confidence keys when responding to any item. A separate analysis examining the effects of test order on overall response confidence (across all three tests) confirmed this. This analysis revealed a significant effect of test order group on overall response confidence [$F(5, 29) = 3.42, p < .05, \eta^2 = .37$], with the S-T-D group showing systematically higher confidence than the other order groups (Fig. 6); post-hoc tests indicated that this group showed significantly greater response confidence than three of the five other groups (Tukey’s HSD, $ps < .05$). It is unclear why this particular order group rated their responses more confidently than the others did; however, given the small sizes of each of the order groups ($N = 5$ in the “standard–distractor-free–target-free” group, $N = 6$ in all other groups), this effect could have easily arisen from a handful of participants simply relying on the extremes of the confidence scale for most responses, a problem not uncommon to confidence ratings during recognition tests. Indeed, when this group was excluded and the confidence analyses above were rerun, there was no evidence for a main effect of test order or a test order $\times$ test type interaction

Fig. 4 Proportions of each response type in both standard and distractor-free recognition conditions (Exp. 3). Confidence ratings range from a high-confidence correct response to a high-confidence incorrect response

Table 4 Mean correct rejection rates and confidence assigned to correct rejections for the three test types and six order conditions (Exp. 3), with standard deviations presented in parentheses

Test Order	Rejection Rate			Rejection Confidence		
	Standard	Target-Free	Mean	Standard	Target-Free	Mean
S-D-T	.81 (.08)	.82 (.18)	.82 (.13)	1.91 (0.16)	2.00 (0.36)	1.96 (0.26)
S-T-D	.88 (.10)	.73 (.16)	.80 (.13)	2.51 (0.29)	2.46 (0.39)	2.48 (0.34)
D-S-T	.79 (.10)	.80 (.09)	.80 (.10)	1.96 (0.30)	2.00 (0.34)	1.98 (0.32)
D-T-S	.77 (.14)	.79 (.11)	.78 (.12)	2.20 (0.43)	2.33 (0.39)	2.26 (0.41)
T-S-D	.71 (.10)	.61 (.04)	.66 (.07)	2.00 (0.22)	1.99 (0.25)	2.00 (0.24)
T-D-S	.73 (.15)	.71 (.20)	.72 (.18)	1.74 (0.43)	1.93 (0.20)	1.84 (0.32)
Mean	.78 (.12)	.74 (.15)		2.06 (0.39)	2.12 (0.37)	

S stands for “standard recognition,” D for “distractor-free recognition,” and T for “target-free recognition.”

when examining hits [test order, $F(4, 24) = 0.98$, n.s.; test type $\times$ test order, $F(4, 24) = 1.11$, $p = .38$, $\eta^2 = .16$] or when examining correct rejections [test order, $F(4, 24) = 1.67$, $p = .19$, $\eta^2 = .22$; test type $\times$ test order, $F(4, 24) = 0.48$, n.s.]. Given this, we concluded that test order did not have an appreciable effect on confidence and that, instead, the S-T-D group’s generally high confidence level was the result of a high proportion of observers simply choosing to use the most extreme confidence keys.

Analysis of initial test performance

Standard vs. distractor-free recognition Although the analyses above did not yield evidence for opportunistic shifts in

response rates across the test formats, one might be concerned that such differences could be obscured by treating test type as a within-subjects factor. More specifically, one might be concerned that effects were present, but were somehow washed out or obscured by carryover effects across the different orderings that each participant experienced. This might be a particularly important factor to consider, given that three different types of tests were administered in this experiment². To address this concern, we focused solely on the initial test of each test order group. This yielded three separate groups (standard group $N = 11$; distractor-free group $N = 12$; target-free group $N = 12$), each of which took a different type of initial test. Performance on this test could not be contaminated by carryover effects from other formats. Using independent-samples t tests, we analyzed whether correct response rates or confidence differed among these three groups. Again, these analyses allowed us to exclude test order as a factor and simply compare performance on each test, as if the experiment were a between-group comparison of each test format—although, of course, the reduced sample size corresponding to each test somewhat weakened power. We considered all participants initially and then excluded the potentially overly confident S-T-D group identified above.

Hit rates did not differ for the standard and distractor-free recognition test groups [.75 vs. .69; $t(21) = 1.06$, n.s.]. When we turn to hit confidence, hits were rendered with significantly more confidence during the initial standard recognition test than during the initial distractor-free test [2.64 vs. 2.29; $t(21) = 2.60$, $p < .05$]. When the data were reconsidered without the S-T-D order participants (who generally had inflated confidence), the hit rates remained relatively unchanged [.76 vs. .69; $t(15) = 0.85$], and the difference in hit confidence between the

Fig. 5 Proportions of each response type in both standard and target-free recognition conditions (Exp. 3). Confidence ratings range from a high-confidence incorrect response (FA) to a high-confidence correct response (CR)

² We thank a helpful reviewer for this suggestion.

Fig. 6 Mean response confidence (1=*low confidence*, 2=*medium confidence*, 3=*high confidence*) across all tests and all response types, separated by six order conditions (Exp. 3). S stands for “standard recognition,” D stands for “distractor-free recognition,” and T stands for “target-free recognition.” Error bars represent one standard error of the mean

standard and distractor-free groups was eliminated [2.40 vs. 2.29; $t(15) = 0.64$].

Standard vs. target-free recognition The next analyses compared correct rejection rates and confidence between the standard and target-free groups. Correct rejection rates were higher for the standard versus the target-free group [.84 vs. .69; $t(21) = 3.69$, $p < .01$]. Likewise, correct rejections were more confident for the standard versus the target-free group [2.24 vs. 1.96; $t(21) = 2.09$, $p < .05$]. When the data were reconsidered without the S-T-D order participants, the correct rejection rates remained significantly different [.81 vs. .66; $t(15) = 2.21$, $p < .05$], while the difference in correct rejection confidence between the standard and target-free groups was eliminated [1.91 vs. 1.96; $t(15) = -0.51$].

Experiment 3 confirms the previous findings that hit rates and hit confidence are generally insensitive to the presence or absence of distractors at test (when instructional demand characteristics are controlled). In contrast, correct rejection rates were not insensitive to test construction. Instead, the data suggest a small but reliable tendency for participants to make fewer correct rejections when the test in fact contained no targets, as compared to a standard recognition test. This effect was marginally significant in the full analysis, and clearly so in the analysis that only considered performance on the first test. This seems to support the motivated-to-recognize account outlined earlier. Importantly, the observed pattern in performance was again inconsistent with the notion of an opportunistic criterion; in fact, the reduction of rejection rates during the target-free

test was the opposite of what would be predicted from an opportunistic criterion. These points will be discussed more fully in the General Discussion section.

General discussion

We examined whether observers opportunistically shift their decision criteria in response to extreme changes in the distributions of old and new items within individual tests. In the case of hits and hit confidence, the findings are clear: The frequency and confidence of recognition reports is largely unaffected by whether or not verbal recognition tests mix old and new items, or are instead constructed exclusively of old items. Furthermore, even though a subtle shift was detected in [Experiment 1](#), [Experiment 2](#) demonstrated that this was likely the result of the different instructions given immediately before the standard and distractor-free test formats. When observers knew that the test would be composed entirely of studied materials, they endorsed fewer items and were less confident in their hits. Because this pattern was eliminated when identical instructions were presented prior to the two types of test in [Experiment 2](#), it appears that the pattern was the result of participants adopting a more rigorous or introspective stance when they knew that all of the items were in fact studied. Nonetheless, even under these circumstances, the criterion shift was not opportunistic, but was in fact inopportune (e.g., a more conservative response criterion during a distractor-free test). Thus, neither experiment supports the idea of an adaptive criterion that improves success rates in response to the relative distributions of old and new items in the test.

In the case of correct rejections, [Experiment 3](#) demonstrated an inopportune shift even with equivalent instructions for the standard and pure-list test variants. Participants correctly rejected new items less often during the target-free test than during the standard recognition test. That is, they demonstrated a somewhat liberal criterion when old items were wholly removed from the test, thus committing more false alarms. Again, this was the opposite of what would occur under an opportunistic criterion shift. Why were participants sensitive to the target-free manipulation but not the distractor-free manipulation under equated test instructions? Our tentative explanation is termed the *motivated-to-recognize account*. Under this account, we assume that observers are not neutral with respect to old and new decisions during tests that are described to them as “memory” tests. In short, we assume that individuals prize recognizing studied materials in tests that are described to them as tests of recognition memory. This assumption is supported in part by functional imaging research demonstrating greater activation during hits than during correct rejections in striatal brain regions linked

to reward and motivation (Han, Huettel, Raposo, Adcock, & Dobbins, 2010; McDermott, Szpunar, & Christ, 2009). Under this account, a distractor-free test is not particularly salient, in that the observers' goal of recognizing a substantial portion of the probes is easily met. In contrast, during target-free tests observers may decide that their memory is functioning quite poorly, because they are not experiencing many of the items as recognized, and given this impression, may strategically adopt a more liberal stance. This tendency, if the account is correct, should be sensitive to the framing of the testing situation by the experimenter. For example, if participants were led to believe that the detection of novelty was the key focus, or perhaps that it was heavily linked to desirable attributes such as intelligence, the account would predict the reverse pattern from the one found in the present data. That is, when instructions were matched, a novelty-focused framing should yield an inopportune criterion shift during distractor-free recognition (unlike in Exps. 2 and 3) and no criterion shift during target-free recognition (unlike in Exp. 3).

In contrast to the motivated-to-recognize account, it is important to consider the potential influence of task instructions on behavior during the target-free test. Prior to the test, participants were informed, somewhat misleadingly, that the test contained "some combination of old and new items," which participants likely interpreted to mean that there were some nonzero quantities of old and new items. Given this interpretation, it would make sense for participants to shift their decision criterion if their recognition expectations were not being met (i.e., not recognizing or rejecting enough items). However, this notion fails to account for the asymmetry present in the data—namely, that participants shifted during a target-free test, but did not shift the decision criterion during a distractor-free test. Given that the motivated-to-recognize theory accounts for the observed patterns in the data, we prefer this interpretation.

Although the findings suggest that observers do not hold recognition and rejection experiences in equal favor, it is important to emphasize that the findings with neither hits nor correct rejections support the idea of an opportunistically adaptive criterion geared toward maximizing success, adding to a growing body of literature demonstrating that irregularities in the distribution of memory evidence within a given test do not serve as salient cues for opportunistic or "adaptive" criterion shifts (Verde & Rotello, 2007).

The findings reported here seem to contrast with the various instances of successful base rate manipulations throughout the literature (i.e., cases in which participants do shift the criterion to maximize correct responding in response to different base rates of old and new items). Crucially, in cases where observers do adaptively position the decision criterion in response to base rate manipulations, they appear to require some external cue to do so, such as explicit instructions informing them that the base rates of old and

new items are wildly different (Healy & Kubovy, 1978; Ratcliff, Sheu, & Gronlund, 1992; Van Zandt, 2000) or performance feedback procedures that likely make such differences readily apparent (Rhodes & Jacoby, 2007). However, simply making participants aware of the base rate differences is clearly not sufficient, because the data of Wallace (e.g., Wallace et al., 1978) and those here do not demonstrate the opportunistic shifts documented in these prior works. The key distinctions seem to be the nature of the test instructions and the difference between knowing that there is a difference in base rates versus knowing and being encouraged to capitalize on that difference in base rates. Specifically, the distractor-free instructions that Wallace et al. used emphasized to participants that, even though all items had been previously studied, participants were only to endorse those items that they actually recognized. That is, his instructions emphasized that participants should report only their subjective experiences in response to the items, as opposed to describing the task as a "test," which arguably would instead emphasize the need to score well. Likewise, our distractor-free instructions in Experiment 1 emphasized that participants should respond with how items subjectively "struck them" in terms of perceived oldness, despite the fact that all presented items were actually old.

Under these conditions, it is clear that participants do not opportunistically adjust the decision criterion. However, if one revisits many of the earlier studies in which opportunistic shifts were achieved with base rate manipulations, it becomes clear that not only were participants made aware of the base rate differences, but that the procedures actively encouraged them to use this information to increase performance scores. For example, in Healy and Kubovy (1978), participants were exposed to 50/50 and 25/75 old-to-new base rate manipulations, and large criterion shifts were observed. However, not only were they given these base rates prior to every test block, but they also were quizzed on this information prior to each test block, (described as a "computer assisted test on the two prior probabilities") and were provided with corrective feedback on every trial, along with accumulated earnings totals as a function of accuracy. All of this presumably suggested to the participants that their goal should be to use the probabilities to maximize the winnings achieved in the experiment. In essence, one could argue that procedures such as these emphasize to participants that the goal is to maximize a performance score rather than faithfully reporting subjective experiences of recognition, a distinction emphasized by Wallace (1978).

These sorts of encouragements are absent in the Wallace procedures (Wallace et al., 1978) and likewise in the present Experiment 1. Here, participants were aware of the construction but encouraged to report their subjective mnemonic experiences. Additionally, we also emphasized that this was the primary purpose of the experiment, in

contrast to emphasizing maximizing correct responding. Indeed, the notion of maximizing correct responding arguably makes no sense when observers explicitly know that the test is composed entirely of studied materials. In Experiments 2 and 3, they were not made explicitly aware of the test list construction, and again they failed to demonstrate an opportunistic criterion shift. These findings somewhat resemble those of Verde and Rotello (2007), who surreptitiously manipulated the strength of targets across different halves of recognition tests. Critically, when transitioning from a test half containing strongly encoded targets to a half containing weakly encoded targets, the participants' criteria remained fixed, suggesting that they were likely insensitive to the manipulation. Thus, when considered in conjunction with the prior findings, the present data suggest at least two factors that influence whether or not observers demonstrate opportunistic shifts as a function of base rate or similar manipulations. The first has to do with individual awareness of the manipulation. Studies that demonstrate opportunistic shifts by manipulating base rates either make this information explicitly known to participants or encourage its detection through performance feedback. The second factor reflects whether or not participants are encouraged to capitalize on base rate imbalances, which necessarily requires that they be made aware of such imbalances. Manipulations such as repeatedly reminding them of such imbalances and providing payment and feedback linked to accurate performance all serve to notify participants that not only are imbalances present, but that they are expected to use these imbalances in an opportunistic fashion to improve net performance on the task. Thus, it appears that participants do not spontaneously detect such manipulations, or if they do, they require some further encouragement or impetus to opportunistically shift the decision criterion.

While the tendency *not* to spontaneously translate base rate information into opportunistic responding might be characterized as somehow nonoptimal, one might instead consider what would occur if, instead, short-term regularities in evidence values within tests led to frequent updating of the decision criterion. Given that targets and lures are randomly intermixed on most recognition memory tests, it is likely that a participant would often encounter runs of several or more items falling reliably above or below the current criterion. If a participant were sensitive to these trends and responded by updating his or her response criterion based on the mistaken belief that the outcomes indicated an overly lax or strict criterion, the net effect would actually be a reduction in net accuracy. That is, trial-to-trial variability in the decision criterion is statistically more likely to lead to errors than successes, and hence drives measured performance downward (for discussions, see Benjamin, Diaz, & Wee, 2009; Han & Dobbins, 2008). Thus, from this perspective, insensitivity to these types of statistical trends is arguably

ideal. Additionally, actively monitoring such trends would arguably be quite cognitively taxing, since it would require continuously monitoring the outcome of at least several previous trials in order to adjust the criterion (Stretch & Wixted, 1998b). The only evidence favoring any active monitoring on the part of participants lies in the findings of Experiment 3, in which observers actually slightly reduced correct rejection responding during the target-free test. Again, however, this is the opposite of an opportunistic approach, and as we note above, arguably reflects the fact that “recognizing” tends to be more valued than “rejecting” on tests described as memory tests. Whether such a shift would remain if the test were described as one of novelty detection, instead of recognition memory, is an interesting question awaiting further research. Under the motivated-to-recognize account, such a manipulation should eliminate the shift, as noted earlier.

Although a criterion insensitive to short-term regularities seems concordant with independence between the criteria and evidence, the present data do not necessarily fit well with the original derivation of the signal detection model, in which the decision axis is not one of sensory or mnemonic strength, but instead one of relative likelihood (Pastore, Crawley, Berens, & Skelly, 2003). To illustrate the difference, the bottom *x*-axis labels in Fig. 7 express the evidence in terms of the logarithm of the ratio of the density of the old distribution in relation to the density of the new distribution at each point on the axis, the *log likelihood ratio*. When this ratio is greater than 0, the evidence favors

Fig. 7 Example of the log likelihood ratio (bottom *x*-axis) and strength (top *x*-axis) models of recognition memory. For any given level of strength, the corresponding likelihood ratio evidence value is computed by dividing the probability density of the target distribution at that location by the probability density of the lure distribution at that location. This value specifies the odds in favor of the item being old given the current evidence distributions. Because the boundaries of the raw likelihood ratio scale are asymmetric [0, infinity], it is often easier to work with the log of the likelihood ratio, which is centered on zero (i.e., even odds in favor of old) and has upper and lower boundaries of positive and negative infinity, respectively

classifying an item as old; when it is less than 0, it favors classifying the item as new; and when it is equal to 0, the item is equally likely to arise from either distribution. In short, on a likelihood ratio axis, the observer has information specifying the odds that an item with a given strength originated from either the target or the lure distribution. There are numerous statistical and practical benefits to a likelihood ratio decision axis, not the least of which is that evidence is expressed in terms of a universal variable that unambiguously signals the statistically optimal response under various conditions (Macmillan & Creelman, 2005). Nonetheless, models assuming this or similar types of decision axes have been criticized, because they arguably require that the observer know the relative locations and shapes of the evidence distributions in order to calculate the likelihood ratio estimate (Criss & McClelland, 2006).

The present study adds to those criticisms of the likelihood ratio evidence assumption. It is unclear how observers would calculate likelihood ratios when entire distributions of evidence are absent from the tests. Put simply, a likelihood ratio evidence variable is undefined in either the target-free or distractor-free tests, if one assumes that it is estimated based on the distribution of evidence during the course of testing. One way that participants might overcome this problem is to assume that they also use data from the study phase in an attempt to gain knowledge about the distribution characteristics of the memoranda (see, e.g., Hintzman, 1994). For example, in the distractor-free tests here, one could argue that the initial experiences with the items during encoding provided enough information for the observer to estimate the location of a novel item distribution. As each item was encountered during study, its novelty in that context could be registered and in turn used to estimate the distributed familiarity of novel items within that context. Given this estimate, it would still be possible to employ a likelihood ratio criterion during the subsequent test, even if new items were not presented. However, even under this scenario, it is still unclear how the observer would estimate the locations of the distributions, because feedback was absent (cf. Brown, 2010). Regardless, the problem of missing distributions is further compounded when one considers the target-free test condition. Under this condition, it is unclear how the observer would actually locate the distribution estimate of memory strengths for the studied items, given that these were never presented at test. Additionally, appealing to processes at encoding is not feasible, since although novelty may be registered fairly automatically during the course of encoding the materials, a bootstrapped estimate of the future strength of the items under some undefined future testing condition seems quite implausible. Thus, the present data, along with the original studies by Wallace (1978), clearly suggest that a likelihood ratio decision axis is

untenable if the evidence distributions have to be estimated by the observer during the course of the experiment. Given the present and the reviewed findings, it seems that a firm conclusion regarding the feasibility of the likelihood ratio decision axis can be reached: Namely, if the estimation of the distributions must take place within the experiment session, the likelihood ratio decision assumption is clearly untenable.

If a formal likelihood ratio account is untenable, it begs the question of how the participants in the present studies demonstrate such remarkable similarity in their response and confidence patterns across standard and pure-list tests. One possibility, suggested by Wixted and Gaitan (2002) on the basis of animal learning research, is to treat the likelihood ratio account as merely an approximation of a reinforcement learning process that yields behavior that appears as though participants are working with formal likelihood ratios. Under this approach, the core decision variable is indeed item strength or familiarity; however, the translation of strength into “old” and “new” judgments is governed by the observer’s learning history. Presumably, acting on low levels of strength as though they were adequate indicators of prior encounter would not be well reinforced outside the laboratory, so observers learn to match their behavior to various levels of strength in a manner that improves the probability of receiving reinforcement and avoiding punishment. Such an account fits well with the present findings, because it predicts that the rendered judgments will be largely insensitive to whether or not targets and lures are presented in a pure or intermixed manner, provided that feedback or reinforcements are also *not* given during testing. For either pure or mixed tests, the responses are instead assumed to be governed by the prior history of reinforcement outside of the experiment. This account is consistent with recent findings suggesting that recognition response bias remains stable within an individual across a variety of domains (Kantner & Lindsay, 2010). Additionally, the account also fits well with recent demonstrations that the criterion can be easily shifted using fairly subtle manipulations of feedback during recognition testing (Han & Dobbins, 2008, 2009), which demonstrates that reinforcement learning mechanisms appear to easily influence recognition criteria. For example, falsely informing participants that one class of errors (e.g., misses) is in fact correct yields a robust shift in the recognition criterion. Additionally, basal ganglia structures thought to be critical for reinforcement learning processes are also active during the rendering of recognition judgments and appear to track the motivational significance of rendered recognition decisions and the provision of feedback (Han et al., 2010). Under this framework, the reason why decisions look so similar under standard and pure-list test conditions is because both tests measure what is essentially habit responding.

Although a reinforcement learning account of initial criterion and confidence mapping seems quite compatible

with the present findings, it is clearly only a tentative account, because it rests on a hypothetical pattern of unobserved reinforcement that governs the mapping between evidence and behavioral choices. Although this account is consistent with research demonstrating that the criterion can be shifted by means of subtle reinforcement mechanisms (Han & Dobbins, 2008, 2009), other manipulations, such as verbal instructions to the observers to simply be cautious or lax in their recognition decisions (Hirshman, 1995), also easily yield criterion shifts. Since the use of instructions to implement response strategies is presumably independent of an incremental reinforcement learning process, it appears that an ultimate understanding of the regulation of recognition criteria will likely require a multiprocess approach.

Conclusions

Despite severe changes in the distributions of evidence, observers largely maintained similar endorsement rates and patterns of confidence assignment during recognition testing. When subtle differences in endorsement rates were found, these indicated that observers were not responding opportunistically, but instead were lowering their correct response rate on pure-list tests relative to standard recognition tests. It is likely that observers enter a recognition memory experiment with established evidence-to-response tendencies, perhaps acquired through lifelong trial-and-error learning (Wixted & Gaitan, 2002). Such an account would predict a high degree of response stability, even in the face of the extreme manipulations used here, that would persist unless some external factor modified the established learning or overrode reliance on what is essentially a habitual response tendency. From this perspective, the use of a likelihood ratio decision axis is at best an approximation of another process that instead maps directly onto mnemonic strength. In terms of decision models of recognition, the likelihood ratio decision axis assumption should be avoided if its use implies that observers' decisions are highly sensitive to the relative distributions of new- and old-item evidence during testing, because the present data and those of Wallace (1978, 1982) demonstrate that they are not sensitive.

Acknowledgements This research was supported by NIH Grant R01 MHO73982.

References

Benjamin, A. S., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language*, 51, 159–172. doi:10.1016/j.jml.2004.04.001.

- Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review*, 116, 84–115. doi:10.1037/a0014351.
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10, 433–436. doi:10.1163/156856897X00357.
- Brewer, W. F., Sampaio, C., & Barlow, M. R. (2005). Confidence and accuracy in the recall of deceptive and nondeceptive sentences. *Journal of Memory and Language*, 52, 618–627. doi:10.1016/j.jml.2005.01.017.
- Brown, S. D. (2010). *The pervasive problem of criterion setting*. Unpublished paper presented at the Annual Meeting of the Psychonomic Society, St. Louis, MO, November.
- Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review*, 7, 26–48.
- Chandler, C. C. (1994). Studying related pictures can reduce accuracy, but increase confidence, in a modified recognition test. *Memory & Cognition*, 22, 273–280.
- Cleary, A. M., Winfield, M. M., & Kostic, B. (2007). Auditory recognition without identification. *Memory & Cognition*, 35, 1869–1877.
- Criss, A. H., & McClelland, J. L. (2006). Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM). *Journal of Memory and Language*, 55, 447–460. doi:10.1016/j.jml.2006.06.003.
- Curran, T., DeBuse, C., & Leynes, P. A. (2007). Conflict and criterion setting in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 2–17. doi:10.1037/0278-7393.33.1.2.
- Dobbins, I. G., Kroll, N. E. A., & Liu, Q. (1998). Confidence-accuracy inversions in scene recognition: A remember-know analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 1306–1315.
- 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*, 31, 1075–1095. doi:10.1037/0278-7393.21.5.1075.
- 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, 703–715. doi:10.3758/MC.36.4.703.
- Han, S., & Dobbins, I. G. (2009). Regulating recognition decisions through incremental reinforcement learning. *Psychonomic Bulletin & Review*, 16, 469–474. doi:10.3758/PBR.16.3.469.
- Han, S., Huettel, S. A., Raposo, A., Adcock, R. A., & Dobbins, I. G. (2010). Functional significance of striatal responses during episodic decisions: Recovery or goal attainment? *The Journal of Neuroscience*, 30, 4767–4775. doi:10.1523/JNEUROSCI.3077-09.2010.
- Hautus, M. J., Macmillan, N. A., & Rotello, C. M. (2008). Toward a complete decision model of item and source recognition. *Psychonomic Bulletin & Review*, 15, 889–905. doi:10.3758/PBR.15.5.889.
- 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.
- Hintzman, D. L. (1994). On explaining the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 201–205. doi:10.1037/0278-7393.20.1.201.
- 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. doi:10.1037/0278-7393.21.2.302.
- Kantner, J., & Lindsay, D. S. (2010). *Is response bias in recognition memory a cognitive trait?* Unpublished paper presented at the Annual Meeting of the Psychonomic Society, St. Louis, MO, November.

- Ley, R., & Long, K. (1987). A distractor-free test of recognition and false recognition. *Bulletin of the Psychonomic Society*, 25, 411–414.
- Ley, R., & Long, K. (1988). Distractor similarity effects in tests of discrimination recognition and distractor-free recognition. *Bulletin of the Psychonomic Society*, 26, 407–409.
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). New York: Erlbaum.
- McDermott, K. B., Szpunar, K. K., & Christ, S. E. (2009). Laboratory-based and autobiographical retrieval tasks differ substantially in their neural substrates. *Neuropsychologia*, 47, 2290–2298. doi:10.1016/j.neuropsychologia.2008.12.025.
- McKelvie, S. J. (1993). Effects of spectacles on recognition memory for faces: Evidence from a distractor-free test. *Bulletin of the Psychonomic Society*, 31, 475–477.
- Morrell, H. E. 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, 1095–1110. doi:10.1037/0278-7393.28.6.1095.
- Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonparametric" *A'* and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review*, 10, 556–569.
- Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10, 437–442. doi:10.1163/156856897X00366.
- Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, 99, 518–535. doi:10.1037/0033-295X.99.3.518.
- 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. doi:10.1037/0033-295X.33.2.305.
- Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence–accuracy relation in eyewitness identification studies. *Psychological Bulletin*, 118, 315–327. doi:10.1037/0033-295X.118.3.315.
- Stretch, V., & Wixted, J. T. (1998a). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 1397–1410. doi:10.1037/0278-7393.24.6.1397.
- Stretch, V., & Wixted, J. T. (1998b). 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. doi:10.1037/0278-7393.24.6.1379.
- Strong, M. H., & Strong, E. K., Jr. (1916). The nature of recognition memory and of the localization of recognitions. *The American Journal of Psychology*, 27, 341–362. doi:10.2307/1413103.
- Underwood, B. J. (1972). Word recognition memory and frequency information. *Journal of Experimental Psychology*, 94, 276–283. doi:10.1037/h0032785.
- Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 582–600. doi:10.1037/0278-7393.26.3.582.
- Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition*, 35, 254–262.
- Wallace, W. P. (1978). Recognition failure of recallable words and recognizable words. *Journal of Experimental Psychology: Human Learning and Memory*, 4, 441–452. doi:10.1037/0278-7393.4.5.441.
- Wallace, W. P. (1980). On the use of distractors for testing recognition memory. *Psychological Bulletin*, 88, 696–704. doi:10.1037/0033-2909.88.3.696.
- Wallace, W. P. (1982). Distractor-free recognition tests of memory. *The American Journal of Psychology*, 95, 421–440. doi:10.2307/1422134.
- Wallace, W. P., Sawyer, T. J., & Robertson, L. C. (1978). Distractors in recall, distractor-free recognition, and the word-frequency effect. *The American Journal of Psychology*, 91, 295–304. doi:10.2307/1421539.
- Wilson, M. (1988). MRC Psycholinguistic Database: Machine-usable dictionary, Version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20, 6–10.
- Wixted, J. T., & Gaitan, S. C. (2002). Cognitive theories as reinforcement history surrogates: The case of likelihood ratio models of human recognition memory. *Animal Learning & Behavior*, 30, 289–305.

Author note

Supplemental materials may be downloaded along with this article from www.springerlink.com.