Memory & Cognition (2019) 47:842–854
https://doi.org/10.3758/s13421-019-00906-8

A strength-based mirror effect persists even when criterion shifts are unlikely

Gregory J. Koop[1] · Amy H. Criss[2] · Angelina M. Pardini[1]

Published online: 21 February 2019

Abstract
In single-item recognition, the strength-based mirror effect (SBME) is reliably obtained when encoding strength is manipulated between lists or participants. Debate surrounds the degree to which this effect is due to differentiation (e.g., Criss *Journal of Memory and Language, 55*, 461–478, 2006) or criterion shifts (e.g., Hicks & Starns *Memory & Cognition, 42*, 742–754, 2014). Problematically, differing underlying control processes may be equally capable of producing an SBME. The ability of criterion shifts to produce an SBME has been shown in prior work where differentiation was unlikely. The present work likewise produces an SBME under conditions where criterion shifts are unlikely. Specifically, we demonstrate that an SBME can be elicited without the typical number of trials needed to adjust one's decision criterion (Experiments 1, 2, and 5) and using encoding manipulations that do not explicitly alert participants that their memory quality has changed (Experiments 3 and 4). When taken in the context of the broader literature, these results demonstrate the need to prioritize memory models that can predict SBMEs via multiple underlying processes.

Keywords Recognition · Strength based mirror effect · Differentiation · Criterion-shifts

Imagine you have just started teaching at a new university when a friend comes to visit and requests a tour of campus. During this tour, a group of young adults passes by, and your friend asks you if you have any of them in class. Feeling somewhat sheepish, you point to a couple of students that look sort of familiar and identify them as being in your class. (Fortunately, your friend has no way of knowing if you're wrong.) After only a week of classes, it remains exceedingly difficult to distinguish between your own students and the other students that look similar. As luck would have it, your friend again passes through town 8 months later, and you again are walking around campus when she asks you to identify any of your students. Having spent an academic year on campus, you quickly and confidently identify only those students that you actually had in class. Just as importantly, you also note that you are much less tempted to misidentify the other students in that group that you *did not* have in class. It strikes you as odd that you can so easily dismiss these other students because, after all, you have had roughly the same amount of experience with them after a year of classes as you did after a week of class (i.e., none). Why now, after a year has passed, has it become easier to dismiss these unknown students?

This opening example is analogous to a phenomenon in recognition-memory research known as the strength-based mirror effect (SBME; Glanzer & Adams, 1990; Glanzer, Adams, Iverson, & Kim, 1993; Stretch & Wixted, 1998). In a typical SBME task, participants study a series of strong words or weak words, and then complete a single-item recognition test over that material (see Fig. 1 for a version of this design used in Experiment 1). At test, participants are generally presented with equal numbers of studied items (targets) and unstudied (items) foils, and asked to identify each as "old" or "new." The SBME describes the finding that strengthening items at study improves performance in two ways. One's ability to correctly identify previously studied items (the hit rate, or HR) improves, whereas the likelihood of incorrectly recognizing unstudied material (the false-alarm rate, or FAR) decreases. The question posed at the end of the opening example is also one that has persisted in the literature: Why do unstudied foils become easier to reject as the contents of memory are strengthened? Why should foils benefit from an encoding manipulation for which they, by definition, were not present?

✉ Gregory J. Koop
gregory.koop@emu.edu

[1] Department of Psychology, Eastern Mennonite University, 1200 Park Road, Harrisonburg, VA 22802, USA

[2] Syracuse University, Syracuse, NY, USA

Significance of the strength-based mirror effect to memory theory

While all contemporary models of memory predict that strengthening items (whether through repetition, duration, or "depth") leads to a higher HR, the accompanying decrease in FAR has been somewhat more contentious. One reason for this debate stems from the use of signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 2004) to analyze recognition performance. Signal detection theory is a measurement model that describes performance in a recognition-memory task along two spectra: discriminability and bias. Discriminability represents the ease with which targets can be distinguished from foils. Targets, by virtue of being studied, generate more mnemonic evidence than foils during the recognition test. As items are strengthened at study, targets acquire increasing amounts of mnemonic evidence and the difference in evidence between targets and foils increases (thereby increasing discriminability). Applications of SDT to memory assume that the mnemonic evidence for foils is unaffected by study.

Bias describes one's general tendency toward giving "old" or "new" responses. In other words, bias reflects the amount of mnemonic evidence an individual requires to call an item "old." This threshold for calling an item "old" is known as a *decision criterion*. For example, an individual who is more "old" biased requires less evidence to call an item "old" and therefore adopts a more liberal criterion.

Returning to the SBME, there exist two accounts for the decrease in FAR for strong items relative to weak items. The traditional explanation, here called the decisional account, assumes that changes in the FAR between strong lists and weak lists is due to metacognitive decisional processes—that is, criterion shifts (Benjamin & Bawa, 2004; Hirshman, 1995; Morrell, Gaitan, & Wixted, 2002; Stretch & Wixted, 1998). Some assume the criterion is set on the basis of the study list (e.g., Hirshman, 1995), and others assume the criterion is set on the basis of the test list (e.g., Verde & Rotello, 2007). In a standard SBME paradigm these hypotheses are indistinguishable because the strength of the targets is the same on study and test. Participants in a strong list condition ostensibly realize the quality of their memory is high and therefore require more mnemonic evidence to call an item "old" (Starns, Ratcliff, & White, 2012; Starns, White, & Ratcliff, 2010; Stretch & Wixted, 1998; Verde & Rotello, 2007). This more conservative criterion results in a lower FAR because, according to the decisional account, mnemonic evidence for foils is only determined by preexperimental familiarity (Hirshman, 1995; Parks, 1966; Stretch & Wixted, 1998). However, the assumption that the mnemonic evidence generated by foils does not change between strong and weak lists has been contested.

We will call this second explanation the mnemonic account. In short, the mnemonic account claims that the match between a test item and the contents of episodic memory produces mnemonic evidence that necessarily depends on the fidelity of the events in episodic memory (e.g., Criss, 2006, 2009, 2010; Shiffrin & Steyvers, 1997). Relative to weak lists, the contents of episodic memory traces are more complete in strong lists. As the contents of memory become more complete, any item presented at test will be less likely to match by chance alone. In other words, foils match the contents of the episodic memory traces less well in strong lists and therefore functionally produce *less* mnemonic evidence. That is, they produce more evidence of *not* being in episodic memory. This process, known as differentiation, is a fundamental characteristic of episodic memory (see Criss & Koop, 2015, for a review; McClelland & Chappell, 1998).

Although these two accounts specify very different mechanisms for the SBME, it is difficult to discriminate between them using the standard pure-strength SBME design (Starns et al., 2012; Starns et al., 2010). One strategy has been to look for mirror effects under conditions where differentiation should not occur (e.g., Hicks & Starns, 2014; Starns & Olchowski, 2015; Starns et al., 2012; Starns et al., 2010). Critically, this body of work shows that criterion shifts alone can be sufficient to produce an SBME. This literature *does not* ask if differentiation alone could also produce an SBME. This is the primary question engaged in this article. To address this question, we first review the literature on when criterion shifts occur (and when they do not).

Under what conditions do criterion shifts occur?

The standard SBME design is a pure-strength, single-item recognition task. That is, a given study–test cycle will include only strong items or weak items, but never both. Under these conditions, a mirror effect can be reliably produced (e.g., Criss, 2006, 2009, 2010; Glanzer & Adams, 1990; Hirshman, 1995; Koop & Criss, 2016; Stretch & Wixted, 1998; among others). However, decisional and mnemonic accounts are confounded in such a design (Starns et al., 2012; Starns et al., 2010). In contrast to pure-strength experiments, a mixed-strength experiment presents strong and weak items within the same study–test cycle (see Fig. 2 for a version of this design used in Experiment 3 of this article) and assumes that a criterion is set on the basis of the test. At first glance, such a mixed-strength design would seem to distinguish between decisional and mnemonic accounts because the overall degree of differentiation would be consistent for all test

items.[1] Unfortunately, there is an obvious problem with this logic—foils are not classified as strong or weak within the experimental design. A foil is by definition unstudied and not *directly* affected by any encoding manipulations.

Stretch and Wixted (1998) addressed this by explicitly cuing anticipated strength. At test, each item was presented in one of two colors. If an item was red (for example), that indicated the item was either a strongly studied target or a foil. If an item was blue, that indicated the item was either a weakly studied target or a foil. The straightforward prediction, then, was that participants should adopt a more stringent criterion for red (strong) foils than for blue (weak) foils. Although the HR was greater for strong items than for weak items, there was no difference in FAR, even when participants were explicitly alerted to the meaning of the color-cuing manipulation. Rather than item color, Morrell et al. (2002) differentially strengthened one of two semantic categories at study. However, participants again failed to show changes to FAR as a function of category strength even when explicitly told that one category would be strengthened. This led Morrell and colleagues to conclude that although possible for participants to shift criteria on an item-by-item basis, "they appear to be remarkably reluctant to do so even when they know they should, and it would be easy for them to do were they so inclined" (p. 1107). Many other studies have also failed to indicate within-list strength-based criterion shifts (e.g., Bruno, Higham, & Perfect, 2009, Experiment 1; Higham, Perfect, & Bruno, 2009, Experiment 2; Verde & Rotello, 2007).

Eliciting criterion shifts within mixed-strength lists is a fickle endeavor but there have been a few demonstrations that participants can flexibly adjust their criterion. The literature suggests two characteristics that increase the likelihood of criterion shifts. The first is that participants are explicitly provided with clearly differing strength expectations. The second is that participants have substantial time (or, more accurately, trials) to adjust the criterion when changes in the testing environment (and strength expectations) are not made explicit.

A study by Hicks and Starns (2014) demonstrates both of these principles. They had participants study a mixed-strength list. At test, strength was cued by color coding and by instruction. Participants were informed that items in red font (for example) should be judged as studied once or as not studied, and items presented in green font should be judged as having been studied four times or not studied. Following a mixed-strength study list, participants completed an 80-item test where item strength was randomly intermixed or like-strength items were grouped into blocks of varying size (40, 20, or 10 items). When like-strength items were blocked *and* participants were clearly alerted to differing strength expectations via instructions and color cues, strength-based criterion shifts (as measured by changes in FAR) were elicited. When items were color cued but presented randomly, criterion shifts were not consistently produced (only one of three experiments showed the effect).

When the like-strength blocks were provided but color cuing (and corresponding instructions) was withheld, false alarms did not differ as a function of strength of the targets in the test block. However, when blocks were 40 items in length, participants that began with a weak block showed a significantly higher FAR than those that began with a strong block (Experiment 1). This finding led the authors to conclude that "participants do not stabilize their criterion in the first 10 or 20 trials, but getting a consistent and high expectation of strength for 40 trials produces a criterion shift" (Hicks & Starns, 2014, p. 751).

Subsequent work has demonstrated an alternative, but conceptually related, means to elicit strength-based criterion shifts. Differences in FAR can be produced if participants are required to use unique responses to expected-strong and expected-weak items at test (Franks & Hicks, 2016; Starns & Olchowski, 2015). Thus, we can conclude that criterion placement requires substantial affordances: manipulations that make participants clearly aware of differences between strong and weak items (Franks & Hicks, 2016; Hicks & Starns, 2014; Starns & Olchowski, 2015), and/or numerous like-strength trials (more than 20; Hicks & Starns, 2014).

These findings fit with the general criterion shift narrative that during a typical SBME paradigm, participants set the criterion on the basis of expected test strength, which differs between pure-weak lists and pure-strong lists. Evidence from alternative bias manipulations like base rate (Estes & Maddox, 1995; Koop & Criss, 2016; Rhodes & Jacoby, 2007) or distractor similarity (Benjamin & Bawa, 2004; Brown, Steyvers, & Hemmer, 2007) support the notion that as changes in the testing environment become more apparent to the participant, criterion shifts become increasingly likely. By providing individuals with abundant affordances (e.g., very explicit cues, pure-strength blocking of significant duration) experimenters can *usually* elicit changes in FARs between weak and strong items when differentiation should be minimized.

Why are pure-strength strength-based mirror effects so reliable?

Given extensive explicit affordances, individuals shift the criterion somewhat reliably. However, the affordances needed to produce within-list criterion shifts are a clear departure from the ease with which pure-list SBMEs have been consistently

[1] Recall that differentiation occurs because test items are compared with the contents of memory acquired during study (see Shiffrin & Steyvers, 1997). In pure-strength experiments, test items are compared with either an entirely strong list (producing poorer matches) or an entirely weak list (producing better matches). Consequently, false-alarm rates will be lower for strong lists than for weak lists. In a mixed-strength design, both strong and weak items are compared back to the same (mixed-strength) contents of memory.

documented over decades of research (e.g., Criss, 2006, 2009, 2010; Glanzer & Adams, 1990; Hirshman, 1995; Koop & Criss, 2016; Stretch & Wixted, 1998). One possibility is that pure-strength lists produce an SBME so reliably because they meet the two criteria for criterion shifts that were discussed above: a sufficient number of trials to establish stable strength expectations, and awareness of expected memory quality at test (see Hicks & Starns, 2014, for a similar explanation).

The first question addressed by the experiments presented here is whether an SBME is still produced in a pure-strength design when the number of trials falls below the number of trials required for establishing a criterion. If an SBME is still evident, it would suggest that criterion shifts are not necessary to produce a mirror effect in a pure-strength single-item recognition study.

Another explanation for the consistency of the pure-strength SBME is that criterion adjustment is not necessary in a pure-strength design. After all, the 40-trial threshold for criterion setting comes from an experiment where the strength of the study list (mixed) and test list (pure) differed. Perhaps participants are much more efficient at setting criteria in pure-strength lists because expectations about memory strength do not change. This account is highly unlikely. Recent data have indicated that participants bring established expectations about memory performance into the experimental setting (Cox & Dobbins, 2011; Koop, Criss, & Malmberg, 2015; Turner, Van Zandt, & Brown, 2011). For example, HRs and FARs are shockingly similar between groups of participants presented with test lists consisting *entirely* of targets or foils and individuals in standard test lists consisting of half foils and half targets (Cox & Dobbins, 2011; Koop et al., 2015). Obviously, participants faced with a test list consisting entirely of targets should have a different understanding of the strength of items at test in an all-target list than in an all-foil list. In fact, participants only dramatically altered their responses when they were provided with feedback that indicated preexperimental expectations no longer held (Koop et al., 2015). These data demonstrated that individuals do not come into the test setting as "blank slates." Participants have a lifetime of experience making recognition decisions and have therefore developed an understanding about what is likely to be an accurate memory and what is not (Wixted & Gaitan, 2002). It is reasonable to assume these "preexperimental priors" will be maintained unless it becomes apparent to the participant they no longer hold (Turner et al., 2011). Thus, for participants to adjust their criterion in a typical SBME design, they must have an accurate understanding about the effects of different encoding manipulations. This claim will be the focus of Experiments 3 and 4.

To summarize, strength-based criterion shifts require two things: sufficient trials for establishing expectations about memory quality, and/or manipulations that make participants acutely aware of differences between strong and weak trials. Experiments 1 and 2 explored whether SBMEs persist with an insufficient number of trials to establish a criterion, while Experiments 3 and 4 examined participants' awareness of the effects of encoding manipulations. Finally, Experiment 5 combines both of these manipulations to assess memory and awareness of encoding manipulations in short study–test lists.

Experiments 1 and 2

In Experiments 1 and 2, we use lists that are so short so as to eliminate criterion shifts. The above discussion focuses on the number of test trials necessary to establish a criterion, but, of course, study trials could also help participants establish an appropriate criterion (e.g., Hirshman, 1995). To eliminate either possibility, we use both short study and test lists. The experiments are very similar and only differ in that Experiment 2 is slightly more difficult by virtue of different encoding tasks and a slightly longer delay between study and test.

Method

Participants Seventy-two introductory psychology students from Syracuse University participated in Experiment 1. Thirty-nine introductory psychology students from Eastern Mennonite University participated in Experiment 2. Participant data were excluded from analysis if they had a d' of less than 0.5 on either of the two study–test cycles (described below). This exclusion criterion resulted in removing three participants from Experiment 1, and two participants from Experiment 2. All participants received partial fulfillment of course requirements in exchange for their participation.

Design and materials In both experiments, participants completed two study–test cycles containing 10 items each. For each participant, one of these study–test cycles was strong and one was weak, with the order randomly determined across participants. Participants in Experiment 1 also completed an additional study–test cycle that followed the two short blocks examined here. These data were collected for a different research project, one focused on theoretical questions outside the scope of the present work and will therefore not be discussed further. Participants in Experiment 2 only completed the two short study–test cycles. The test phase was single-item recognition, where participants were asked to make old/new decisions on five targets and five foils. The order of targets and foils at test was randomized. Thus, the data analyzed here come from a 2 (block strength: strong vs. weak) × 2 (item type: target vs. foil) repeated-measures design.

Word stimuli were pulled from a pool of 800 high normative frequency words between four and 11 letters in length

(median = 5) and ranging between 12.99 and 9 log frequency (M = 10.46) in the Hyperspace Analog to Language Corpus (Balota et al., 2007). For each participant, a subset of 30 words were selected and randomly assigned to condition. Stimulus presentation and recording of responses were conducted with the Psychtoolbox add-on for MATLAB (Brainard, 1997; Kleiner et al., 2007).

Procedure Upon arrival to the experiment, all participants were given an informed consent form. Next, participants read instructions that informed the participants that they would study a list of words and later have their memory for those words tested. At study, participants were given either a weak or strong encoding prompt for each trial depending on the strength of that particular cycle. The weak encoding task for Experiment 1 asked participants to indicate whether or not the word contained the letter *e*. For strong encoding trials, participants indicated whether or not they considered the word to be pleasant (Fig. 1).

In Experiment 2, the weak encoding task asked participants whether or not the stimulus was written in red, whereas the strong encoding task asked participants whether the stimulus was easy to imagine. For all prompts, yes and no responses were indicated by a single key press, ("z" or "?" key, counterbalanced between participants). The study phase was self-paced with the lone constraint that a response could not be entered until a minimum of 1.5 seconds after stimulus presentation.

After each study phase, there was a math distractor task. In Experiment 1, this task lasted for 45 seconds, whereas in Experiment 2 it lasted for 90 seconds. After completing the distractor task, participants were given instructions for the test phase. At test, participants were presented with a single test word and asked to indicate whether that word was "old" (previously studied word) or "new" (word that was not studied). Participants indicated their choice by clicking on "old" or "new" response boxes located in the upper left and right corners of the screen (left/right order was counterbalanced between participants). After providing a recognition judgment, the stimulus disappeared, and participants clicked on a start button at the bottom of the screen to begin the next trial. Following the first recognition test, participants were allowed to take a short break (if necessary) prior to beginning the second study–test cycle. After completing all tests, participants were asked if they had any questions and were thanked for their participation.

Results

All analyses were conducted using JASP (JASP Team, 2018). In addition to reporting standard frequentist statistics we also report Bayes factors (BF). Bayes factors provide a continuous

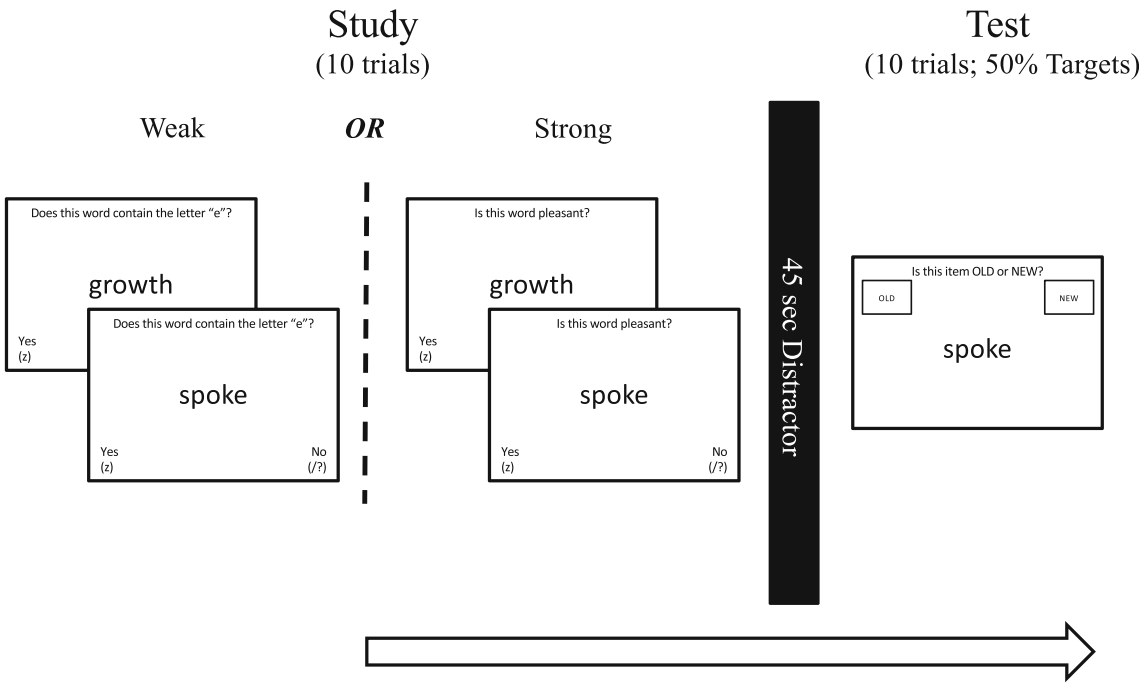

Fig. 1 Pure-strength study–test cycle in Experiment 1. Participants completed either 10 items using a weak encoding task (two leftmost study panels) or a strong encoding task (two rightmost study panels), followed by a 45 second distractor task, and 10 single-item test trials. A second study–test cycle followed the first and used whichever encoding task participants did not see during Block 1. Experiment 2 used the same design with different encoding tasks and a longer distractor. In the weak condition, participants were asked, "Is this word written in red?" In the strong condition, participants were asked, "Is this word easy to imagine?"

estimate of relative evidence. In particular, as presented here BF_1 indicate the ratio of evidence for the model with an effect compared with the null model. Values greater than 1 indicate evidence for a model with an effect, and values below 1 indicate support for the null model.

In order to explore whether a strength-based mirror effect was obtained under conditions where criterion shifts would not be expected, we first conducted a 2 (strength: strong vs. weak) × 2 (trial type: target vs. foil) repeated-measures ANOVA. In Experiment 1, the data (see Table 1) demonstrated a Strength × Trial Type interaction, $F(1, 68) = 14.80$, $p < .001$, $\eta_p^2 = .18$, $BF_1 = 72.26$.[2] This interaction was due to an increase in HR from weak to strong conditions accompanied by a decrease in FAR. Planned comparisons confirmed an increase in HR from weak ($M = .88$, $SE = .02$) to strong ($M = .97$, $SE = .01$) conditions, $t(68) = 3.41$, $p = .001$, $d = .41$, $BF_1 = 23.17$. There was a numerical decrease in FAR between weak ($M = .07$, $SE = .01$) and strong ($M = .05$, $SE = .01$) conditions, but it was not statistically significant, $t(68) = 1.35$, $p = .182$, $d = .16$, $BF_1 = .31$.

In Experiment 2, the data again demonstrated a Strength × Trial Type interaction, $F(1, 35) = 59.43$, $p < .001$, $\eta_p^2 = .62$, $BF_1 = 1.86e + 9$. Strong lists elicited a higher HR and a lower FAR than weak lists. Planned comparisons again confirmed an increase in HR from weak ($M = .72$, $SE = .03$) to strong ($M = .98$, $SE = .01$) lists, $t(36) = 7.51$, $p < .001$, $d = 1.24$, $BF_1 = 1.78e + 6$. Unlike Experiment 1, FARs were also reliably lower for strong lists ($M = .03$, $SE = .01$) relative to weak lists ($M = .08$, $SE = .02$), $t(36) = 2.05$, $p = .048$, $d = 0.34$, $BF_1 = 1.14$.

Generally speaking, this experimental design is a challenge because limiting memory to such a short list necessarily results in near ceiling performance, making it difficult to detect changes in FAR. Therefore, we performed an exploratory analysis that collapsed across these two highly similar studies to see whether the associated increase in power would provide clarity, especially with regard to the FAR. As expected, the 2 (experiment) × 2 (strength) × 2 (trial type) mixed-factors ANOVA revealed a three-way interaction, $F(1, 104) = 17.50$, $p < .001$, $\eta_p^2 = .14$, $BF_1 = 229.89$. As is apparent from Table 1, this interaction is the product of the typical SBME interaction being more pronounced in Experiment 2 than in Experiment 1. However, the direction of the interaction is identical. This combined data set demonstrates a reliable difference between strong ($M = .04$, $SE = .01$) and weak ($M = .07$, $SE = .01$) FAR, $t(105) = 2.26$, $p = .026$, $d = 0.22$, $BF_1 = 1.22$, and strong ($M = .97$, $SE = .01$) and weak ($M = .83$, $SE = .02$) HR, $t(105) = 6.75$, $p < .001$, $d = 0.66$, $BF_1 = 1.18e + 7$.

[2] Here, the BF represents support for an interaction model (Strength × Trial Type) relative to a model only containing main effects.

Table 1 Hit and false-alarm rates in Experiments 1 and 2

	Hit rate		False-alarm rate	
	Weak	Strong	Weak	Strong
Experiment 1	.88 (.02)	.97 (.01)	.07 (.01)	.05 (.01)
Experiment 2	.72 (.03)	.98 (.01)	.08 (.02)	.03 (.01)

Note. Standard error in parentheses

Discussion

We observed the descriptive SBME pattern of higher HR and lower FAR for strongly encoded lists indicating that the SBME can be elicited even under conditions where criterion shifts are highly unlikely. However, the small magnitude of the BFs suggest that the evidence for differences in the FAR was not strong. We see these experiments as a "proof of concept" and will return to a more rigorous (and preregistered) evaluation of this short-list SBME in Experiment 5.

Although a substantial body of literature suggests it takes more than 10 trials to establish firm expectations about memory quality at test, it could be possible that participants were able to establish expectations about memory quality in the short study lists. In other words, it is possible that participants quickly established an accurate expectation about the strength of the upcoming test list (but see Turner et al., 2011, for evidence that participants bring preexisting memory expectations into the experimental context). This would require that participants can somewhat accurately evaluate memory fidelity for individual study items. In Experiments 3 and 4, we provide an empirical test of how these expectations develop over the course of study and test, using list lengths that are more typical of SBME experiments.

Experiments 3 and 4

To best address how test expectations develop, we look to studies assessing metaknowledge. Benjamin (2003) examined individuals' expectations regarding the recognizability of low-frequency and high-frequency words. During study, Benjamin asked participants to rate the likelihood that they would recognize a studied item on the subsequent memory test. Three separate studies indicated that, in general, participants *incorrectly* expected that they would have better memory for high-frequency words than for low-frequency words. The inability of participants to grasp—prior to test—the effects of word frequency on recognition lead Benjamin (2003) to speculate that people may often have "poor self-assessment of one's own memory ability and, by extension, of the effects of different variables on one's memory" (p. 304). Obviously,

if such a claim were true in the standard SBME paradigm, this would question the viability of criterion shifts to produce all SBMEs. It seems reasonable to assume that people are aware that repetition helps memory. However, an SBME is observed not just when items are strengthened by repetition but also through levels of processing, as in Experiments 1 and 2 (see also Glanzer & Adams, 1990; Kiliç, Criss, Malmberg, & Shiffrin, 2017; Koop & Criss, 2016). We collect memory predictions in mixed-strength study lists (Experiment 3) and pure-strength study lists (Experiment 4) to establish whether participants are aware that different encoding conditions lead to different levels of subsequent memory.

Method

Participants Introductory psychology students participated in Experiments 3 and 4. Thirty-one students from Eastern Mennonite University participated in Experiment 3 in exchange for partial course credit. Forty-four students from Syracuse University participated in Experiment 4 and were compensated with partial course credit. As in the first two experiments, all participants that did not achieve a d' above 0.5 were excluded from analyses. This resulted in excluding one participant from Experiment 3, and one participant from Experiment 4.

Design and materials In Experiments 3 and 4, participants completed two study–test cycles. Study lists consisted of 30 words, and test lists consisted of 60 words (30 targets and 30 foils). In Experiment 3, we presented mixed-strength study lists. In each study list, half of the words were presented with the weak encoding task ("Does this word contain the letter *e*?") and half of the words were presented with the strong encoding task ("Is this word pleasant?"). Strong and weak items were randomly intermixed at study and test. Study lists in Experiment 4 were pure strength and the encoding tasks were identical to Experiment 3. Each participant completed one weak study–test cycle and one strong study–test cycle. The order in which participants encountered strong and weak blocks was randomly assigned across subjects.

Stimuli were pulled from a pool of 424 medium normative frequency words between three and 13 letters in length (median = 6) and ranging between 13.22 and 5.19 log frequency (M = 8.87) in the Hyperspace Analog to Language Corpus (Balota et al., 2007). For each participant, a subset of 180 words was randomly selected from this pool and randomly assigned to strength condition (weak vs. strong). Stimuli were presented and responses were recorded using the Psychtoolbox add-on for MATLAB (Brainard, 1997; Kleiner et al., 2007).

Procedure The procedure for Experiments 3 and 4 is depicted in Fig. 2. First, participants were instructed that they would be asked to study lists of words and later complete a test of their memory for those words. Participants completed two study–test cycles. The critical addition to Experiments 3 and 4 was that we also collected participants' predictions about their ability to later recognize each studied word (1 = *I won't recognize*, 9 = *I will recognize*; Benjamin, 2003). These predictions were collected on each study trial immediately after participants responded to the encoding task. All other details for the study phase and subsequent distractor task were identical to Experiment 1.

The test phase in Experiment 3 was procedurally identical to that of the previous experiments, with the exception of length and that study was mixed strength. Experiment 4 had one additional change. Because participants experienced pure-strength lists in Experiment 4, it was possible to collect weak and strong postdictions at test (Benjamin, 2003). Whenever participants in Experiment 4 provided a "new" response at test, they were asked to respond to the question "How likely would you have been to remember this word if you had actually studied it?" by providing a rating on a 1–9 scale (1 = *I am sure I would NOT recognize this word*; 9 = *I am sure I WOULD recognize this word*). Each test word remained on-screen during the postdiction phase.

Results

We first examined the participants' accuracy data (see Table 2) to verify that the encoding manipulation had the expected effect. In Experiment 3, participants showed a higher HR to strongly encoded items (M = .94, SE = .01) than to weakly encoded items (M = .82, SE = .02), $t(29) = 6.36$, $p < .001$, d = 1.16, BF_1 = 28795.70. The mixed-strength design of Experiment 3 means that it is not possible to compare weak and strong FAR. For Experiment 4, we conducted a 2 (strength: weak vs. strong) × 2 (trial type: target vs. foil) repeated-measures ANOVA. As expected, there was the Strength × Trial Type interaction that is characteristic of an SBME, $F(1, 42) = 47.92$, $p < .001$, $\eta_p^2 = .53$, $BF_1 = 1.69e+5$. Strong HRs (M = .94, SE = .01) were higher than weak HRs (M = .85, SE = .02), $t(42) = 6.10$, $p < .001$, d = .93, BF_1 = 54362.28, whereas strong FARs (M = .07, SE = .01) were lower than weak FARs (M = .14, SE = .02), $t(42) = 3.73$, p = .001, d = .57, BF_1 = 49.15.

Having confirmed that the encoding manipulations had the intended effect, we turn attention to the question of whether participants were *aware* of how encoding would affect later memory (see Fig. 3). In other words, did participants expect to have better memory for items presented with the strong versus weak encoding task?

Experiment 3 had a single study list with strong and weak encoding tasks intermixed at study. We evaluated participants' predicted recognizability for all strongly studied items and all weakly studied items. Ratings of strongly studied items (M =

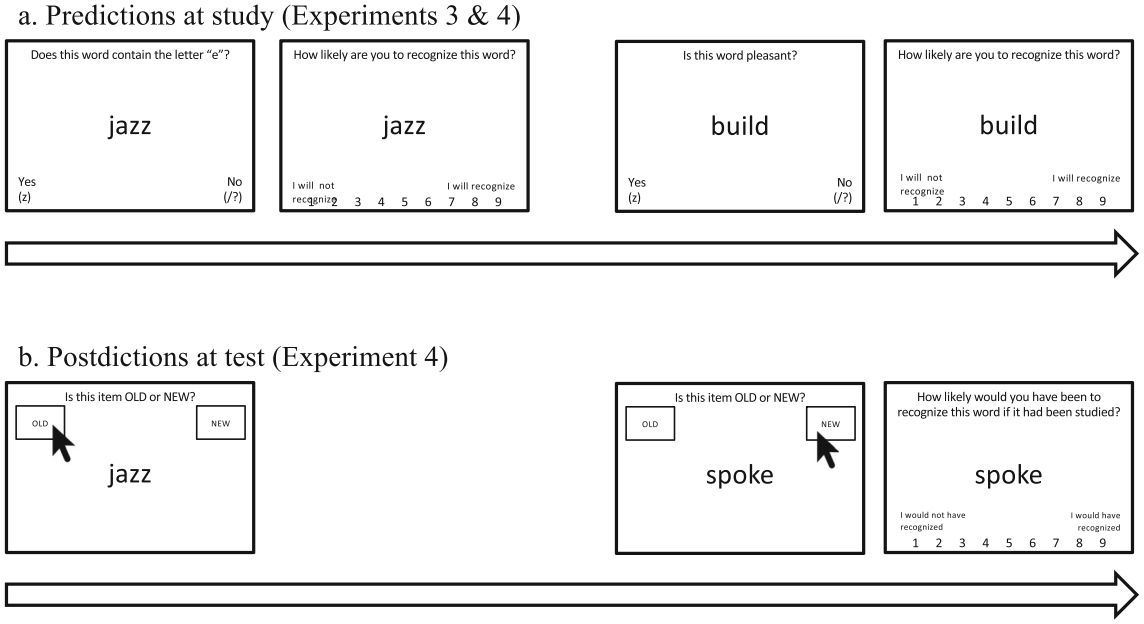

Fig. 2 Procedure for collecting predictions (top panel; Experiments 3 & 4) and postdictions (bottom panel; Experiment 4 only). Predictions were collected following every trial during the study phase. Postdictions were only collected following a "NEW" response. In the bottom panel, no postdiction is collected for "jazz," because the participant identified it as an old word. A postdiction is collected for "spoke" because it was identified as a new word

6.84, $SE = 0.26$) did not reliably differ from ratings of weakly studied items ($M = 6.79$, $SE = 0.29$), $t(29) = 0.61$, $p = .550$, $d = 0.11$, $BF_1 = 0.23$. In Experiment 4, encoding task was manipulated between lists. We again compared predicted recognizability for strong and weak items. Participants in Experiment 4 showed slightly higher predictions for strong items ($M = 6.34$, $SE = 0.22$) than for weak items ($M = 6.04$, $SE = 0.25$), $t(42) = 2.06$, $p = .046$, $d = 0.31$, $BF_1 = 1.12$.

We also collected postdictions in Experiment 4. Each time participants provided a "new" response at test, they were asked how likely they would have been to remember that word if it had actually been presented. Participants showed greater postdicted confidence for strong correct rejections ($M = 5.95$, $SE = 0.27$) than for weak correct rejections ($M = 5.64$, $SE = 0.27$), $t(42) = 2.84$, $p = .007$, $d = 0.43$, $BF_1 = 5.41$. However, there was not a reliable difference between strong misses ($M = 4.97$, $SE = 0.49$) and weak misses ($M = 5.00$, $SE = 0.32$), $t(22) = 0.20$, $p = .84$, $d = 0.04$, $BF_1 = 0.22$. Notably, the average confidence ratings for misses are lower than those of correct rejections. This means that participants indicated that they would be more likely to remember unstudied foils than *items they actually studied*. One distinct possibility is that participants simply reported confidence in their response rather than postdiction judgments (or they conflated the two). A reviewer suggested an interesting alternative interpretation. He suggested that lower ratings for misses than correct rejections indicates that metamemory judgments are quite accurate in the sense that participants accurately indicate that they would not have remembered exactly those items that they forgot (e.g., misses). In either event, these results are not informative for the present research question without additional research.

Table 2 Hit and false-alarm rates in Experiments 3 and 4

Block order and experiment	Hit rate		False-alarm rate	
	Weak	Strong	Weak	Strong
Experiment 3 (mixed-strength)	.82 (.02)	.94 (.01)	.08 (.02)	
Experiment 4 (pure-strength)	.85 (.02)	.94 (.01)	.14 (.02)	.07 (.01)

Note. Strong and weak study trials are intermixed in Experiment 3, therefore it is not possible to define weak or strong false alarm rates. Standard error in parentheses

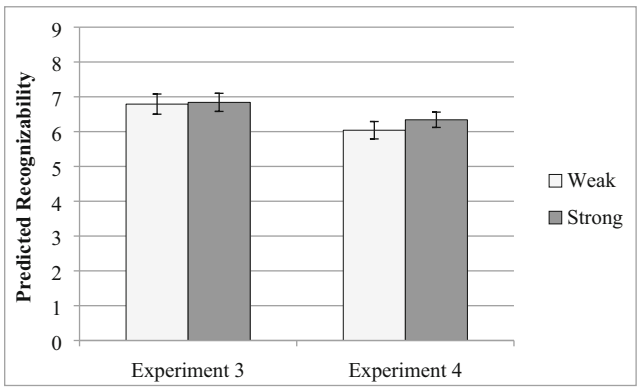

Fig. 3 Mean predicted recognizability for studied items in Experiments 3 (mixed-strength lists) and 4 (pure-strength lists). Error bars are ±1 *SE*

Discussion

The goal of Experiments 3 and 4 was to assess participants' awareness of the effects of different encoding manipulations. We investigated whether participants have accurate expectations about the effects of encoding manipulations on memory quality, as predicted by the decisional account. Remarkably, the data suggest no difference in predicted recognizability between weak encoding tasks and strong encoding tasks in Experiment 3 even though participants' accuracy was dramatically higher for strong items. It appears as though participants do not notice clear distinctions in memory quality at a trial-by-trial level. Experiment 4 did show a difference between predicted recognizability for strong and weak items. However, the small magnitude of the BF suggests minimal evidence for the observed difference in the predicted memory. We will return to this with a large *N*, preregistered evaluation of memory expectations in Experiment 5.

Interestingly, we note that hit rates in Experiments 3 and 4 are extremely similar. However, the ratings of predicted memorability are higher for Experiment 3 than Experiment 4. In other words, quite different memorability ratings do not correspond to differences in accuracy. This is another indication that participants are not well calibrated with respect to assessing future memory or how encoding might affect later memory.

Our measure of predicted memory is somewhat similar to judgments of learning (JOL). Current theorizing on JOLs attributes them to fluency in processing the individual items and beliefs about what affects memory (e.g., Dunlosky, Mueller, & Tauber, 2015; Koriat, 1997). The role of beliefs has largely been tested in terms of properties of the items (e.g., font size), whereas here we manipulated the encoding task that is common to all items. Consistent with prior work, memory predictions do not reflect differences in quality between encoding tasks like those used here. For example, Begg, Duft, Lalonde, Melnick, and Sanvito (1989) provided individuals with interactive or separate imagery-based encoding tasks. Although memory performance differed between groups, memory predictions did not.

Recall that strength-based criterion shifts require two things: sufficient trials for adjustment and accurate expectations about memory quality on the part of participants. Experiments 1 and 2 showed preliminary evidence for an SBME with insufficient trials to establish a criterion. Experiments 3 and 4 showed that participants do not have accurate expectations about memory quality. In our final study, we collect additional data about participants' memory expectations by asking a single question about predicted memory for the entire set of targets. After all, it is possible that a post-study estimate might better characterize expectations about the encoding conditions absent judgments about the specific target items.

Experiment 5

In Experiment 5, we combine the short list design of Experiments 1 and 2 while also assessing participants' awareness of the specific encoding manipulations used therein. If participants continued to show an SBME while simultaneously failing to note the mnemonic consequences of weak and strong encoding tasks, then a pure criterion shift account of all SBMEs would be highly unlikely (assuming the standard assumption that criterion shifts are active control processes). On the other hand, if participants accurately assess differences in memory strength even after short study lists, this could provide grounds for revisiting assumptions about the speed with which criterion shifts can occur. Experiment 5 was a preregistered study that was identical to two additional experiments (5a and 5b) that appeared in a previous draft of this manuscript. Results from those studies can be found in the supplementary materials posted at (https://osf.io/bv6c3/). Preregistration for Experiment 5 can also be found there.

Method

Participants In order to determine our sample size for Experiment 5, we performed a power analysis using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007). We selected a sample size of 120 participants because it would give us above $1 - \beta = .9$ with an effect size of $d = .3$ (roughly the effect size on FAR from Experiment 2). To ensure that we would accrue 120 participants after no-shows, cancellations, and exclusions, we posted many more than 120 sessions at Syracuse University and Eastern Mennonite University. In total 176 individuals participated in the experiment.[3] All participants completed the experiment before we looked at any data. Participants were compensated with partial fulfillment of course requirements.

Design and materials Experiment 5 is a complete replication of Experiment 2, with the addition of a single question about the quality of participants' memory immediately prior to the test phase (described more fully below).

Procedure Participants received the same instructions and completed the same study–test cycles as in Experiment 2.

[3] G.K. and A.H.C. had an interesting conversation about whether to report the first 120 participants so as to remain faithful to the preregistered sample size or to report all participants. In the end, we agreed that it was not sensible to discard the contribution of a large number of volunteers because we pessimistically scheduled too many appointments. As Alexander DeHaven wrote, "preregistration is a plan not a prison" (https://cos.io/blog/preregistration-plan-not-prison/). Finally, note that the pattern of results do not change with the smaller sample size.

Following the 90-second distractor task and test instructions, participants were asked an additional question about their perceived memory quality. The question was as follows:

> The test will be 10 words in length. Before starting the test, we would like you to estimate how well you will do. If you believe you will give the correct "OLD" or "NEW" response to all 10 items, you would type "10." If you feel like you will be completely guessing, you can expect to get around five answers correct, and should enter "5."

Following this question, participants then completed the 10 single-item recognition test trials just as described in Experiment 2. After completing both a strong study–test cycle and a weak study–test cycle (counterbalanced across participants), participants were thanked and then dismissed.

Results

Participants were excluded using the same criterion ($d' < .5$ in each study–test cycle) as the previous studies. This resulted in excluding 22 participants. Additionally, one individual gave a memory prediction response outside of the 0–10 scale and was therefore excluded from analysis. In total, data from 153 participants were analyzed.

We analyzed accuracy data using a 2 (strength: strong vs. weak) × 2 (trial type: target vs. foil) repeated-measures ANOVA and paired-samples *t* tests. Predicted recognition performance following weak and strong study lists was compared using a paired-samples *t* test. These analyses were preregistered. In addition, we included Bayes factors for all analyses, which we did not preregister.

Participants showed a Strength × Trial Type interaction, $F(1, 152) = 193.98$, $p < .001$, $\eta_p^2 = .56$, $BF_1 = 8.37e + 30$. HRs were higher on strong blocks ($M = .97$, $SE = .01$) than on weak blocks ($M = .75$, $SE = .02$), $t(152) = 13.48$, $p < .001$, $d = 1.09$, $BF_1 = 3.64e +24$. There was also a reliable difference in FAR between strong ($M = .04$, $SE = .01$) and weak blocks ($M = .10$, $SE = .01$), $t(152) = 4.26$, $p < .001$, $d = 0.34$, $BF_1 = 412.29$. Thus, the SBME was present for short study and test blocks.

An analysis of participants' predicted memory showed a small difference between predictions following a strong block ($M = 7.29$, $SE = 0.16$) and those following a weak block ($M = 6.94$, $SE = 0.16$). The statistical analysis of this effect is mixed, $t(152) = 2.05$, $p = .042$, $d = 0.17$, $BF_1 = 0.70$, with the *p* value indicating support for this difference and the BF indicating no evidence for an effect and in fact weak evidence for a null effect.

Discussion

Experiment 5 was designed to assess whether individuals show an SBME under conditions where a criterion shift is unlikely and without demonstrating awareness of differing memory quality between the two encoding tasks. The results from Experiment 5 clearly showed strong evidence for an SBME under conditions where criterion shifts would not be expected. Whether participants are aware that different encoding tasks result in differences in memory quality is ambiguous. Given the strong evidence of an SBME, it is particularly striking that there was trivial evidence for differences in the predicted memory. If participants are basing a criterion shift on the outcome of encoding, then the cognitive system must be magnifying small differences in expected memory to rather large differences in the decision space.

General discussion

The experiments presented here have demonstrated that it is possible to elicit an SBME under conditions inhospitable for criterion shifts. We observed an SBME even when participants had few items on which to establish a criterion. We used encoding manipulations that participants did not consistently think affected memory. We demonstrated these findings in the first four experiments and then combined them in a preregistered study with a large sample size. Collectively, this demonstrates the presence of an SBME under conditions where criterion shifts would not be predicted. Further, participants do not accurately adjust their expectations about memory quality in response to levels of processing manipulations. Even if participants could set a specific criterion for the list after exposure to only a few items, they would set an inappropriate criterion (to generate an SBME) because they estimate that encoding tasks produce minimal differences in memory accuracy.

Prior research has established that an SBME can be found when differentiation is not present. This, of course, is consistent with all models of memory because no models dispute the possibility of a criterion or the ability of participants to modify a criterion to suit the needs of the individual or context. Here, we find that an SBME is observed under conditions that would not seem to support a criterion shift of the sort required to produce an SBME. This does not imply that differentiation is responsible for the pattern of HR and FAR; there could be some alternative mechanism that produces an SBME. However, our research does suggest that a criterion shift is unlikely to be the single mechanism responsible for this pattern of data.

Implications for strength-based mirror effects

Because our aim in the present work was to address the debate between decisional and mnemonic accounts of SBMEs, we have ignored a more nuanced perspective on criterion shifts. Recent work has raised the possibility that individuals can adjust decision criteria on an item-by-item basis, but making the decision as to what strength to expect is arduous (Starns & Olchowski, 2015). Thus, rather than a failure to elicit criterion shifts, much of the literature reviewed in the introduction could potentially be framed as failures to effectively manipulate strength *expectations*.

Starns and Olchowski (2015; see also Franks & Hicks, 2016) produced item-by-item shifts in FARs by making encoding strength covary with the side of the screen on which items were presented and requiring participants to use different response keys for strong and weak items. For example, when an item was presented on the right side of the screen, it was either a strong target or a foil, whereas items presented on the left side of the screen were either weak targets or foils. Providing an "old" response then required the use of different keys for items presented on the left or right side of the screen.

While this work may very well demonstrate that participants often do not shift their strength expectations (rather than criteria) in typical mixed-strength recognition studies, it does not affect the interpretation of our results. First, the fundamental assumption is that without external affordances, participants take time to adapt to a new decision environment. Based on this work, one presumes that the literature demonstrating that criterion shifts take time (e.g., Brown & Steyvers, 2005; Hicks & Starns, 2014; Verde & Rotello, 2007) could easily be reframed as slow decisions to adopt different strength expectations. However, the fact remains that differences in FAR are only observable on an item-by-item basis when significant affordances are provided. The affordances may include things like color cuing (Hicks & Starns, 2014; Starns & Olchowski, 2015; Stretch & Wixted, 1998), pure-strength blocking at test (Hicks & Starns, 2014; Starns et al., 2012; Verde & Rotello, 2007), or forcing individuals to acknowledge strength distributions through different response keys (Starns & Olchowski, 2015). In short, without explicit cues, it takes time for people to adapt to changing strength environments. Thus, our results challenge the notion that any adjustment of memory strength expectation occurred during the shortened pure-strength SBME design.

Again, one might ask the question: If it is so hard to elicit this adjustment (whether criterion shift or decision about expected strength), why is it that pure-strength lists reliably produce an SBME without going to any great lengths to alert (or force) participants to acknowledge differences in strength between lists? One possibility is that the encoding tasks provide this explicit affordance. After all, if items are strengthened by repetition, participants "should certainly expect to have better memory after an entire list of words studied five times than after an entire list of words studied once" (Starns & Olchowski, 2015, p. 57). However, all the experiments presented above used less intuitive strength manipulations than mere repetition. Experiments 3, 4, and 5 directly tested this assumption. Those data suggested that participants *do not* consistently form dramatically different expectations for strong encoding tasks and weak encoding tasks. One possible explanation for why repetition leads to clear memory expectations, whereas our strength manipulations do not, is because memory expectations are influenced by the ease with which study items are processed (Begg et al., 1989). Repetition manipulations make items easier to process and therefore have a much larger impact on memorability expectations than do the manipulations used in this work. For example, both strong and weak tasks are relatively easy for participants to perform and therefore do not have significant effects on expected memorability. This account is very speculative, and future work could examine this possibility in greater detail. In summary, although the results shown by Starns and Olchowski (2015) certainly demonstrated that criterion shifts can produce a mirror effect, it is premature to assume that such item-by-item criterion shifts underlie *all* demonstrations of SBMEs.

Typically, data demonstrating SBMEs have fallen into two competing (and often mutually exclusive) camps: the decisional account and the mnemonic account. This debate has generated a significant amount of data. Some of these data have produced an SBME under conditions not predicted by the mnemonic account, whereas the present data produced an SBME under conditions not predicted by the decisional account. After surveying this literature, we believe it will be most fruitful to take a "both/and" approach to the SBME rather than "either/or." Taken as a whole, the literature indicates that neither a pure-decisional nor a pure-mnemonic account explains all SBMEs. We also suggest that these data should lead to changes in terminology. Rather than speaking of *the* SBME as if this is a single phenomenon, these results suggest it is more appropriate to discuss *an* SBME.

Implications for memory theory

While SBMEs have produced a significant amount of research, the effect is only interesting insofar as it tell us something about the nature of the memory processes used to produce it. In other words, "What drives the SBME?" is not a particularly important question, whereas "What do SBMEs tell us about the nature of memory?" is. Focusing too narrowly on SBMEs may lead to an increasingly compartmentalized memory literature—a broader problem that has led to a fair amount of handwringing (Criss & Howard, 2015; Hintzman, 2011; Malmberg, 2008). In short, theorists have raised concerns about the generalizability of memory models. At times it

appears as though accounts are created ad hoc for individual tasks even though similar mechanisms should underlie performance of the memory system in a number of domains. We briefly highlight this debate because one argument occasionally used to support a pure-decisional account is that it is parsimonious or commonsensical. Although a pure-decisional account may seem to be a relatively straightforward explanation of SBMEs (though differentiation is not a particularly complex explanation either), we see it as contributing to the fractured landscape of memory models.

The critical question, then, is whether these data advance our broad understanding of memory, or are we simply "asking what causes some characteristic twitch in the data" (Hintzman, 2011, p. 267). Our hope is that rather than merely cataloguing memory effects and developing ad hoc models to explain them, we can identify models that do reasonably well at accommodating data from a wide variety of tasks. In the present context, this means identifying a model that can accommodate the "both/and" approach rather than only a "pure decisional" or "pure mnemonic" approach. In other words, a model should include a mechanism for decisional processes (i.e., an adjustable criterion) as well as incorporating mnemonic processes like differentiation exactly as suggested by Atkinson and Shiffrin (1968).

We briefly highlight the retrieving effectively from memory (REM; Shiffrin & Steyvers, 1997) framework for its ability to use common mechanisms to produce human data across a variety of tasks (Malmberg, 2008). Unlike signal detection theory, REM is a true process model that provides an account of how memories are encoded, stored, and retrieved. For example, versions of REM have been applied to recognition (Shiffrin & Steyvers, 1997), free recall (Lehman & Malmberg, 2013), and cued recall (Diller, Nobel, & Shiffrin, 2001; Wilson & Criss, 2017). Although this work does not necessarily fulfill Hintzman's (2011) call to study memory "in the wild," we believe the success of REM across a number of different experimental tasks begins to speak to more general characteristics of memory, like differentiation. Concerning SBMEs, REM incorporates both decisional processes and differentiation, which could conceivably cover the breadth of data collected on the SBME to date.

Finally, we see the present work as addressing a common problem also noted by Atkinson and Shiffrin (1968)—specifically, the fact that multiple control processes may give rise to similar patterns of memory performance. For example, many SBME studies used fairly transparent strengthening operations like repetition. Use of this type of strengthening task most likely elicits metacognitive criterion shifts that are absent for encoding processes like those used in the present experiments. It has taken a sizable amount of research to establish that some SBMEs may be the product of repetition strategies and subsequent criterion shifts, whereas others (like depth of processing) may somewhat automatically elicit an SBME via differentiation. While the present work indicates that a depth of processing manipulation produces SBMEs without a criterion shift, it is still unclear exactly what people are doing during such encoding tasks. Future work will need to more effectively model what, exactly, individuals are doing during such manipulations.

Acknowledgements Data and supplementary materials are available at (https://osf.io/bv6c3/).

Experiment 5 is preregistered at (https://aspredicted.org/b9ne5.pdf). We thank the following students for assistance collecting data: Michael Austin, Lara Weaver, Andrew Peltier, Sophi Hartman, and Olivia Dalke.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 89–195). New York, NY: Academic Press.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459.

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language*, *28*, 610–632.

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition*, *31*, 297–305.

Benjamin, A. S., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language*, *51*, 159–172.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Brown, S., & Steyvers, M. (2005). The dynamics of experimentally induced criterion shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 587–599.

Brown, S., Steyvers, M., & Hemmer, P. (2007). Modeling experimentally induced strategy shifts. *Psychological Science*, *18*, 40–45.

Bruno, D., Higham, P. A., & Perfect, T. J. (2009). Global subjective memorability and the strength-based mirror effect in recognition memory. *Memory & Cognition*, *37*, 807–818.

Cox, J. C., & Dobbins, I. G. (2011). The striking similarities between standard, distractor-free, and target-free recognition. *Memory & Cognition*, *19*, 925–940.

Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory and Language*, *55*, 461–478.

Criss, A. H. (2009). The distribution of subjective memory strength: List strength and response bias. *Cognitive Psychology*, *59*, 297–319.

Criss, A. H. (2010). Differentiation and response bias in episodic memory: Evidence from reaction time distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 484.

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

Criss, A. H., & Koop, G. J. (2015). Differentiation in episodic memory. In J. Raaijmakers, A. H. Criss, R. Goldstone, R. Nosofsky, & M. Steyvers (Eds.), *Cognitive modeling in perception and memory: A*

Festschrift for Richard M. Shiffrin (pp. 112–125). New York, NY: Psychology Press.

Diller, D. E., Nobel, P. A., & Shiffrin, R. M. (2001). An ARC–REM model for accuracy and response time in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(2), 414–435.

Dunlosky, J., Mueller, M. L., & Tauber, S. K. (2015). The contribution of processing fluency (and beliefs) to people's judgments of learning. In D. S. Lindsay, A. P. Yonelinas, H. I. Roediger, D. S. Lindsay, A. P. Yonelinas, & H. I. Roediger (Eds.), Remembering: Attributions, processes, and control in human memory: Essays in honor of Larry Jacoby (pp. 46–64). New York, NY: Psychology Press.

Estes, W. K., & Maddox, W. T. (1995). Interactions of stimulus attributes, base rates, and feedback in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 1075–1095.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Franks, B. A., & Hicks, J. L. (2016). The reliability of criterion shifting in recognition memory is task dependent. *Memory & Cognition*, *44*, 1215–1227.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 5–16.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review*, *100*(3), 546–567.

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

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*, 742–754.

Hintzman, D. L. (2011). Research strategy in the study of memory: Fads, fallacies, and the search for the "coordinates of truth". *Perspectives on Psychological Science*, *6*(3), 253–271.

Higham, P. A., Perfect, T. J., & Bruno, D. (2009). Investigating strength and frequency effects in recognition memory using Type-2 signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 57–80.

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.

JASP Team. (2018). JASP (Version 0.9)[Computer software]. Retrieved from https://jasp-stats.org/

Kiliç, A., Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2017). Models that allow us to perceive the world more accurately also allow us to remember past events more accurately via differentiation. *Cognitive Psychology*, *92*, 65–86.

Kleiner, M., Brainard, D. H., Pelli, D. G., Broussard, C., Wolfe, T., & Niehorster, D. (2007). What's new in Psychtoolbox-3? *Perception*, *36*(14), 14.

Koop, G. J., & Criss, A. H. (2016). The response dynamics of recognition memory: Sensitivity and bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(5), 671–685.

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.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126,* 349–370.

Lehman, M., & Malmberg, K. J. (2013). A buffer model of memory encoding and temporal correlations in retrieval. *Psychological Review*, *120*(1), 155–189.

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

Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology*, *57*(4), 335–384.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *105*(4), 724–760.

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

Parks, T. E. (1966). Signal-detectability theory of recognition-memory performance. *Psychological Review*, *73*, 190–204.

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.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*, 145–166.

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.

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

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

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

Turner, B. M., Van Zandt, T., & Brown, S. (2011). A dynamic stimulus-driven model of signal detection. *Psychological Review*, *118*, 583–613.

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition*, *35*, 254–262.

Wilson, J. H., & Criss, A. H. (2017). The list strength effect in cued recall. *Journal of Memory and Language*, *95*, 78-88.

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*(4), 289–305.