August 2014

An Investigation of the Basis of the Strength-Based Criterion-Shift

James E. Olchowski
University of Massachusetts Amherst

An investigation of the basis of the strength-based criterion-shift

A Thesis Presented

by

JAMES E. OLCHOWSKI

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE

May 2014

Psychology

Cognitive Psychology

An investigation of the basis of the strength-based criterion-shift

A Thesis Presented

By

JAMES E. OLCHOWSKI

Approved as to style and content by:

_____
Jeffrey J. Starns, Chair

_____
Erik Cheries, Member

_____
Caren Rotello, Member

_____
Melinda Novak, Department Head
Psychology

ABSTRACT

AN INVESTIGATION OF THE BASIS OF THE STRENGTH-BASED CRITERION-SHIFT

MAY 2014

JAMES E. OLCHOWSKI, B.A., MCGILL UNIVERSITY

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

In recognition memory, participants often fail to change their criterion for making a "studied" response from one trial to the next based on learning strength, even when they are given obvious cues to identify each test item as studied often ("strong") or studied a single time ("weak") (e.g., Stretch & Wixted, 1998). In three experiments we tested the hypothesis that participants produce robust item-by-item shifts only when responding did not involve significant response interference (Simon, Acosta, Mewaldt, & Speidel, 1976). In our three experiments, participants studied lists of words studied once (weak) or five times (strong). In Experiment 1, both strong and weak words appeared at test under the questions "Was this studied at all?" or "Was this studied five times?" Participants were randomly assigned to conditions using two keys to respond "yes" or "no", or using four keys with one set of "yes" and "no" per question. Four-key participants were expected to shift their criteria while 2-key participants could not due to response interference, though results showed that both conditions were capable of criterion-shifting. In Experiment 2 test items appeared on either the left or the right side of the screen; only strong words appeared on the right and only weak words on the left. Participants went through one study-test cycle with four response keys, and one with two. Regardless of the testing conditions, participants did not shift their criteria in the 2-key condition while participants in the 4-key condition did shift their criteria. Finally, Experiment 3 fully crossed 2 or 4 key conditions with blocked or unblocked presentation of test items. Previous experiments have found both number of response keys and blocking of presentations to have an effect on ability to criterion-

shift (Hicks & Starns, 2014; Starns & Olchowski, submitted; Verde & Rotello, 2007). Experiment 3 confirmed that number of response keys has a significant effect on criterion-shifting and that it is separate from any effect of blocking. All three experiments suggest that response interference is not the driving force behind criterion-shifting. A new explanation is proposed.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

AN INVESTIGATION OF THE BASIS OF THE STRENGTH-BASED CRITERION-SHIFT

1.1 Introduction

Memory guides our actions and teaches us about the world. However, our memories do not consist of perfect recollections of prior events. Memories are instead reconstructions possessing varying degrees of fidelity. Events that we are confident have been accurately remembered tend to be those that have a level of importance and relevance. The accuracy of these significant memories can vary, as can our confidence in them. How and why our judgments about our own memory vary according to the evidence we have on hand is not completely understood. Researchers have not yet outlined the complete process that allows people to distinguish an item stored in memory from a new item with familiar characteristics. In fact, it is still not possible to completely describe how it is determined that a memory is strong enough to decide that an item has been recognized.

Encouragingly, it is often possible to make quite reasoned judgments about memories. An example of our ability to judge our own capacity to remember is the so-called "mirror effect" (Glanzer & Adams, 1985). The generic mirror effect is most easily understood by looking at a specific form of mirror effect involving hit rates (HR) and false-alarm rates (FAR) in a recognition task. Participants in a standard recognition task are asked to study a list of words and are then given a test list consisting of both words that were studied and new words. Their task is then to determine whether or not each test item appeared on the previously studied list by answering "studied" or "not studied". The HR consists of the proportion of "studied" responses for items that were on the previous list (targets), and the FAR consists of the proportion of "studied" responses for items that were *not* on the previous list (lures). When participants are tested on a list that consists entirely of items studied more often ("strong" items) they not only

have the higher HR that would be expected from stronger items, they also have a lower FAR for strong items than they do for items studied less often ("weak" items). Weak items on the other

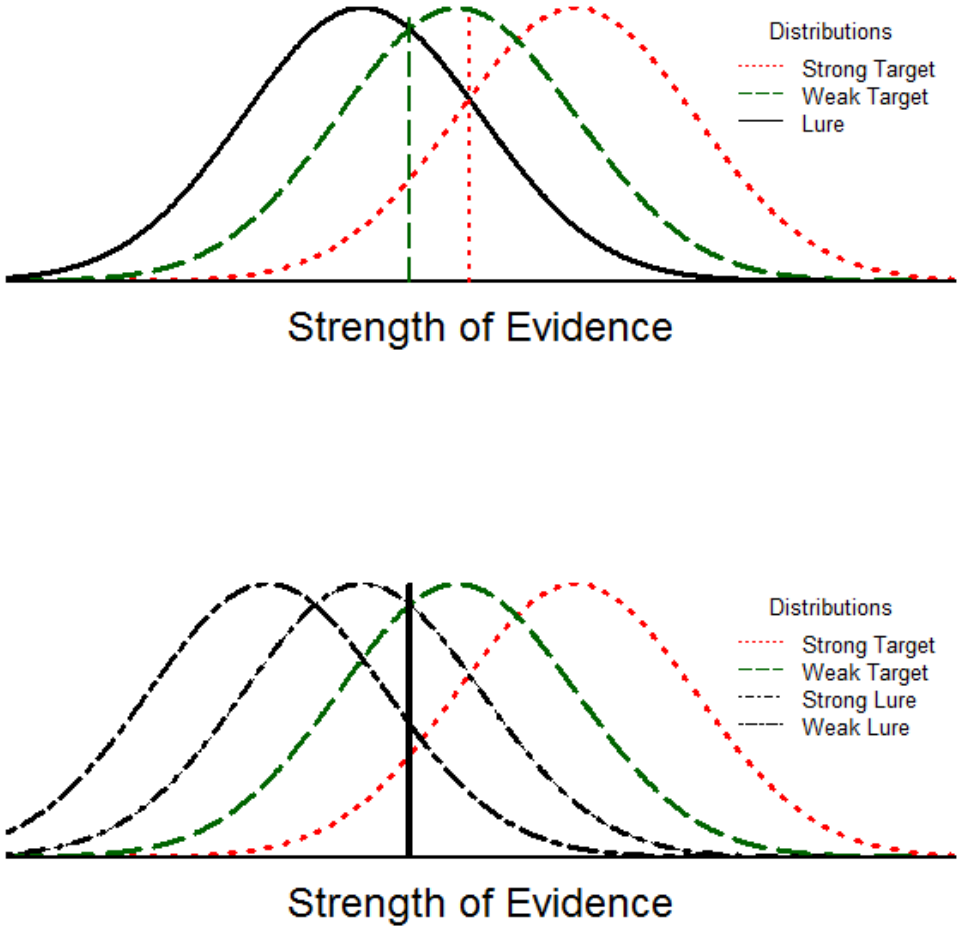

Figure 1: The multiple criteria Signal Detection (Top panel) vs. the single criterion Differentiation (Bottom panel) explanations of the mirror effect

hand have a higher FAR and a lower HR, producing "mirrored" data. (Cary & Reder, 2002; Glanzer, Adams, Iverson, & Kim, 1993; Singer 2009; Singer & Wixted, 2006; Starns, Ratcliff, & White, 2012; Stretch & Wixted, 1998; Verde & Rotello, 2007). There are some dual-process accounts of the mirror-effect that explain the difference in FAR for strong and weak items by

appealing to data from a remember-know task (Cary & Reder, 2002). However, there is reason to believe that the data are better explained by a single-process signal-detection model of memory, in part because remember-know data can be easily explained via a more parsimonious single-process model if one assumes that remember tracks a strong criteria for memory and know represents a weak criteria. (Hirshman, 1997; Wixted & Stretch, 2004).

Signal-detection theory explains mirror-effect data by way of the relationship between the memory evidence provided by the strong and weak items and the memory evidence provided by lure items. The assumption is that each test item has a value of "memory strength" for a participant and studied items have strength scores that are on average higher than the scores of items that were not studied. But due to variation in the memory evidence for individual items, some lures have a higher strength value and some targets have a lower strength value as seen in Figure 1, top panel (see above). Participants in a test of recognition memory establish a cut-off point for the strength of evidence, or a "response criterion," above which they will say "studied" and below which they will say "not studied". The response criterion represents the understanding participants have of how strong they expect studied items to be. This criterion can change depending on the information available to the participants (Benjamin & Bawa, 2004; Brown & Steyvers, 2005; Brown, Steyvers, & Hemmer, 2007; Lages & Treisman, 1998; Singer, Gagnon, & Richards, 2002; Stretch & Wixted, 1998; Tanner & Swets, 1954; Treisman & Williams, 1984). According to this model a change in FAR between the two strength categories demonstrates that participants adjust the response criterion based on strength, which in turn produces the lowered FAR for strong items seen in the mirror effect data. A participant making fewer "yes" responses to lure items when they expect to have strong memory versus weak memory indicates that the participant is evaluating those lures against a stricter criterion.

Some prior experimenters have proposed that the mirror effect is the result of a

differentiation rather than a criterion-shifting process. The differentiation account states that participants maintain a single criterion, but lures become progressively lower in memory strength and more dissimilar from studied items with more extensive learning and therefore the difference in FAR is entirely due to the separation between the strong-item and weak-item lures, as can be seen in Figure 1, bottom panel (see above; Criss, 2006, 2009, 2010). While the differentiation effect posits that the mirror effect is due to participants studying lists that differ in strength, there have been multiple recent experiments that have eliminated the possibility of differentiation effects by giving participants the same mixed study lists and telling them either only strong or only weak items would appear at test. These experiments find evidence of the mirror effect, suggesting that study lists consisting of entirely strong or weak items contributing to a differentiation effect are not behind the mirror effect data. (Dobbins & Kroll, 2005; Starns, Ratcliff, & White, 2012; Starns, White, & Ratcliff, 2010; 2012).

When participants are allowed to study some items at length and some items extremely briefly, their response criterion changes (Dobbins & Kroll, 2005; Hicks & Starns, 2014; Shiffrin, Huber, & Marinelli, 1995). If participants are accurately informed that they will be presented with a test list that consists entirely of strong items, they will set a high criterion for the amount of evidence they will require to state that an item was studied. When participants are presented with a list of entirely weak items they have exactly the opposite inclination (Hirshman, 1995; Verde & Rotello, 2007). Weak items will tend to have been not learned very well by participants, therefore participants will require less evidence to state that an item was studied. While showing a purely strength-based criterion-shift is difficult, it has some advantages over other designs such as those using emotional valence and other varieties of manipulation that are more complex than pure strength manipulation. These manipulations lead down more complex interpretive paths. An explanation of pure strength-based criterion-shift would explain most other observed criterion-

shift results, while an explanation of emotionally-valenced criterion-shift might depend on aspects unique to the emotional valence manipulation for instance.

The ability to adjust the response criterion with test item strength does have some limitations. Many experimenters have tried and failed to get participants to shift their criterion on a trial-by-trial basis (Stretch & Wixted, 1998; Verde & Rotello, 2007). When participants are presented with sufficiently large blocks of strong items followed by weak items, they are capable of shifting from a strict criterion for the strong item blocks to a lax criterion for the weak item blocks. Participants only display this ability to shift the criterion for responding if they are also provided with feedback about performance (Verde & Rotello, 2007) or cues that explicitly differentiate the strong blocks from the weak (Hicks & Starns, 2014). When those blocks are too short, the criterion shift is effectively eliminated even when items are still explicitly differentiated (Hicks & Starns, 2014). Participants given blocks that are too small appear to use the same criterion for all test items presented rather than considering the appropriate criterion strength for each test item. A possible explanation is that due to the items being presented without explicit signaling, participants are simply unaware of the fact that there are multiple strengths being presented and so see no reason to shift the criterion. A study by Stretch and Wixted (1998) that addressed this possibility makes that proposition seem less likely. They found that even marking strong and weak items with distinctive colors (red for strong, green for weak) did not allow participants to effectively shift their criterion in response to strength cues when presented with test items in a random order, which of course frequently produces blocks as short as a single item.

In contrast, some studies have produced evidence that participants can shift their criterion according to strength cues when presented with randomly ordered items at test (Bruno, Higham & Perfect, 2009; Dobbins & Kroll, 2005; Singer, 2009, semantic orienting task conditions; Singer & Wixted, 2006, Experiments 3 & 4). However, one thing these studies have in common is that they

apply additional semantic category markers above and beyond strength to test items. For example, in one condition of the Singer (2009) task participants were instructed to rate words for pleasantness at study in addition to studying them either five times or once. Criterion shifting was found in the condition which performed the pleasantness rating, but not those participants in the control condition who had only studied some items five times and some items once without making any such rating. Participants in these experiments are not only informed of how often they have studied a given test item, they are also informed of the fact that a test item belonged to one or the other semantic category when it was studied. This means that it is questionable whether the criterion-shift observed in these experiments shares a mechanism with criterion-shifts that occur solely based on strength cues. It is possible that participants relied on the semantic category markers to make criterion-shift decisions in these experiments and the strength markers were secondary or unused.

A recent experiment by Starns and Olchowski (submitted) reliably induced criterion shifts with random presentation of test items using strength cues alone, without category differences. The effect was produced via a minor alteration to the experimental paradigm used by Stretch and Wixted (1998). The experiment consisted of a replication of that paradigm with the addition of a third response key at test. The participants were informed of the strength of test items in the same manner as participants in the original Stretch and Wixted experiment: Participants were told that the red items would either have been studied five times or not at all and the green items would either have been studied once or not at all. At test participants were instructed to indicate whether or not a given test item had been studied by using the response keys available. In addition to the color cue, items were presented on the left side of the screen if weak and the right side of the screen if strong as an additional strength indicator, and participants were informed of this additional explicit cue. As in the Stretch and Wixted experiment, participants were not

specifically told to shift their criterion in response to these cues and both study and test items were presented in a random order. Participants were only made aware of the strength difference during study and the cues that indicated strength at test. Unlike the Stretch and Wixted study, participants were told to respond "studied" using a different key for strong and weak items. Participants used "S" for a weak-cued "studied" response and "L" for a strong-cued "studied" response. Regardless of the cue on screen, participants always used the space bar to respond "not studied". With the addition of separate response keys for strong and weak items, participants displayed a significant difference in FAR between the strong items presented on the left and weak items presented on the right that can be explained by criterion-shift. When participants in a further experiment were given only two response keys the FAR difference disappeared. The fact that participants were capable of trial-by-trial criterion-shifting in 3-key conditions but not in 2-key conditions raises the question of what aspect of the 3-key condition is allowing participants to make these criterion-shifts.

Some experimenters have suggested that strength-based criterion-shifting is difficult to observe when test items are presented in a random order because criterion-shifting takes too much effort for participants to repeatedly perform (Benjamin, Diaz, & Wee, 2009; Curran, Debuse, & Leynes, 2006; Morrell, Gaitan, & Wixted, 2002). Experiment 3 of Starns and Olchowski (submitted) tested this concept directly. Experiments 1 and 2 showed that it was possible to induce criterion shifts in the Stretch and Wixted (1998) paradigm by giving participants an additional response option in addition to the explicit strength cues of color and side. Having shown that it is possible to induce a criterion shift, the effort hypothesis could be tested directly. Experiment 3 split participants into two groups. The control was the exact same as the prior 3-key condition, aside from a lack of color cuing on test items (i.e. strength was signaled solely via side of the screen). In the experimental condition, strong and weak items were

displayed with equal frequency on both sides of the screen. In the experimental condition participants were thus switching keys dependent on which side of the screen items were presented without having any reason to shift their criterion based on strength differences. The results showed evidence of criterion-shift only in the control condition, as expected. The results also showed no significant difference in response time between the participants in the control and experimental conditions, even though the former showed criterion-shift and the latter did not. The fact that there was no difference in response times suggests that the shifting of a criterion does not require that participants expend any kind of extra effort, as that would be expected to cause a longer response time.

Earlier experiments have suggested that a key element underlying the ability to criterion-shift when test items are presented in a random order at test is awareness of the strength manipulation. Rhodes and Jacoby (2007) conducted an experiment in which participants studied a list of words and were tested on recognition memory. At test words appeared on either the left or the right hand side of the screen and participants were told to respond with either two keys (studied or not studied) or four keys (studied left/right, not studied left/right) depending upon the condition. The test list consisted of both targets and lures, with the proportions presented on each side of the screen manipulated. One side was mostly targets, and the other side was mostly lures. Participants in the 4-key condition displayed an FAR difference, which suggests that a criterion shift occurred. Rhodes and Jacoby hypothesized that providing additional response keys made participants more likely to notice that the proportion of targets differed between the sides of the screen. This was supported by post-experimental questionnaires asking whether or not participants noticed that the proportion of studied items presented at test differed dependent on the side of the screen. Verde and Rotello (2007) made a similar proposal in their examination of criterion-shifting. While they were unable to induce participants to shift in most of their

experimental conditions, they did have some success when participants were not only provided with large blocks of strong and weak items, but with accurate feedback about whether they had properly identified a given item as studied or not studied. Verde and Rotello explained their criterion results as being due to participants being made more aware of the fact that there were multiple categories of test item and responding with a criterion shift only when they were sufficiently aware that there were multiple item types presented at test.

However, both Stretch and Wixted (1998) as well as Starns and Olchowski (submitted) provide some evidence that awareness of the test structure does not necessarily produce criterion shifts. In both of these studies, participants were made aware of the experimental manipulation. Yet in both the 2-key condition of Starns and Olchowski (submitted) and the Stretch and Wixted (1998) experiments which used two response keys, participants did not display criterion-shift. Participants in the 3-key condition of Starns and Olchowski (submitted) did display a criterion-shift. The 3-key conditions did not contribute to additional awareness of the experimental manipulation in the manner proposed by Rhodes and Jacoby (2007), as participants in all conditions were made aware of the manipulation through explicit pre-test instructions. There was therefore no opportunity for participants to discover the manipulation during the test phase, yet the 3-key condition produced a major difference in FAR while the 2-key condition did not. It is therefore reasonable to propose that criterion-shifting is not solely a function of whether or not the participant is aware of the fact that there are multiple strength categories being presented at test. Participants seem to be able to shift easily and efficiently when given more than two keys to respond, and to be unwilling or unable to shift when given only two keys. This is true regardless of whether participants are made entirely aware of the experimental manipulation. What is still unclear is why this is the case. The process of criterion-shifting needs to be investigated in more detail if we want to better understand how participants are making memory judgments.

The key question under consideration is why participants are capable of shifting their criterion on a trial-by-trial basis so easily when an apparently minor alteration is made to the experimental setup. A possible explanation is that participants who do not shift are experiencing response interference. Simon, Acosta, Mewaldt, and Speidel (1976) showed that participants who are given response options that match the side on which a stimulus is presented respond faster than participants whose response options do not match the stimulus location,, something that has come to be known as the "Simon Effect." In a typical experiment, Simon et al. had participants respond with a key on the left or the right side of a response box depending on the color of the light light currently illuminated in front of them. A red light indicated that the participant was to respond with the right button, while a green light indicated that the participant was to respond with the left button.. Simon et al then placed the colored lights on either the same side as the button they signaled (red light on the right, green light on the left) or on the opposite side of the button they signaled (red light on the left, green light on the right). The irrelevant cue of light location was found to interfere with participant responses and cause response time to slow in the opposite-side light condition. This is a pertinent example of how a response conflict can have a measurable effect on the ability to make even a very simple decision. Participants were required to use controlled processing to hit a button on the right in order to indicate that a light had lit up on the left (Simon et al. 1976; Simon & Berbaum, 1990).

Conflict between responses on successive trials could produce interference in the same manner as conflict between response and stimulus attributes (Clare & Lewandowsky, 2004; Curran, Debuse, & Leynes, 2006). Participants in an experiment like Stretch and Wixted's (1998) paradigm are required to use responses that overlap completely even when the response is intended to indicate something other than what it indicated on the last trial ("old" in response to a medium-strength item versus "new" in response to a medium-strength item). For example,

imagine that trial N is a weak-cued trial and the item has a moderate level of memory strength. The participant applies a lax criterion and presses the key indicating that the item has been studied. Trial N+1 is another moderate-strength item, but appears with the strong cue. To successfully shift the response criterion when responding to trial N+1 the participant must inhibit the impulse to make the same response they just made for a very similar stimulus and instead hit the key indicating that the item was not studied. If a participant were provided with separate keys to indicate that a strong versus a weak item was studied then the overt response made on trial N would not be available on trial N+1 and the need for controlled processing would be alleviated. In other words, having less overlap in responses should reduce interference across trials and facilitate criterion shifts. In this account, criterion shifts in and of themselves do not necessarily require effort. The onerous aspect of the task is the need to overcome the interference inherent in using the same response to mean different things from one trial to the next. The goal of this project is to test this response interference account with the three experiments described below.

1.2 Experiment 1

Our first experiment used a paradigm adapted from the original Starns and Olchowski (submitted) Experiments 1 and 2 and modified to strongly induce participants to shift their criterion. Participants studied a list consisting of nouns, verbs, and adjectives. Strong items appeared on the study list five times, while weak items appeared only a single time. At test, they were presented a mixed list of strong and weak items. Each item would appear on either the left or the right hand side of the screen, selected at random and appearing underneath one of two questions. A word that was studied a single time should receive a "yes" response if it appeared underneath the question "Was this studied at all?" and a "no" response if it appeared underneath the question "Was this studied five times?" We ran two conditions. In the 2-key condition, the "yes" and "no" responses were the same regardless of the question being asked. In the 4-key

condition, the “yes” and “no” responses were unique for each question, with a different one for each side of the screen. This paradigm strongly implied that correct responding required an appropriate criterion shift. It was our expectation that the additional time pressure would exacerbate response interference in the 2-key condition, reducing or eliminating evidence of criterion-shifting in that condition. We also expected that participants in the 4-key condition would experience little to no response interference and therefore be capable of performing criterion shifts easily.

Evidence of criterion-shifting was expected to appear in the difference between the FAR for items appearing under the “Studied five times” and “Studied at all” questions in the 4-key condition. We expected that participants in the 4-key condition would make fewer “studied” responses overall for items in the “Studied five times?” question than they did in the “Studied at all?” question. This change in responding would produce a drop in FAR, as well as a corresponding drop in both strong and weak HR. In contrast, we expected the 2-key condition to show similar responding to both “Studied five times” and "Studied at all" question items, reflecting a lack of ability to shift criterion according to the cue.

1.2.1 Methods

Participants. Seventy-eight University of Massachusetts undergraduate students participated, randomly assigned to each experimental condition. The 4-key condition included 38 participants, and the 2-key condition included 40 participants. Participants were tested individually and they earned extra credit in psychology courses as compensation for their time. In the 4-key condition, our response data indicate only 11 participants ever pressed the wrong key, and only 27 out of 13464 trials were affected, these trials were dropped from analysis. Eight low-performance participants were removed, either due to an overall accuracy below 60% on the memory test or an accuracy below 90% on the last half of the Response Practice portion. This left

34 4-key and 36 2-key participants in the final data.

Materials. Stimuli consisted of nouns, verbs and adjectives randomly selected from a pool of 859 low-frequency words generated for use in memory experiments. Forty of these words were assigned to a short practice phase and 408 words were assigned to the experimental phase that provided the data which were analyzed. The practice study list consisted of 10 strong targets and 10 weak targets. The strong targets were each presented five times throughout the list and the weak targets were presented once, for a total of 60 study presentations. The practice test items consisted of all the words from the practice study list in addition to 20 lure items, 10 lures each for the ”Studied five times” and "Studied at all" questions.

Study lists for the experimental phase consisted of 34 strong target words and 34 weak target words. The strong target words were each presented five times at study and the weak target words were each presented a single time, for a total of 204 presentations. The test lists for the experimental phase included all of the words on the study list as well as 68 lures. Thirty-four lures were presented with a "five times" question and 34 were presented with a "at all" question. The order of the words in each list was randomized for each participant. Each participant completed three study/test phases, with the order of the words in each list randomized for each phase but no words repeated across cycles.

This experiment, as well as Experiments 2 and 3, used an X-Box 360 controller for participant responding. A controller was chosen instead of a keyboard because it serves to restrict the number of possible responses on the part of the participant while allowing for ergonomic multi-key responding using “triggers” and “bumpers”. In addition, the controller offers options for future investigation of differences in performance based on the physical structure of responding. We used the program AutoHotKey to create a script that allowed the X-Box 360 controller keys to function as responses.

Design and Procedure. In this experiment, participants were read the instructions by an experimenter at the same time as they read them on the screen. Participants were given a response practice phase at the start of the experiment designed to acquaint them with the responses they would use at test. This phase involved seeing 50 presentations each of the possible item types for a total of 200 presentations. These were not actual test items, but symbols representing the different classes of item such as "5x" for an item studied five times. Participants were instructed to respond as though these were actual test items with the appropriate key, and were shown an error message if they pressed an incorrect response key.

Participants were then given a short practice phase in which they were introduced to the experiment and had a chance to ask questions. The second practice phase was exactly the same as the experiment proper, but the study and test lists were shortened to 60 study items and 40 test items so that participants could proceed through the practice in only a few minutes. Participants were asked to pay attention to the lists of study words because their memory would be tested later. Study items remained on the screen for 900 ms, with 100 ms of blank screen between each item. For all lists the words were presented in white text in the center of a black computer screen.

Following the study phase, participants performed a 2-back task as a distraction for approximately 30 seconds. A random sequence of numbers ranging from 1 to 9 appeared on the screen one at a time with a new digit appearing every second, for a total of 30 digits. Participants were told to press the "/" key every time they saw a number that matched the number appearing 2 items back in the sequence.

The test list was presented directly after the 2-back task. Test stimuli were presented on either the left or the right hand side of the computer screen. Both sides of the screen presented items that had been studied once or five times, as well as lures. Participants were asked to respond to the questions "Was this studied at all?" for items on one side and "Was this studied

five times?" for items on the other. If a test item was presented on the side of the screen under the "at all" question, they were instructed to respond with "yes" if it was studied either once or five times and "no" if it was not studied at all. If a test item was presented under the "five times" question, they were told to respond "yes" only if the item was studied five times and "no" otherwise. Participants in each condition were also informed that they were under time pressure and that if they took more than 1800 ms to respond they would receive a "too slow" message. In addition, they were told that if they responded too quickly, faster than 500 ms, they would receive a "too fast" message. Each of these error messages remained on screen for 1500 ms as an incentive to avoid encountering them.

Participants who used 4 response keys responded using both "triggers" and both "bumpers" on a standard X-Box 360 wired controller. These keys differ primarily in their location, with the "triggers" being below the "bumpers". All four keys are located on the "shoulders" of the controller, ergonomically designed for use with four fingers at the same time. in the 4-key condition, bumpers were used to indicate "studied" while triggers were used to indicate "not studied". Words appearing on the left were responded to using the left bumper and trigger while words appearing on the right were responded to using the right bumper and trigger. If a participant incorrectly used the right trigger or bumper for a word presented on the left or the left bumper or trigger for a word presented on the right, they received feedback indicating they had used incorrect keys in their response and reminding them which keys should be used.

For the participants who used two response keys, responses were made using only the bumper and trigger on a single side. The side used for responding in the 2-key condition was randomly chosen for each participant at the beginning of the experiment, but the participant always used only that side to respond with either the left or right bumper and trigger meaning "studied" and "not studied" respectively. Instructions were identical for 2-key participants, except

for the identification of the response keys. Participants were informed of the percentage of their responses which were correct at the end of each study-test cycle.

1.2.1.1 Results

We used $\alpha = .05$ for all statistical tests. The critical data considered were the FARs and HRs for strong and weak items shown with the "Was this studied at all?" and "Was this studied five times?" questions. We ran t-tests on the difference between the "Studied at all" and "Studied five times" test questions for both 2-key and 4-key participants as well as a 2 × 2 ANOVA to investigate the possibility of interactions. If the response interference account is an accurate interpretation of why participants did not shift their criterion trial-by-trial in previous experiments, then participants in the 2-key condition should display a significantly smaller difference between the "Studied at all" and "Studied five times" test questions than the participants in the 4-key condition.

Table 1 shows the proportion of "yes" responses – p("yes") – for for strong, weak, and lure items under the strong and the weak questions for both the 2-key and 4-key conditions. The HR was in all cases higher for strong items than weak items in both the 2-key and the 4-key conditions. The HR was also higher for both strong and weak targets in the "Studied at all" question condition. The difference in p("yes") between the "Studied five times" and "Studied at all" questions for the lure data which was the focus of our interest was significant in both the 2-key and the 4-key data.

Table 1
Proportion of "Yes" Responses in Experiment 1

	Five Times Question			At All Question		
	Strong	Weak	Lure	Strong	Weak	Lure
4-key Cond	0.569 (.03)	0.247 (.03)	0.092 (.01)	0.736 (.03)	0.473 (.03)	0.204 (.02)
2-key Cond	0.703 (.02)	0.302 (.02)	0.115 (.02)	0.826 (.02)	0.547 (.03)	0.219 (.02)

Standard errors are noted in parenthesis.

A 2(question) × 2(key condition) ANOVA on the lure item data revealed that the "Studied

at all" question condition was significantly more likely to result in a "yes" to a lure item from participants than the "Studied five times" question , $F(1, 68) = 106.27$, $p < .001$. Further, the value of partial eta-squared ($\eta^2 = .61$) suggests a large amount of variance is accounted for by the question condition. Both the 4-key and the 2-key conditions had similar FAR, indicating that there was not a significant difference in FAR depending on key condition, $F(1,68) = .652$, $p = .422$. No significant interaction was found between the question asked and the key condition, $F(1, 68) = .141$, $p = 0.709$. In both of these cases, partial eta-squared results ($\eta^2 = .009$ and $\eta^2 = .002$ respectively) indicated that they accounted for very little of the overall variance.

Participants in the 4-key condition responding to lure items presented under the "Studied at all" question gave "yes" responses more often (M = .204, SD = .08) than those responding to lure items under the "Studied five times" question (M = .092, SD = .10). A similar pattern was seen with participants in the 2-key condition when lures were presented under the "Studied at all" (M = .219, SD = .15) versus the "Studied five times" (M = .116, SD = .10) questions. Paired-samples *t*-tests confirmed that the effect of question type on the p("yes") to lure items was significant in both the 4-key, $t(33) = 8.55$, $p < .001$, and 2-key experimental conditions, $t(35) = 6.43$, $p < .001$, with the p("yes") to lure items in each case being higher when participants were asked the "Studied at all" question. Cohen's d in both cases (d = 1.21 and d = .84 respectively) indicates a large effect.

We also investigated differences in HR for strong items. A 2(question) × 2(key condition) ANOVA on the HR data for strong items revealed that the "Studied at all" question condition was significantly more likely to result in a "yes" to a strong item from participants than the "Studied five times" question , $F(1,68) = 376.28$, $p < .001$. Partial eta-squared ($\eta^2 = .847$) suggest that the majority of the variance was due to this effect of question asked. In addition to the effect of question asked, which key condition a participant was in contributed to the difference in hit rate

with HR significantly higher in the 2-key condition, $F(1,68) = 8.495$, $p < .01$. A significant interaction between the question asked and the key condition was also found with HR being lower for the 4-key participants in response to the "Studied five times" question but equalizing somewhat in response to the "Studied at all" question, $F(1,68) = 4.432$, $p = 0.039$. Partial eta-squared suggests both of these effects accounted for only a small amount of the variance ($\eta^2 = .111$ and $\eta^2 = .061$ respectively).

We explored the interaction further by examining the effect of question asked on strong-item HR in each key condition and the effect of key condition on HR for strong items in each question condition. Paired-sample t-tests showed that there were significant effects of which question was asked on HR for strong items in both 4-key, $t(33) = 7.259$, $p < .001$, and 2-key, $t(35) = 6.151$, $p < .001$, conditions. Cohen's d in both cases indicates a large effect of question asked on HR (d = 0.92 and d = 0.86 respectively). Further *t*-tests found that there was an effect of key condition on HR for strong items in the "five times" question condition, $t(33) = 2.86$, $p < .01$, as well as in the "at all" question condition, $t(33) = 2.32$, $p < .05$. In each case, Cohen's d indicates a moderate effect of key condition on p("yes") to targets (d = 0.64 and d = .059, respectively).

Finally, we investigated differences in p("yes") for weak items. A 2(question) × 2(key condition) ANOVA on the p("yes") data for weak items revealed that question type had a significant effect on the probability of a "yes" for weak items, $F(1,68) = 418.646$, $p < .001$. This effect of question asked again accounted for the majority of variance ($\eta^2 = .86$). Which key condition a participant was in contributed to the difference in hit rate with HR significantly higher in the 2-key condition, $F(1, 68) = 5.059$, $p < .05$. However, it there was no interaction between the question asked and the key condition $F(1,68) = 0.388$, $p = 0.536$. The effect of the question asked on p("yes") for weak items was significant in each key condition. Both the key condition and the interaction accounted for only a fraction of the variance ($\eta^2 = .069$ and $\eta^2 = .006$

respectively). Participants in the 4-key condition answered "yes" more often to weak items when asked the "Studied at all" question (M = .476 , SD = .18) than when asked the "Studied five times" question (M = . 247 , SD = .15). Similar data was found for participants in the 2-key condition when asked "Studied at all" (M = .547, SD = .16) versus "Studied five times" (M = . 302, SD = .14). Paired-sample t-tests showed that in both 4-key, $t(33) = 8.041$, $p < .001$, and 2-key, $t(35) = 7.685$, $p < .001$, conditions there was a significant effect of which question was asked on the p("yes") for weak items. Cohen's d again indicated a large effect of question asked on HR (d = 1.39 and d = 1.65 respectively).

In addition to the analyses based on FAR data, we performed analyses on the criterion parameter in a signal detection model to address potential issues with using FAR difference as a measure of change in criterion. Due to the fact that FAR is a proportional measure, the number of "studied" responses to lure items, this measure does not conform to the assumptions made by the tests shown above. Ideally, when fitting to an ANOVA model, data are expected to follow a continuous, unbounded normal distribution. Proportion data necessarily cannot go below 0 or above 1. In contrast, the criterion parameter is not limited in this way.

We fit each participant's data with a standard signal detection model. Our prediction function took in four parameters: the mean of the strong target distribution, the mean of the weak target distribution, the strong question criterion, and the weak question criterion. This function then returned the proportion of responses in each of the twelve possible categories; yes and no responses to a strong, weak, or lure item under either the "Studied at all?" or the "Studied five times?" question. We found the best fitting parameters by minimizing chi-squared. The overall fit was acceptable with a mean chi-square across all participants of 2.46. A chi-square test showed a significant deviation between the observed and predicted response frequencies for only six participants. These six participants represented about 8% of the sample, compared to the 5%

rejection rate expected under the null hypothesis that the empirical data were generated by the signal detection model.

The best-fitting strong question and weak question criterion parameters were analyzed in the same manner as the FAR for lure items. A 2(question) × 2(key condition) ANOVA on the criterion data revealed that the "Studied at all?" question condition resulted in a significantly lower criterion (M = .857, SD = .43) from participants than the "Studied five times?" question (M = 1.424 , SD = .52) , $F(1, 68) = 153.05$, $p < .001$. Further, the value of partial eta-squared ($\eta^2 = .69$) suggests a large amount of variance is accounted for by the question condition. The 4-key and the 2-key conditions had similar change in criterion values, indicating that there was not a significant difference in criterion depending on key condition, $F(1,68) = .433$, $p = .513$. No significant interaction was found between the question asked and the key condition, $F(1, 68) = 1.896$, $p = 0.173$. In both of these cases, partial eta-squared results ($\eta^2 = .006$ and $\eta^2 = .027$ respectively) indicated that they accounted for very little of the overall variance.

1.2.1.2 Discussion

The results of this experiment do not support the claim that response interference is a significant barrier to trial-by-trial criterion-shifts. Participants made aware of the need to shift performed equally well in both the 2-key and 4-key conditions. If the response interference hypothesis were accurate, participants should have been deficient in shifting their criteria when lacking proper response options even though the manipulation made criterion-shifting both obvious and necessary for proper performance. Based on the results seen in Experiment 1, it seems that extra response options offered in Starns and Olchowski (submitted) may be facilitating performance in some way unrelated to a reduction in response interference.

It is worth noting, however, that participants are also given a much stronger incentive to shift in both the 2-key and 4-key conditions of this experiment than they were in the Starns and

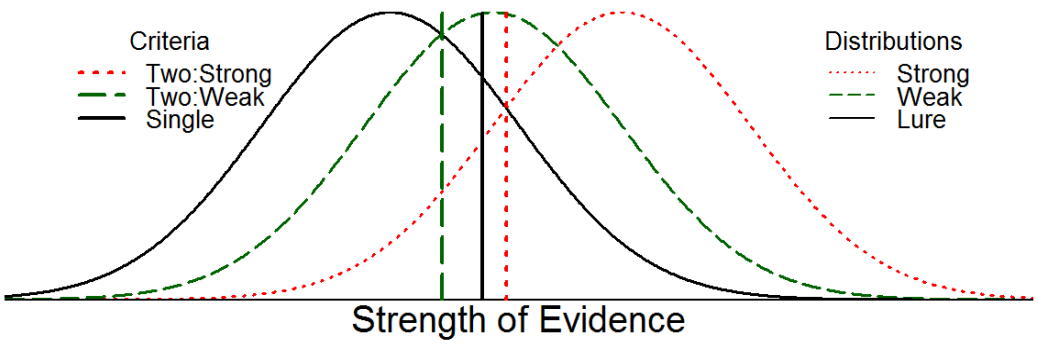

Figure 2: Single vs. Multiple Criteria in Unforced Shift Conditions. d' Strong: 1.8, d' Weak: .8; Single Criterion: .71354, Two Criteria: Strong = .9, Weak = .4

Olchowski (submitted) experiments. While some past experimenters have focused on the costs of making a criterion-shift (Morrel, Gaitan, & Wixted, 2002; Benjamin, Diaz, & Wee, 2009; Stretch & Wixted, 1998), few have examined the benefits of a criterion-shift from the perspective of a participant. In a standard criterion-shift experiment, successfully shifting the criterion for each item often provides an insignificant increase in overall accuracy.

Figure 2 (see above) shows example lure, weak, and strong distributions. Criteria have been placed at the optimal position for responding "yes" or "no" in an experiment like Stretch and Wixted's (1998) work, such that they maximize the likelihood of making a "studied" response to targets while minimizing the possibility of a "studied" response to lures in a standard criterion-shift experiment. At test, these items would appear on either the left or the right hand side of the screen with strong items or lures appearing on one side and weak items or lures appearing on the other. Unlike the current experiment, participants would not see any weak items on the strong side or vice versa, and would not see any questions above either side encouraging a criterion-shift response. Thus this figure represents possible behavior by participants who are not forced to respond with strength-based criterion-shifting (Stretch & Wixted, 1998; Starns & Olchowski,

submitted). The black criterion in the figure is placed at the ideal location for a single criterion, while the red and green criteria are placed at the ideal position for two criteria, each one maximizing correct responses to only weak or only strong items respectively.

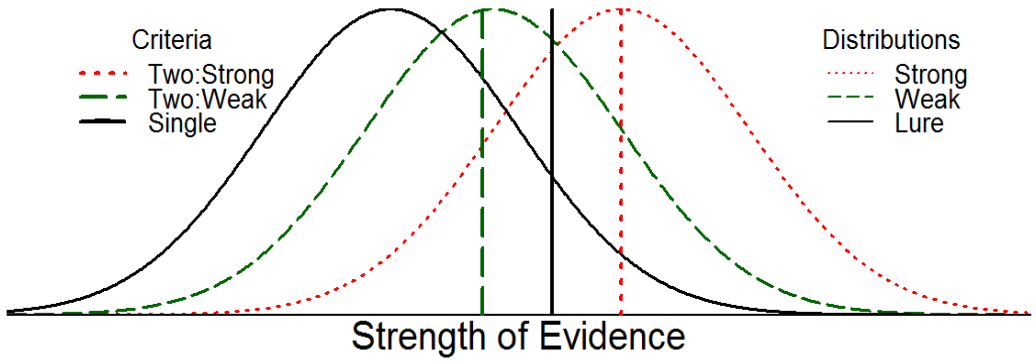

Figure 3: Single vs. Multiple Criteria in Forced Shift Conditions. d' Strong: 1.8, d' Weak: .8; Single Criterion: 1.257, Two Criteria: Strong = 1.801, Weak = .71354

A participant using two optimal criteria for responding and shifting them depending on the item presented would find that out of a 140 item test list they would respond correctly to approximately 103 items for a total accuracy of approximately 74%. A participant using a single criterion and not shifting it at all over the course of the test would respond correctly to approximately 102 items for an approximate accuracy of 74%. Even setting aside whether or not criterion-shifting requires effort on the part of the participant, the choice to shift or maintain a single criterion is at best a equivalent from the perspective of the participant if their goal is to maximize accuracy or minimize misses.

Figure 3 (see above) presents similar example lure, weak, and strong distributions but with optimal criteria designed to mimic what would be appropriate for the current experiment. The optimal criteria in this case are calculated by maximizing the number of correct responses possible while taking into consideration the differing base rates in the two conditions. The key

difference is that in the current experiment, both strengths of item can appear underneath each question cue. That is, the participant is compelled to answer "was this studied five times?" to weak items in addition to strong items and lures. Because answering "yes" to a weak target in the "Studied five times" question condition was an incorrect response, there is a much greater impetus to use a correct criterion in each cue condition. This is most easily seen by observing the differences in weak item responses. In this experiment, a participant using a single ideal criterion and not shifting their criterion with strength would respond correctly to a weak item at best 50% of the time, approximately 17 out of 34 weak targets. Whatever criterion the participant uses, the number of correct responses to weak targets in the weak question will be exactly offset by incorrect responses to the weak targets in the strong question. On the other hand, a participant using the two criteria proposed in figure 2 would respond accurately to approximately 24 out of the 34 weak targets, or 70%. This is a large increase in accuracy due to criterion-shifting.

In the conditions previously run, participants lacked any compelling reason to shift based on their understanding of performance. While accuracy as measured by HR does tend to increase slightly in the course of actual experiments when participants use two appropriate criteria, participants often do not see a major benefit. On the other hand, in the current experiment participants in both the 2-key and 4-key conditions had a distinct reason to shift appropriately. Lack of criterion-shifting in this experiment would produce notably impaired performance relative to accurate criterion-shifting. It is not surprising that they would not shift their criterion until it had a significant impact upon accuracy.

It is possible that the design of experiment 1 provides a much greater incentive for participants to respond with the appropriate criterion-shift than in the designs of prior experiments (Stretch & Wixted, 1998; Starns & Olchowski, submitted). If this is the case, it may be that the observed criterion-shift in the 4-key conditions of previous experiments might arise

not from the response options making participants more able to shift but from giving them a more significant incentive to shift. Part of this effect can be attributed to the fact that failure to respond with the appropriate keys in the 4-key condition immediately triggers an error screen. The additional length this adds to the experimental condition might provide incentive above and beyond the accuracy feedback. However, this incentive would primarily serve to encourage participants to choose the correct key for responding, not force a criterion-shift. This suggests then that if a participant is encouraged to use appropriate keys to respond, criterion-shifting occurs regardless of the presence of an accuracy benefit. This may point to the decision to acknowledge the cue being the real difficulty for participants inherent in the task, rather than the criterion-shift per se.

1.3 Experiment 2

In Experiment 2, we turned to the possibility that participants who are capable of performing trial-by-trial shifts with 4 response keys can continue to perform those shifts when subsequently restricted to 2 keys. We hypothesized that even when participants were shown to be capable of criterion-shifting, exacerbating response interference should remove their ability to shift. Similar to Experiment 1, this study was designed to establish whether the number of response keys drove the observed criterion-shift effect regardless of participant awareness or the instructions received. All participants went through a complete 4-key study/test cycle followed by a brief pause for further instruction and a full 2-key study/test cycle. Participants were further randomly assigned to three separate categories of 2-key responding that were expected to create different levels of response interference. The three categories of 2-key responding were 2-key with no response-stimulus interval (RSI), 2-key with a blank RSI of 1200 ms, and 2-key with an RSI of 1200 ms in addition to a label accurately denoting the strength cue of the next item that would appear. The 1200 ms RSI was expected to reduce the experience of response interference

because participants had a period of rest in which to readjust what the "studied" key meant given the knowledge that two criteria are necessary to properly respond. Further, it was expected that the addition of a label that acted as a preview for the strength of the next item to appear on the test would reduce response interference to a great degree. The preview should have allowed participants to re-adjust the context in which their next response occurred, understanding that while they might be hitting the same key it would indicate a different response.

1.3.1 Methods

Participants. One-hundred and forty-three University of Massachusetts undergraduate students participated, with one third assigned to each experimental condition at random. Forty-six were assigned to the 2-key condition with no RSI, 47 were assigned to the condition with a 1200 ms blank RSI, and 50 to the condition with a 1200 ms RSI which included a preview indicator. Participants were tested individually and earned extra credit in psychology courses as compensation. Thirty-five low-performance participants were removed, either due to an overall accuracy below 60% or an accuracy below 90% on the last half of the Response Practice portion. This left 35 2-key with no RSI, 36 2-key plus RSI, and 37 2-key plus RSI and indicator participants in the final data.

Materials. The test lists in this experiment displayed strong targets only on the right-hand side of the screen and weak targets only on the left-hand side of the screen. Otherwise materials were identical to Experiment 1.

Design and Procedure. Participants were randomly assigned to each of the three conditions and controllers were again used, in the same manner described in Experiment 1. Participants were given a practice phase at the start of the experiment designed to acquaint them with the responses they would use at test. This phase involved seeing 50 presentations of each of the possible item types for a total of 200 presentations. These were not actual test items, but

phrases representing items such as “5x” for an item studied five times. Participants were instructed to respond as though these were actual test items with the appropriate key, and were shown an error message if they pressed an incorrect response key. Participants in all three conditions also completed a brief practice phase before the experimental phase during which they had the opportunity to go through a full, although short, study/test cycle and ask any questions they might have before beginning the study list proper. Test items were presented in a random order. After the test phase, participants saw feedback on-screen with their percentage of correct responses; this remained on screen for 4000 ms.

After feedback was presented for the practice cycle, participants were given a full study/test cycle using the four controller keys to respond. Feedback was then presented for this first study/test cycle for 4000 ms. After this feedback, participants viewed a screen with instructions informing them of another study/test pairing, but telling them that they would be using only the bumper and trigger on one side of the controller to respond “studied” and “not studied”. The side of the controller used for responding in this portion was chosen at random for each participant. The instructions also made it clear that the strength cues remained valid: words on the right were strong test items and words appearing on the left were weak test items. The experimenter read these instructions to the participant as well to insure that any questions the participant had about the changing conditions could be answered. Pressing any key other than the appropriate bumper or trigger in this portion resulted in an error message. Participants in the condition with no RSI were presented test items in exactly the same manner as the 4-key condition, one after the other. Participants in the 1200 ms RSI condition were informed in the additional experimental instructions that they would experience 1200 ms of a black screen between test items. Participants in the 1200 ms RSI + Indicator condition were informed by the additional experimental instructions that they would experience 1200 ms of a screen containing a

series of asterisks denoting the side of the screen where the next test item would appear.

1.3.1.1 Results

Table 2 shows the proportion of "yes" responses of participants in each condition. In all cases, the p("yes") for strong targets was higher than that for weak targets. The measure of interest, the difference in p("yes") between strong-cued and weak-cued lure items, differed between the 4-key and 2-key data. In the 4-key experimental phase, all conditions saw a higher proportion of "yes" responses to weak-cued lure items than strong-cued lure items. In each condition, the subsequent 2-key phase of the experiment saw weak-cued lure items with a p("yes") almost equal to the p("yes") for strong-cued lure items.

Table 2
Proportion of "Yes" Responses in Experiment 2

	4-key Data				2-key Data			
	Strong Target	Weak Target	Strong Lure	Weak Lure	Strong Target	Weak Target	Strong Lure	Weak Lure
No RSI	0.82 (.01)	0.57 (.02)	0.15 (.01)	0.22 (.01)	0.80 (.01)	0.54 (.01)	0.28 (.02)	0.29 (.01)
RSI	0.87 (.01)	0.61 (.01)	0.17 (.01)	0.25 (.01)	0.81 (.01)	0.59 (.01)	0.29 (.02)	0.31 (.01)
RSI + Indicator	0.81 (.01)	0.55 (.02)	0.20 (.02)	0.23 (.02)	0.75 (.01)	0.52 (.02)	0.32 (.02)	0.31 (.02)
Overall	0.83 (.01)	0.57 (.01)	0.17 (.01)	0.23 (.01)	0.79 (.01)	0.55 (.01)	0.30 (.01)	0.30 (.01)

Standard errors are noted in parenthesis.

For the 2-key condition, a 2(strength) × 3(RSI condition) ANOVA on p("yes") for the target item data found an effect of strength with strong targets receiving significantly more "yes" responses than weak targets, $F(1, 105) = 386.459$, $p < .001$. Partial eta-squared suggests that a majority of variance is accounted for by this effect ($\eta^2 = .786$). But in each RSI condition, the mean p("yes") for target items remained essentially the same. No significant effect was found for RSI condition, $F(1, 105) = .978$, $p = .380$. Target strong and weak items remained effectively the same across all combinations of strength and RSI condition. There was therefore no significant interaction, $F(2, 105) = 1.480$, $p = .232$. Little of the variance was explained by either the effect

of RSI condition or the interaction between RSI condition and strength ($\eta^2 = .018$ and $\eta^2 = .027$ respectively).

For the 4-key condition, a 2(strength) × 3(RSI condition) ANOVA on p("yes") for the target item data again found an effect of strength, with strong targets receiving significantly more "yes" responses than weak ones, $F(105) = 351.219$, $p < .001$. And again, a majority of variance is accounted for by this effect ($\eta^2 = .770$). Again in each RSI condition, the mean p("yes") for target items remained the same and no significant effect was found for RSI condition, $F(1, 105) = 2.600$, $p = .079$. Target strong and weak items again remained the same across all combinations of strength and RSI condition. There was therefore no significant interaction, $F(2, 105) = .117$, $p = .890$. Little of the variance was explained by either the effect of RSI condition or the interaction between RSI condition and strength ($\eta^2 = .047$ and $\eta^2 = .002$ respectively).

On the other hand a 2(strength) × 3(RSI condition) ANOVA on p("yes") for the lure items in the 2-key data found that across all conditions the probability of a participant responding "yes" did not alter. This ANOVA did not find any significant effect of strength, $F(1, 105) = 3.146$, $p = .079$, RSI condition, $F(1, 105) = .038$, $p = .962$, nor any interaction between strength and RSI condition, $F(2, 105) = .376$, $p = .687$. In each case, very little of the variance in responding was accounted for ($\eta^2 = .029$, $\eta^2 = .001$, and $\eta^2 = .007$ respectively).

As in Experiment 1, we also estimated the response criterion used by each participant in each of the strength-cue conditions. However for this experiment we simply computed the negative z-score of the FAR to find the criterion. The target data could not contribute to the criterion calculation, because each type of target appeared only in one strength-cue condition. We also applied a correction for values of zero in the data in order to get meaningful criterion measures for those values (Snodgrass & Corwin, 1988). A 2(strength) × 3(RSI condition) ANOVA on the participant response criteria in the 2-key data found that across all conditions the

criterion for responding "yes" did not alter. In the 2-key condition the criterion did not vary considerably depending upon whether a lure was presented with a strong (M = .769, SD = .47) or a weak (M = .703, SD = .47) cue. We did not find any significant effect of strength, $F(1, 105) = 3.157$, $p = .078$, RSI condition, $F(2, 105) = .022$, $p = .978$, nor any interaction between strength and RSI condition, $F(2, 105) = .662$, $p = .518$. In each case, very little of the variance in responding was accounted for ($\eta^2 = .029$, $\eta^2 = .000$, and $\eta^2 = .012$, respectively).

A 2(strength) × 3(RSI condition) ANOVA on the p("yes") for participants when responding to lure items in the initial 4-key condition found a very different picture. In the 4-key condition the p("yes") varied considerably depending upon which strength cue a lure was presented with. A main effect of strength was found, $F(1, 105) = 38.432$, $p < .001$. This effect of strength accounted for a moderate amount of the variance ($\eta^2 = .268$). There was no difference between RSI conditions, which is to be expected as all participants experienced the same 4-key condition regardless of their RSI condition assignment. Still, it underscores that there were likely no effects of an accidentally biased sample. The ANOVA found no significant effect of RSI condition, $F(1, 105) = .804$, $p = .487$, or significant interaction between RSI condition and strength, $F(2, 105) = 2.024$, $p = .137$. Neither the effect of RSI condition nor the interaction contributed much at all to variance in responding ($\eta^2 = .015$ and $\eta^2 = .037$ respectively).

We also found the participant criterion for the 4-key condition and performed a similar ANOVA. This 2(strength) × 3(RSI condition) ANOVA on the participant criterion when responding in the initial 4-key condition found a very similar picture to the FAR data. In the 4-key condition the criterion varied considerably depending upon whether a lure was presented with a strong cue (M = 1.248, SD = .57) or a weak cue (M = .897 SD = .46), $F(1, 105) = 48.041$, $p < .001$. This effect accounted for a significant amount of the variance ($\eta^2 = .314$). There was no difference between RSI conditions, which is to be expected as all participants experienced the

same 4-key condition regardless of their RSI condition assignment. Still, it underscores that there were likely no effects of an accidentally biased sample. The ANOVA found no significant effect of RSI condition, $F(2, 105) = .773$, $p = .464$, or significant interaction between RSI condition and strength, $F(2, 105) = 1.816$, $p = .168$. Neither the effect of RSI condition nor the interaction contributed much at all to variance in responding ($\eta^2 = .015$ and $\eta^2 = .033$, respectively).

1.3.1.2 Discussion

Participants in each of the three conditions were expected to demonstrate varying rates of contraction in the difference between their FAR for strong and weak items between the 4-key and 2-key phases. The participants who transitioned to a 2-key response with no RSI were expected to display the same inability to shift their criteria as was seen in prior experiments like Starns and Olchowski Experiment 1A or Experiment 2 (submitted), because they would be transitioning into the maximum possible response interference. Participants who transitioned from the 4-key condition to a blank RSI of 1200 ms were expected to experience some reduction in their response interference, thereby allowing some degree of criterion-shift which should be evident as a difference between strong and weak FAR. Although they were still required to respond "studied" using the same key regardless of item strength, the blank screen between test items was expected to allow them time to separate each response from the next. This separation was expected to alleviate some of the putative response interference because each response would take place in at least a slightly different context from the last due to the responses overlapping less in time. Participants who were provided with both an RSI of 1200 ms and a preview of item strength were expected to have their response interference significantly reduced and therefore provide an even larger criterion-shift. The label denoting the strength category of the next item combined with extra time to re-orient the response was expected to allow participants to treat the next "studied" response as either a "strong studied" or "weak studied" response as they would be

aware of the category before the response needed to be made. This was expected to provide evidence of the effect being due to response interference as well as demonstrating that response interference was continuous.

The requirement to respond with only 2 keys in this case did cause participants to abandon trial-by-trial criterion-shifting as predicted but did not show a continuous impairment in criterion-shifting that could be attributed to differing levels of response interference. However, the predicted indication of continuous response interference that could be mediated by experimental manipulations did not appear. Response interference may be at work only in certain circumstances, or a different process might underpin the presence or absence of criterion-shifting.

1.4 Experiment 3

In our third experiment we examined blocked and unblocked test lists with both 2-key and 4-key conditions in order to determine whether number of response keys and blocking affect criterion-shifting in the same underlying manner. We hypothesized that participants would be able to shift their criteria trial-by-trial in both blocked and unblocked conditions when provided with 4 response keys, but would be unable to shift their criteria in the unblocked 2-key condition. Blocked presentations of strong and weak-cued items had previously been shown to allow criterion-shifting based on strength in experiments using only 2 keys for responding (Verde & Rotello 2007, Hicks & Starns, 2014). However, experiments have also shown that as the blocks become smaller the ability to shift is eliminated (Stretch & Wixted, 1998). We expected that this inability to shift could be explained as the effect of response interference. We hypothesized that as participants were required to switch from “strong studied” to “weak studied” with greater frequency, they become unable to consider these responses separately and are thus unable to shift their criterion to one that is appropriate for each response. However, we predicted that if participants were given 4 keys to respond, they would have no trouble with either blocked or

unblocked presentations. As presentation becomes more random, 4-key participants would still have two separate keys with which to say “strong studied” and “weak studied”. Therefore, we expected a significant difference in the FAR between strong and weak test items in both of the 4-key conditions and in the blocked 2-key condition. We expected no significant difference in FAR between the strong and weak test items in the 2-key unblocked condition. As a consequence of these predictions, we expected an interaction between FAR difference and key condition to appear.

1.4.1 Methods

Participants. Participants were University of Massachusetts undergraduate students. One hundred sixty-two were randomly assigned to each of the blocked and unblocked 2-key and 4-key conditions, resulting in 41 blocked 2-key participants, 41 blocked 4-key participants, 39 unblocked 2-key participants and 41 unblocked 4-key participants. Ten low-performance participants were removed, either due to an overall accuracy below 60% or an accuracy below 90% on the Response Practice portion. This left 41 2-key blocked, 38 4-key blocked, 37 2-key unblocked, and 36 4-key unblocked participants in the final data. The higher proportion of dropped participants in the 4-key conditions (eight of ten) is likely due to the harder response practice portion assigned these participants. In the 4-key response practice condition participants must switch their responding based upon which side of the screen an item appears. This task might be more difficult when it is first introduced.

Materials. Stimuli in all conditions consisted of nouns, verbs, and adjectives randomly selected from the same pool of 859 low-frequency words as experiments 1 and 2. Forty words were assigned to a short practice phase and 300 words were assigned to the experimental phase of each condition. Each practice study list consisted of 10 strong targets and 10 weak targets. All of the strong targets were presented five times over the course of the list and the weak targets were

each presented a single time, for a total of 60 presentations in each practice phase. The practice test items consisted of all the words from the practice study list plus 20 lure items, 10 lures each for strong and weak cues.

The study list for the experimental phases consisted of 40 strong target words and 40 weak target words presented in a random order. The strong target words were each presented five times, and the weak target words were presented a single time each for a total of 240 presentations. The test lists included all the words on the study lists as well as 80 lure items, with 40 presented with a strong cue and 40 with a weak cue for a total of 160 test items. In the unblocked conditions the orders of the study and test lists were randomized for each participant. In the blocked-presentation conditions the test words appeared in blocks of 20 items, alternating between strong items and weak items. The strength category that began the blocked presentations was randomized for each participant. Items were presented in exactly the same manner in all conditions, with the exception of blocking. Every participant in the blocked conditions was informed that the items at test would be presented in blocks.

Design and procedure. Participants were randomly assigned to each of the four conditions. Participants were presented with instructions to read on the screen that the experimenter read aloud at the same time. Participants were given a practice phase at the start of the experiment designed to acquaint them with the responses they would need at test. This phase involved seeing 25 presentations each of the possible item types for a total of 100 presentations. These were not actual words, but phrases representing items, such as “studied five times” for an item studied five times. Phrases were used in place of the symbols used in the prior experiments, as it was found to reduce confusion in participants during the explanation of the response practice portion. Participants were instructed to respond as though these were actual test items with the appropriate response key, and were shown an error message if they pressed an incorrect response

key.

Participants in all four conditions completed a brief practice phase before the experimental phase during which they had the opportunity to go through a full, although short, study/test cycle and ask any questions they might have. In each condition, participants were then presented with a list of study words and asked to pay attention to the list because their memory for the words would be tested later. Study items remained on the screen for 900 ms, with 100 ms of blank screen between each item. For the study and test lists, all of the words were presented in white text on a black computer screen.

Following this distraction task, participants were presented with the test list. In all conditions, strong items were presented on the right hand side of the screen while weak items were presented on the left hand side of the screen. Test items were presented at random in the unblocked conditions and presented in 20-item blocks in the blocked conditions. Participants in the unblocked conditions received instructions identical to those seen in Experiment 2. Participants in the blocked conditions received identical instructions with the addition of a line informing them that test items would appear in blocks of 20 items.

1.4.1.1 Results

Table 3 shows the p("yes") for strong and weak targets and lures for participants in each condition of Experiment 3. In all four conditions, p("yes") was higher for strong-cued targets than weak-cued targets. For all conditions excepting unblocked presentation with two response keys, the p("yes") for weak-cued lure items was higher than the p("yes") for strong-cued lure items.

Table 3
Proportion of "Yes" Responses in Experiment 3

	Blocked Presentation				Unblocked Presentation			
	Strong Target	Weak Target	Strong Lure	Weak Lure	Strong Target	Weak Target	Strong Lure	Weak Lure
4-key	0.81 (.01)	0.60 (.01)	0.19 (.01)	0.29 (.01)	0.82 (.01)	0.6 (.01)	0.16 (.01)	0.24 (.01)
2-key	0.81 (.01)	0.57 (.01)	0.17 (.01)	0.22 (.01)	0.85 (.01)	0.59 (.01)	0.22 (.01)	0.23 (.01)

Standard errors are noted in parenthesis.

A planned contrast of the difference in p("yes") for strong vs weak lure items (FAR difference) produced the expected result of mean FAR difference being far lower in the unblocked 2-key condition than any other in the experiment. The mean FAR difference of the unblocked 2-key condition (M = .017, SD = .09) was contrasted with the average of the blocked 2-key (M = .081, SD = .12), blocked 4-key (M = .101, SD = .13), and finally unblocked 4-key (M = .435, SD = .09) conditions. It was revealed that the mean difference in FAR was significantly lower in the blocked 2-key condition, $t(148) = 2.776$, $p < .01$. Cohen's d for this contrast indicates a mild effect ($d = 0.45$).

As in our Experiment 2 analysis, we also analyzed response criterion estimates as well as applying a correction for values of zero in the data in order to get meaningful criterion measures for those values (Snodgrass & Corwin, 1988). A similar planned contrast of the difference in strong vs weak lure item criteria in produced the expected result of the mean difference being far lower in the unblocked 2-key condition than any other in the experiment. The mean criterion difference of the unblocked 2-key condition (M = .108, SD = .39) was contrasted with the average of the blocked 2-key (M = .163, SD = .38), blocked 4-key (M = .361, SD = .50), and finally unblocked 4-key (M = .308, SD = .44) conditions. It was revealed that the mean difference was significantly lower in the blocked 2-key condition, $t(148) = 2.063$, $p < .05$. Cohen's d for this contrast indicates a moderate effect ($d = 0.34$).

A 2(strength) × 2(presentation condition) × 2(key condition) ANOVA on the p('yes") for target items revealed a significant effect of strength, $F(1, 148) = 477.154$, $p < .001$. This effect of strength accounts for the majority of the variance in the data ($\eta^2 = .763$). But the ANOVA revealed no significant effect of presentation, $F(1,148) = .492$, $p = 0.484$, or key condition, $F(1,148) = .033$, $p = .856$. Neither of these main effects contribute more than an extremely small

amount to variance in the data (η^2 = .003 and η^2 = .000 respectively). In addition no significant interaction was found between strength and key condition, $F(1,148) = 2.875$, $p = .092$, between the presentation and the key condition, $F(1, 148) = .299$, $p = 0.586$, between strength and presentation, $F(1,148) = .434$, $p = .511$, or between strength, presentation, and key condition, $F(1,148) = .187$, $p = .666$. None of these interaction effects displayed a partial eta-squared suggesting they explain much of the variance in p("yes") (η^2 = .019, η^2 = .002, η^2 = .003, and η^2 = .001 respectively).

1.4.1.2 Discussion

The participants in each condition save for the unblocked 2-key condition were expected to demonstrate a significant difference between their FAR for strong and weak items at test. The participants who were placed in the 2-key unblocked condition were expected to display the same inability to shift trial-by-trial seen in the 2-key condition of Starns and Olchwski (submitted, exps 1A & 2). Participants in each of the other three conditions were expected to display the ability to make trial-by-trial criterion shifts as evidenced by a significant difference between strong and weak FAR. All conditions save unblocked 2-key should have had significantly less response interference either due to physical difference which is unaffected by blocking, or due to differences in timing of responses which is affected by how large the blocks are. This prediction was borne out by the planned contrast between unblocked 2-key and the average of the 3 other conditions.

For participants with a sufficient number of response options, blocking was not expected to improve or degrade performance of the criterion-shift. Participants using only two keys were expected to significantly improve their performance of the criterion-shift so long as test items were blocked. If these factors influenced criterion-shifting, it would suggest that participants can shift easily without response interference and are able to overcome it when given sufficient

opportunity to re-orient their responding and remove overlap between strong and weak "studied" responses via blocking test items by strength. This experiment offered no significant evidence for blocking adding any additional performance increase to the 4-key condition.

This set of results demonstrates in a controlled randomized experiment that the number of keys used for responding is a relevant factor controlling the ability to shift, which was the predicted result. However, along with the previous experiments discussed it does not support the initial hypothesis that response interference is the driving force behind the criterion shift observed in both blocking and key manipulation experiments. Experiment 1 found 2-key participants were capable of trial-by-trial criterion-shifting when forced and Experiment 2 found that manipulations reducing response interference had no effect on ability to criterion-shift. The current experiment found that while both blocking and key manipulation affected ability to criterion-shift as predicted, there was not the expected interaction between the two. Given these findings it is probable that something other than response interference is behind the effects blocking and number of keys have on the ability to criterion-shift. Based on this experiment, as well as the experiments detailed above, response interference is not likely the underlying factor driving criterion-shift.

1.5 General Discussion

These experiments were designed to help uncover what allows participants to shift their criterion trial-by-trial using arbitrary strength cues and to provide an explanation for the process that can prevent this criterion-shifting from happening. Prior experimentation had already revealed several factors that influence ability to shift. For instance, participants require some cue at test to differentiate between strength categories. If all they have is an internal measure of the memory strength elicited by the test item but no cue (such as color-coding) to act as an indicator of whether that value was produced in response to a weak or a strong item, participants are unable

to shift without being given test items in blocks, with extensive feedback, or both (Verde & Rotello, 2007; Rhodes & Jacoby, 2007). We know that participants are also better able to shift if they are made aware of the presentation manipulation. If participants are told explicitly that red is strong and green is weak, they do not need to go through a laborious process of learning the association themselves before criterion-shifting (Starns & Olchowski, submitted).

We posited that participants also need the "space", actual or conceptual, to avoid overlapping responses that lead to sustained response interference and an inability to shift effectively in unblocked strength-cued trials without additional assistance (time, semantic category markers, extensive feedback, etc). We predicted that without sufficient response space, participants would collapse their responding along a single middle-ground criterion regardless of the other sources of information to which they were exposed. We believed that with sufficient response space even minimally informative testing environments would give rise to accurate and efficient criterion-shifting on an trial-by-trial basis, and investigated this hypothesis in-depth over the course of this experimental series.

The experiments described above were designed to provide a clear method of evaluating whether or not response interference was in fact the driving force behind the behavior of participants in strength-based criterion-shift experiments. These experiments taken as a whole provide a strong indicator that response interference plays little if any role in the behavior observed in strength-based criterion-shift experiments. Our experimental methodology took the inducement of criterion-shifting as a given, due to the previous successes we have had with inducing shifts using multiple response keys (Starns & Olchowski, submitted). This series of experiments was thus able to put the focus squarely on whether or not manipulation of conditions germane to response interference would in any manner affect the ability of participants to shift. In each case it was found that there was no significant impact of those experimental manipulations

on the ability of participants to criterion-shift.

These investigations have served to further expand our knowledge of the criterion-shift process and provided us with a more accurate understanding of memory. The fact that response interference did not appear to affect the ability of participants to perform a criterion-shift raises new questions about what constitutes the most salient difference between 2-key and 4-key conditions and what promotes or impedes trial-by-trial criterion-shifting. This series of experiments has made a clear case that these differences lie somewhere other than our initial hypotheses supposed.

One possibility that would make sense in light of the data seen in previous experiments showing an attentional aspect to criterion-shifting (DeCarlo, 2007; Rhodes & Jacoby, 2007 ; Verde & Rotello, 2007) is that participants are actually being asked in each of these experiments to make two separate decisions before responding. The first decision involves figuring out which cue is present for a given item, and the second involves deciding whether a test item is old or new. Participants need to make both decisions in order to effectively shift their criterion. The first decision is mediated somewhat by the presence and type of cue. Blocking could be considered a relatively ineffective cue, as it requires a significant amount of time to begin associating items presented in a block with a given strength. Color on the other hand would be an example of a cue that is quickly and effectively grasped by a participant, particularly if the colors chosen are very distinct. What underpins the variety of results seen in previous experiments and in this paper might be the fact that this cue decision is optional in most cases while the old/new decision is not.

It is possible that participants default to the strategy that requires least effort given no inducement to the contrary. It may be important to note that in most criterion-shift experiments, criterion-shifting does not actually improve accuracy to a great degree, as was described in the discussion portion of Experiment 1. In the absence of detailed feedback, the difference between

performing and not performing a criterion-shift is minimal for a participant. Given that appropriately noting the strength cue prior to responding rarely improves performance as measured by overall accuracy visibly, participants might be ignoring strength cues and responding solely by evaluating the memory strength of each item, effectively picking a middle-ground criterion. If this is the case, it would stand to reason that the underlying explanation for the presence of criterion-shift in some experimental conditions is that participants are forced to make the cue decision. Given four response keys, participants do not have the option of not making a cue decision. If they were to respond as though the cue was irrelevant and chose to use either the left or the right-hand response keys at random they would encounter a number of error messages and take much longer to complete the task. These error messages, rather than the accuracy feedback, could be considered the relevant feedback for keeping participants on-track in their criterion-shifting. Ultimately, this series of experiments considered in concert with previous research may point to criterion-shifting being an essentially automatic process that is nonetheless gated by a decision that is often optional and that participants are often not interested in making unless forced. We hypothesize that participants will naturally adopt appropriate criteria in these experiments if the cue decision is somehow forced upon them in the test environment. This can consist of the sort of forced criterion-shifting seen in Experiment 1, where participants were expressly told to make the cue decision at each test item, or it can be left to participants if a random response is actively more trouble than acknowledgment of the existence of categories cued in some sufficiently obvious manner as seen in Experiments 2 and 3.

Criterion-shifting is a useful window into a complex aspect of memory function with important consequences for everyday life. Memory depends on the ability to set a standard that matches the event being described. For example, one should require extreme amounts of evidence evidence to answer “yes” to the question “Did you see that dinosaur on Main Street?” while

answering “yes” to “Did you see that car on Main Street?” would require little. Remarkable events require remarkable amounts of evidence to support their having occurred. It offers a simple method to understand the decisions people make about the strength of their memories. Understanding why participants will only shift their criterion when presented with test information under certain conditions will help us construct a more accurate model of memory as a whole. Though it seems obvious to suppose that participants are making judgments of memory strength simply via getting a “reading” on the strength of evidence a given item elicits, prior work in criterion-shifting makes it clear that this is not an accurate way to interpret the process. Participants with exactly the same ability to differentiate between strong and weak items at test, and even participants made explicitly aware of the relationship between cues and test item strength, behave in vastly different manners depending on minor alterations to experimental conditions. We thought it was possible that this behavior could be explained as a function of how free the participant is to respond in a manner that allows them to establish multiple criteria. Our current set of experiments has set us on the path to a more coherent understanding of how this response environment is interacting with the ability to criterion shift, allowing us to make better sense of how human beings understand their own memories.

BIBLIOGRAPHY

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

Benjamin, A. S., Diaz, M., & Wee, S. (2009) Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116*, 84-115.

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.

Cary, M., & Reder, L. M., (2002). A dual-process account of list-length and strength-based mirror effects in recognition. *Journal of Memory and Language, 49,* 231-248.

Clare, J., & Lewandowsky, S. (2004) Verbalizing Facial Memory: Criterion Effects in Verbal Overshadowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 30*, 739-755.

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*, 484-499.

Curran, T., Debuse, & C., Leynes, P. A. (2006) Conflict and Criterion Setting in Recognition Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 33*, 2-17.

DeCarlo, L. T. (2007) The Mirror Effect and Mixture Signal Detection Theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 33*, 18-33.

Dobbins, I. G., & Kroll, N. E. A. (2005) Distinctiveness and the Recognition Mirror Effect: Evidence for an Item-Based Criterion Placement Heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 31*, 1186-1198.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13*, 8-20.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The Regularities of Recognition Memory. *Psychological Review, 100,* 546-567.

Hicks, J. L., & Starns, J. J. (2014). Strength cues and blocking at test promote reliable within-list criterion shifts in recognition memory. *Memory & Cognition*, *Online First Publication.*

Hirshman, E., & Master, S. (1997). Modeling the conscious correlates of recognition memory: Reflections on the remember-know paradigm. *Memory & Cognition,* 345-351.

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., Huber, D. E., Marinelli, K. (1995) Effects of Category Length and Strength on Familiarity in Recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* 21(2), 267-287.

Simon, J. R., Acosta, E., Mewaldt, S. P., & Speidel, C. R. (1976). The effect of an irrelevant directional cue on choice reaction time: Duration of the phenomenon and its relation to stages of processing. *Perception & Psychophysics, 19*, 16-22.

Simon, R., & Berbaum, K. (1990) Effect of Conflicting Cues on Information Processing: the 'Stroop Effect' vs. the 'Simon Effect'. *Acta Psychologica, 73,* 159-170.

Singer, M. (2009). Strength-based criterion shifts in recognition memory. *Memory & Cognition, 37*, 976-984.

Singer, M., Gagnon, N., & Richards, E. (2002) Strategies of Text Retrieval: A Criterion Shift Account. *Canadian Journal of Experimental Psychology. 56*, 41-57.

Singer, M., & Wixted, J. (2006). Effect of delay on recognition decisions: Evidence for a criterion shift. *Memory & Cognition, 34*, 125-137.

Starns, J. J., & Olchowski, J. (submitted). Shifting the criterion is not the difficult part of trial-by-trial criterion shifts in recognition memory. Manuscript submitted to *Memory and Cognition*.

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.

Starns, J. J., White, C. N., & Ratcliff, R. (2012). The strength-based mirror effect in subjective strength ratings: the evidence for differentiation can be produced without differentiation. *Memory and Cognition, 40*, 1189-1199.

Stretch, V., & Wixted, J. (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.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34-50

Tanner, W. P. Jr., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61*, 401-409.

Treisman, M., & Williams, T. C. (1984). A Theory of Criterion Setting With an Application to Sequential Dependencies. *Psychological Review, 91,* 68-111.

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review. 11,* 616-641.

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