Why Are Trial-by-Trial, Strength-Based Criterion Shifts Hard to Observe? Is the Difficulty in the Mental Process Itself or in the Typical Cued Criterion Method?

JERWEN JOU, ERIC E. ESCAMILLA, ANDY U. TORRES,
ALEJANDRO ORTIZ, and MARIA S. MATOS
University of Texas–Rio Grande Valley

In the cued criterion recognition paradigm (Stretch & Wixted, 1998b), trial-by-trial memory strength–based criterion shifts have been an elusive phenomenon. Often the criterion shifts do not occur. We suggest that the frequent failure in making criterion shifts in the literature is caused by participants' failure to understand the rationale of the task as typically presented in an abstract format. In this study, participants studied words once or thrice and were asked at test to either classify the probes into "new," "seen once," or "seen 3 times" categories by pressing the corresponding keys or to make an old/new binary decision followed by an item presentation frequency, confidence, or a memory quality judgment. No memory strength cues were provided, and only one set of new items served as distractors for strong and weak targets. Robust trial-by-trial criterion shift was observed. We concluded that no cues distinguishing between strong and weak probes are necessary for obtaining this type of criterion shift when the tasks are designed to make good pragmatic sense for the participants. The reason why this type of criterion shift is typically hard to obtain in the cued criterion paradigm is not that the process itself is difficult but that the cued criterion method is hard for the participants to understand.

KEYWORDS: criterion setting in recognition, criterion shifts in recognition, strength-based criterion shift, mixed-strength list test

Decision making and criterion setting are inseparable parts of memory in general and of recognition memory in particular (Cho & Neely, 2013; Healy & Jones, 1973; Healy & Kubovy, 1978; Hirshman, 1995; Jou et al., 2018; Jou, Flores, Cortex, & Leka, 2016; Koriat & Goldsmith, 1996). Two factors can influence recognition memory: sensitivity and decision criterion (also known as response bias). Sensitivity is the true ability to discriminate an experienced event from an unexperienced event. The response bias or criterion is a tendency toward responding positively or negatively to the stimulus, which is supposed to be independent of sensitivity (Green & Swets, 1966). The decision criterion is a threshold that observers place somewhere along the strength-of-evidence continuum, which they use to evaluate the test item for a "yes/no" (or "old/new") decision. If the test item affords an amount of evidence that meets or exceeds that threshold, a "yes" ("old") response is made. If the accrued evidence fails to meet the criterion, a "no"

American Journal of Psychology
Winter 2020, Vol. 133, No. 4 pp. 453–471 • © 2020 by the Board of Trustees of the University of Illinois

This content downloaded from
75.14.0.129 on Thu, 14 Jan 2021 19:51:39 UTC
All use subject to https://about.jstor.org/terms

("new") response is made. When the criterion is set high, the observer needs more evidence for a positive response. When it is set low, the observer needs less evidence for a positive response. When the criterion is set high, fewer items fall above the criterion, and consequently, fewer hits and fewer false alarms (FAs) are made, and vice versa. Thus, when a criterion is shifted from a lower point to a higher point on the strength-of-evidence dimension, the false alarm rate (FAR) will decrease and the hit rate (HR) will also decrease if the memory strength of the studied items remains unchanged (Figure 1).

In signal detection theory (Green & Swets, 1966; Macmillan & Creelman, 2005), recognition accuracy or sensitivity is measured as a d' score, which is the z score distance between the means of the memory strength distributions for the new and old items (see Figure 1). The d' is calculated from the HR and the FAR. The higher the HR and the lower the FAR, the better the discrimination between the old and new items, and hence the higher the sensitivity or d'. Also, observers tend to use a higher criterion for items with better-quality or stronger memory evidence, typically as a result of a stronger encoding (if something is well learned, people need more evidence for a "yes" response). In contrast, when items afford weak memory evidence because of weaker encoding, observers tend to use a more lax criterion (based on less memory evidence) for a "yes" response (Healy & Jones, 1973; Hicks & Marsh, 1998; Hirshman, 1995; Jou et al., 2016; Singer, 2009; Singer & Wixted, 2006; Verde & Rotello, 2007).

Previous studies have shown that lists of words that are all strongly encoded (e.g., with repeated or longer presentations) produce a higher HR and a lower FAR than lists of words that are all weakly encoded (with fewer or briefer presentations) (e.g., Hirshman, 1995; Stretch & Wixted, 1998b, Experiment 1). Because strongly encoded lists produce a higher HR but a lower FAR than weakly encoded lists, the reversal of the relative magnitudes of these two measures across the two encoding conditions is known as the strength-based mirror effect (Bruno et al., 2009; Kilic & Oztekin, 2014; Stretch & Wixted, 1998b). It is attributed to the shifting of the criterion from a lower (more lax) memory strength value for the weakly encoded item distribution to a higher (stricter) value for the strongly encoded item distribution

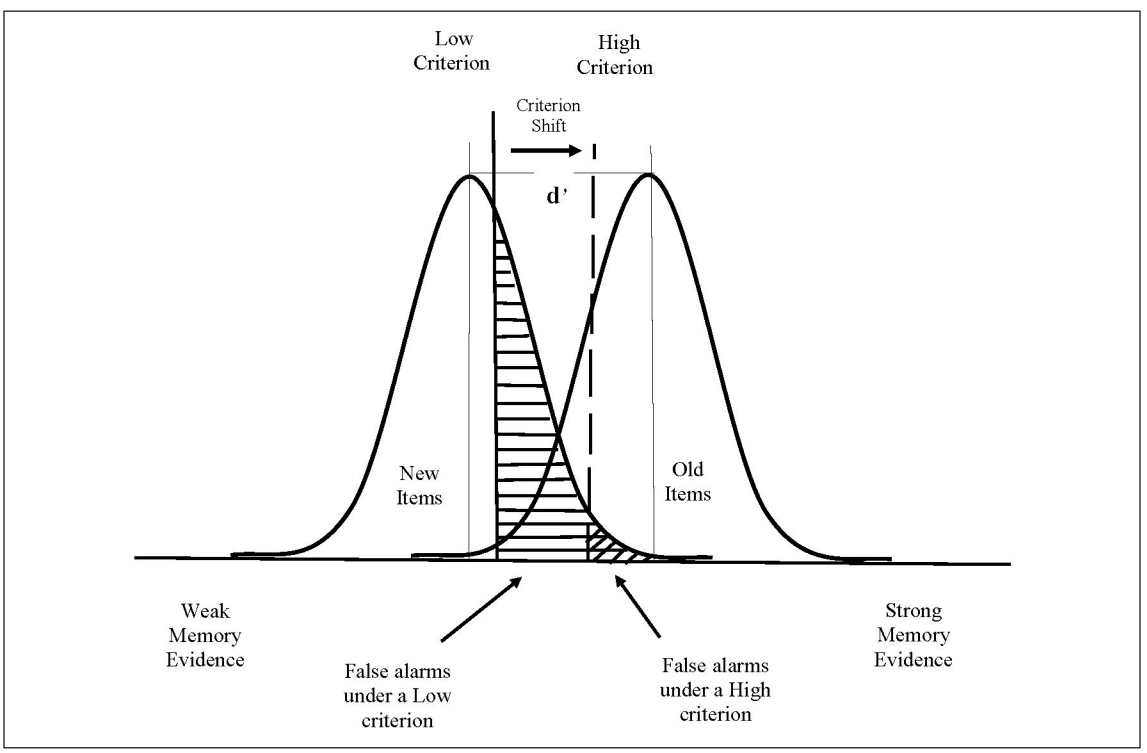

FIGURE 1. The decrease in false alarms as a result of an upward decision criterion shift

This content downloaded from
75.14.0.129 on Thu, 14 Jan 2021 19:51:39 UTC
All use subject to https://about.jstor.org/terms

on the strength-of-evidence axis (Stretch & Wixted, 1998b).

According to this view, the location on the memory strength axis of the new item distribution does not change because increasing the memory strength of the old items should not change the familiarity level of the new items.[1] However, under this circumstance the HR will increase only if the effect of the enhanced encoding more than offsets the decrease in HR due to the rightward shift of the criterion. Therefore, the sure consequence of a criterion shift from a lower to a higher value is the shrinking of the area of the new item distribution that falls above the criterion, that is, the FA area (again, assuming that the prior familiarity level of the new items stays the same). See Figure 2 for an illustration. When the FARs of the strongly and weakly encoded items are the same, this result is taken as evidence of the absence of a criterion shift (Stretch & Wixted, 1998a).

However, typically the strength-based mirror effect occurs reliably only when the encoding manipulation is conducted between lists (between-list or pure list manipulation in which all the words in a list are encoded either uniformly strongly or weakly and tested in separate tests) (Stretch & Wixted, 1998a). When the strongly and weakly encoded words are mixed together in one list at study and tested within a single test (referred to as the within-list or mixed-list manipulation), the findings about criterion shifts have been mixed (for a review of the mixed findings, see Bruno, Higham, & Perfect, 2009, or Starns & Olchowski, 2015). In many cases, although the HR of the strong encoding condition increased, the FARs for the strongly and weakly encoded items were the same, indicating an absence of a strength-based, trial-by-trial criterion shift in the mixed list condition (Higham et al., 2009; Kent et al., 2018, Experiment 3 using unrelated words among other stimuli; Morrell et al., 2002, Experiment 1; Singer, 2009, under a rote learning condition; Singer & Wixted, 2006, Experiments 1 and 2; Starns et al., 2006; Stretch & Wixted, 1998a Experiments 2–5).[2]

The question of why people are disinclined to make a strength-based, trial-by-trial criterion shift when strongly and weakly encoded items are intermixed within the same list has puzzled many researchers and generated debates and competing explanations (again, see Bruno et al., 2009, or Starns & Olchowski, 2015). One explanation for the lack of evidence of a within-list, strength-based criterion

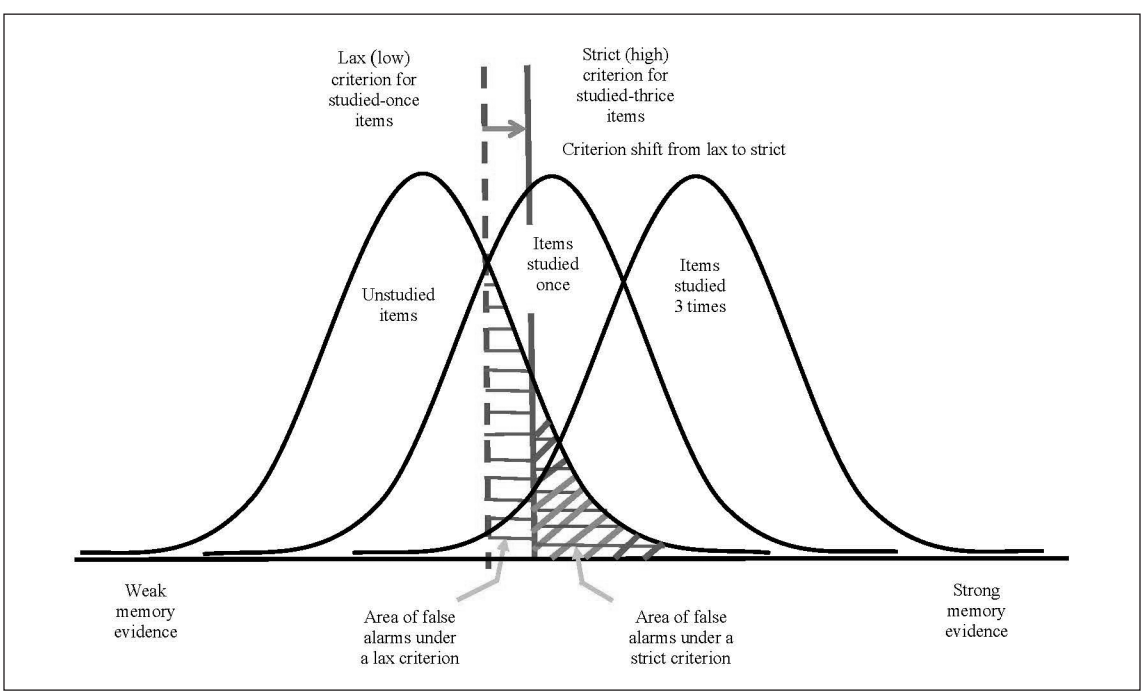

FIGURE 2. A strength-based upward criterion shift from "1" responses to "3" responses in the test word sorting tasks, Experiments 1 and 2

This content downloaded from
75.14.0.129 on Thu, 14 Jan 2021 19:51:39 UTC
All use subject to https://about.jstor.org/terms

shift is that a trial-by-trial criterion shift is a cognitively effortful process and that participants lack the incentive to make a constant recalibration of their decision criterion, resulting in their adopting a single criterion for both strongly and weakly encoded items in the same list and test (Dobbins & Kroll, 2005; Gillund & Shiffrin, 1984; Hockley & Niewiadomski, 2001; Morrell et al., 2002; Rhodes & Jacoby, 2007; Stretch & Wixted, 1998a). Another view attributes the absence of trial-by-trial criterion shifts to participants' unwillingness to process the memory strength cues provided during studying and testing, for example, with the new items associated with the strongly encoded targets displayed in red and those associated with the weakly encoded targets displayed in green (Starns & Olchowski, 2015). A third account attributes it to a subjective perception of memorability of the studied items. According to this account, when people feel confident about their memory, they ignore the memory strength cues at test and therefore do not shift criteria; when they do not feel confident about their memory (e.g., as a result of a large number of nonwords on the study list or of a very weak encoding), they pay attention to the memory strength cues during a test as a guide for their responses and therefore make criterion shifts (Bruno et al., 2009). A fourth account is that the information available at test is not relevant or useful for the purpose of making accurate responses. When a corrective feedback that is informative as to the accuracy of performance is provided, participants make trial-by-trial criterion shifts (Kent et al., 2018; Verde & Rotello, 2007).

Our View on the Issue of the Trial-by-Trial, Strength-Based Criterion Shifts

Changing decision criteria on a case-by-case (trial-by-trial) basis is actually a common practice in daily life. For example, when teachers assign grades A, B, C, D, and F to students' work, what they do is equivalent to using different fixed criteria to evaluate individual pieces of work. Similarly, when employers rank order job candidates on the basis of their qualifications, they actually judge each individual candidate with a different criterion. When farmers sort eggs into grades A, B, and C, they are using a different criterion for each egg. Because people have a well-established schema of sorting or grading performance or objects based on the qualities of these items, they should not find it difficult to perform a classification task for a memory event according to the individual item's memory quality or strength. So why is it so difficult to achieve strength-based, trial-by-trial criterion shifts in the Stretch and Wixted (1998b) cued criterion paradigm, and how plausible are the aforementioned accounts for the frequent failures in obtaining the mixed-list criterion shifts?

A review of the literature indicates that when words of different frequencies (Stretch & Wixted, 1998a), sentences from stories (Kintsch, Welsch, Schmalhofer,& Zimny, 1990), and taxonomic category words (Singer, 2009; Singer & Wixted, 2006) are used in the study and as test materials, strength-based (including differential strengths resulting from study–test delay manipulation), trial-by-trial criterion shifts are often obtained. Other studies found trial-by-trial criterion shifts when the quality or other attributes of the test probes (instead of the encoding strength or quality) were manipulated. For instance, when remembered scenes were either degraded or not degraded in a recognition test, Kent et al. (2018) observed trial-by-trial criterion shifts, with higher criteria used for intact test scenes and lower criteria used for degraded scenes. Similarly, in a study where participants studied faces without wearing sunglasses but were tested with some faces without sunglasses and some faces with sunglasses, a mirror effect was obtained, signaling trial-by-trial criterion shifts (Hockley et al., 1999). Also, when the base rate of the target word occurrence at test was manipulated and cued, trial-by-trial criterion shifts were observed (North et al., 2018). If the failure in trial-by-trial criterion shifts is caused by this process' high cognitive demands, then why did this type of shift occur in the aforementioned studies?

On the other hand, when unrelated words are used as study materials and differential repetitions as strength manipulations, mixed-list criterion shifts often fail to obtain. For example, Stretch and Wixted (1998a) in their Experiments 2–5 used a mixed-list strength manipulation in which strong and weak items were mixed together in the same list, coupled with a cued criterion procedure in which the old and the new items were each divided into two differently colored "strong" and "weak" subsets. They consistently failed to obtain trial-by-trial, strength-based criterion shifts. Thus, whether or not mixed-list,

This content downloaded from
75.14.0.129 on Thu, 14 Jan 2021 19:51:39 UTC
All use subject to https://about.jstor.org/terms

strength-based criterion shifts are observed seems to be contingent on the type of the materials and the method used. As will be shown later in the article, however, the type of materials may not be the main cause for the lack of mixed-list, trial-by-trial criterion shifts.

As noted, new items of the same word frequency or of the same taxonomic category as the old items and unread sentences from the same story as the read sentences can be recognized as belonging to the same class of words or to the same story, based on the semantic or thematic relatedness. When the manipulation of memory strength is made between different taxonomic categories or between different stories, the semantic themes of the categories or stories provide useful cues to participants for the strength-based class identification of the new items. But there are no such conceptual or semantic attributes in unrelated new words that can bind the new and old words into one memory strength–based class of materials. Different coloring (Stretch & Wixted, 1998a) is a perceptual feature and an ineffective means of grouping the new and differentially strengthened old words into one strength-based class. The fact that when semantic or frequency information (or other kinds of manipulations as in the aforementioned studies) is used, mixed-list criterion shifts often occur suggests that it may not be because the trial-by-trial criterion shifting itself is a cognitively laborious process but because the color or other perceptual cues used to mark the unrelated words as strength-based classes cannot effectively convey the information of the strength class membership with which the new words are cued to be associated. However, the semantic unrelatedness between the targets and distractors may not be the only or major cause of the lack of strength-based, trial-by-trial criterion shifts. The other possible factor leading to the failure of the trial-by-trial criterion shifts in the cuing paradigm may be the way Stretch and Wixted's (1998a) paradigm operationally defines criterion shifts.

To our knowledge, the aforementioned cued criterion method was first used by Stretch and Wixted (1998a) and later widely adopted by many researchers investigating this issue, and it has become a benchmark for determining the presence or absence of strength-based, trial-by-trial criterion shifts (Bruno et al., 2009; Singer, 2009; Singer & Wixted, 2006; Starns & Olchowski, 2015; Stretch & Wixted, 1998a). Some researchers studying the subject of strength-based, trial-by-trial criterion shifts elaborated the original cuing procedure and instructions to a painstakingly detailed level. For example, in their experiments Starns and Olchowski (2015) told their participants, "Some of the test words will appear in green on the left side of the screen. Words in green will either be words that you studied a single time on the study list or words that were not on the study list at all. You will hit the 's' key if you think the word was studied once or the SPACE bar if you think the word was not studied at all. Some of the words will appear in red on the right side of the screen. Words in red will either be words that you studied 5 times or words that were not studied at all. You will hit the 'l' key if you think the word was studied 5 times or the SPACE bar if you think it was not studied at all" (p. 53). As Koop et al. (2019) commented on these elaborate cuing procedures aimed at guiding participants to achieving the trial-by-trial criterion shifts, "However, the fact remains that differences in FAR are only observable on an item-by-item basis when significant affordances are provided. The affordances may include things like color cuing (Hicks & Starns, 2014; Starns & Olchowski, 2015; Stretch & Wixted, 1998), . . . or forcing individuals to acknowledge strength distributions through different response keys (Starns & Olchowski, 2015)" (p. 852). Thus, Starns and Olchowski, as well as some other researchers, seemed to go to great lengths to find a breakthrough within the Stretch and Wixted's (1998a) original framework. Although they obtained trial-by-trial criterion shifts in that study, participants might not know why they were asked to do such complex things.

It has been suggested that the crucial aspect of the method that has led to a frequent failure in observing trial-by-trial criterion shifts is its arbitrary dividing of the new item set into two subsets, with each subset marked with perceptual, nonconceptual features such as colors or test item display locations on a computer monitor screen. The intention is that participants will use these distinguishing features as a basis for adopting different criteria in judging these two subsets of new items. We will refer to this widely adopted method as the cued criterion paradigm. We suggest two possible reasons why the cued criterion paradigm has typically failed to produce strength-

This content downloaded from
75.14.0.129 on Thu, 14 Jan 2021 19:51:39 UTC
All use subject to https://about.jstor.org/terms

based, trial-by-trial criterion shifts. First, as noted earlier, colors and display locations are not natural "glues" that can cognitively bind the old and new words into a memory-strength-based class. Word frequency (Glanzer & Adams, 1990; Singer, 2009) and taxonomy (Cho & Neely, 2013), on the other hand, are conceptually based attributes of words that can more naturally serve as such a "glue" to bind the old and new words into one coherent, strength-based class of words. Consider an analogy from a social judgment task. We typically categorize people by features such as gender, age, race, and profession, to name a few. These features are "natural" social classifiers. In the cued criterion paradigm, researchers might have one randomly selected group of people wear red shirts (or stand at one location) and another group wear green shirts (or stand at another location) and then require that participants treat people dressed in one shirt color as a unified group having a specific social characteristic and treat people dressed in a different shirt color as another unified group having a different specific social characteristic. When this is done, people may very well fail to make the social classification even when the color or location differences are made very conspicuous. In other words, participants cannot understand and use the researcher's cuing features as intended.

We suggest that participants should be able to perform strength-based, trial-by-trial criterion shifting when the task makes sense to them. In many reasoning and judgment studies, participants are often found to perform the task poorly (making logically incorrect judgments) when the reasoning problem is presented in an abstract, formal, logical format. But when the problem is presented in the form of a pragmatic issue that participants can relate to their everyday life experiences, their performance greatly improves (Cheng & Holyoak, 1985; Cheng, Holyoak, Nisbett, & Oliver, 1986). For example, in Wason's (1966) card selection task, when participants were asked to turn over the minimum number of cards among the four cards ("4," "A," "7," "K") necessary to verify whether the rule "If a card has a vowel on one side, then it has an even number on the other side" was followed, only 4% of the participants turned over the two correct cards ("A" and "7"). However, when the logically equivalent problem was presented in the form of pragmatic permission rules (e.g., minimum age for drinking and required postage for a sealed letter), the performance improved to 72% correct (Griggs & Cox, 1982; Johnson-Laird et al., 1972). Another similar example concerns a preference reversal phenomenon known as the framing effect (Tversky & Kahneman, 1981), in which people make contrary decisions on the logically same problem depending on whether it is presented in a gaining or losing perspective. However, Jou et al. (1996) showed that the framing effect vanished when the judgment problem was instantiated in a scenario that participants could understand on the basis of their pragmatic life experiences. In short, a task must conform to a scenario that participants can relate to as something meaningful in their everyday life. Forming word strength classes between studied and unstudied semantically unrelated words by font color or display location may not be a very "user-friendly" format of presenting the intended logic.

We believe that another reason for the frequent failures to observe strength-based, trial-by-trial criterion shifts in the cued criterion paradigm is that participants may actually be making such criterion shifts in those experiments but that the researchers may be searching for the evidence "in the difficult-to-find-it place," so to speak. In the cued criterion paradigm, the absence of a strength-based, trial-by-trial criterion shift is inferred from the equal FARs between the two differently cued subsets of new items that come from exactly the same distribution of memory strength values. This equal FARs from the two arbitrarily designated new item subsets may not indicate a lack of trial-by-trial criterion shifts. We suggest that each "yes" response is based on a different amount of evidence supporting the positive response and that participants know whether a given "yes" response is a solid or a shaky one. We provide evidence for this assumption later in the article.

The Present Approach

Just as in everyday life the trial-by-trial criterion shifts can be easily found, they can be easily found in lab studies. That is, the commonly used confidence rating and memory quality assessment using remember/know judgments actually may demonstrate that people can perform trial-by-trial criterion shifts. If participants are aware that some of their "yes" responses in a test are based on more evidence than others and can indicate it, then they are internally us-

This content downloaded from
75.14.0.129 on Thu, 14 Jan 2021 19:51:39 UTC
All use subject to https://about.jstor.org/terms

ing different criteria for individual responses. In fact, some studies on frequency judgments (Hintzman, 2004a), judgments of recency (Hintzman, 2004b, 2005), and source monitoring (Hicks & Starns, 2006a, 2006b) apparently demonstrated trial-by-trial, strength-based criterion shifts even though these authors were focusing on other issues. For example, the appendix of Hintzman's (2004a) study (p. 350, Table A3) indicated that the recognition FAR in misidentifying a new item as having appeared multiple times was lower than the FAR in misidentifying a new item as having appeared once. Given that the single- and multiple-occurrence items were mixed in the test, this is evidence of strength-based, trial-by-trial criterion shifts.[3]

However, to our knowledge researchers focusing on the cued criterion paradigm never brought the aforementioned point to attention in the literature. Similarly, researchers taking the memory rating approach never applied the method to resolving the controversy of trial-by-trial criterion shifts. Judgments of frequency or recency (Hintzman, 2004a, 2004b, 2005) and source monitoring work (Hicks & Starns, 2006a, 2006b)[4] are relevant to this issue, but to our knowledge nobody has indicated in the literature that they are relevant to resolving the issue of trial-by-trial criterion shifts.

We proposed to assess the occurrence or nonoccurrence of strength-based, mixed-list criterion shifts by examining the frequency of FAs participants make when they report that their "yes" response is strong versus when they report their "yes" response is weak. This method does not involve dividing the new item distribution into two subdistributions and marking them with different colors. With this method, participants indicate whether a given FA is made as a result of recognizing a new item as a strong target (equivalent to using a strict criterion) or as a weak target (equivalent to using a lax criterion). The FAR will be assessed on the basis of an undifferentiated, single new item distribution by calculating the proportion of the "strong" and "weak" FAs, respectively, within the single new item set. We call this method the uncued criterion paradigm. We think that these procedures remove the major stumbling block in the cued criterion paradigm.

We used two procedures, a one-step and a two-step response, to make participants indicate which word class they think the accepted word belongs to. In Experiment 1, a one-step response was used. We designated different response keys for participants to use to sort the test items into three categories ("press '3' key if you think you saw the word 3 times at study, and '1' key if you saw the word once and '0' key if you think you did not see the word at study"). This method was basically a frequency judgment task (Hintzman, 2004a). By requiring the participants to sort the test items into the three classes, we were asking participants to "naturally" apply multiple response criteria by using memory strength of the test words as the basis of their response criterion. The crucial diagnostic index for a criterion shift now becomes whether participants make fewer FAs when they make "3" responses than when they make "1" responses. It does not matter to what specific new items the FA response is made. The FAR can be calculated by dividing the number of FAs in the "3" response category and that in the "1" response category, respectively, by the total number of new items. The crucial difference between this approach and the cued criterion paradigm is that in the cued criterion paradigm, the FAR of each class of items is based on the frequency of the FAs made to each differentially cued subset of new items, whereas in the approach we used, it is based on the frequency of *any* new items mistakenly classified either as a strong target or as a weak target. See the graphic representation of the single, whole new item set in Figure 3.

According to this notion, the cued criterion paradigm fails to detect the trial-by-trial criterion shifts because that paradigm is looking into the "difficult-to-find-evidence place" for evidence of criterion shifts. If, instead, participants were asked to give the strength class of an accepted item as they judged it, they would reveal the criterion they used in accepting the item. And the criteria that they reveal they adopt for different accepted items will be different.

We also used a two-step response procedure in Experiment 2 in which participants first made a traditional binary "yes/no" response and then a post-recognition frequency judgment. In Experiment 3, a "yes/no" response was followed by a confidence rating. In Experiment 4, a "yes" response was followed by a memory quality (remember/know) judgment. By asking participants to provide a post-recognition judgment on the quality of the accepted probe item,

This content downloaded from
75.14.0.129 on Thu, 14 Jan 2021 19:51:39 UTC
All use subject to https://about.jstor.org/terms

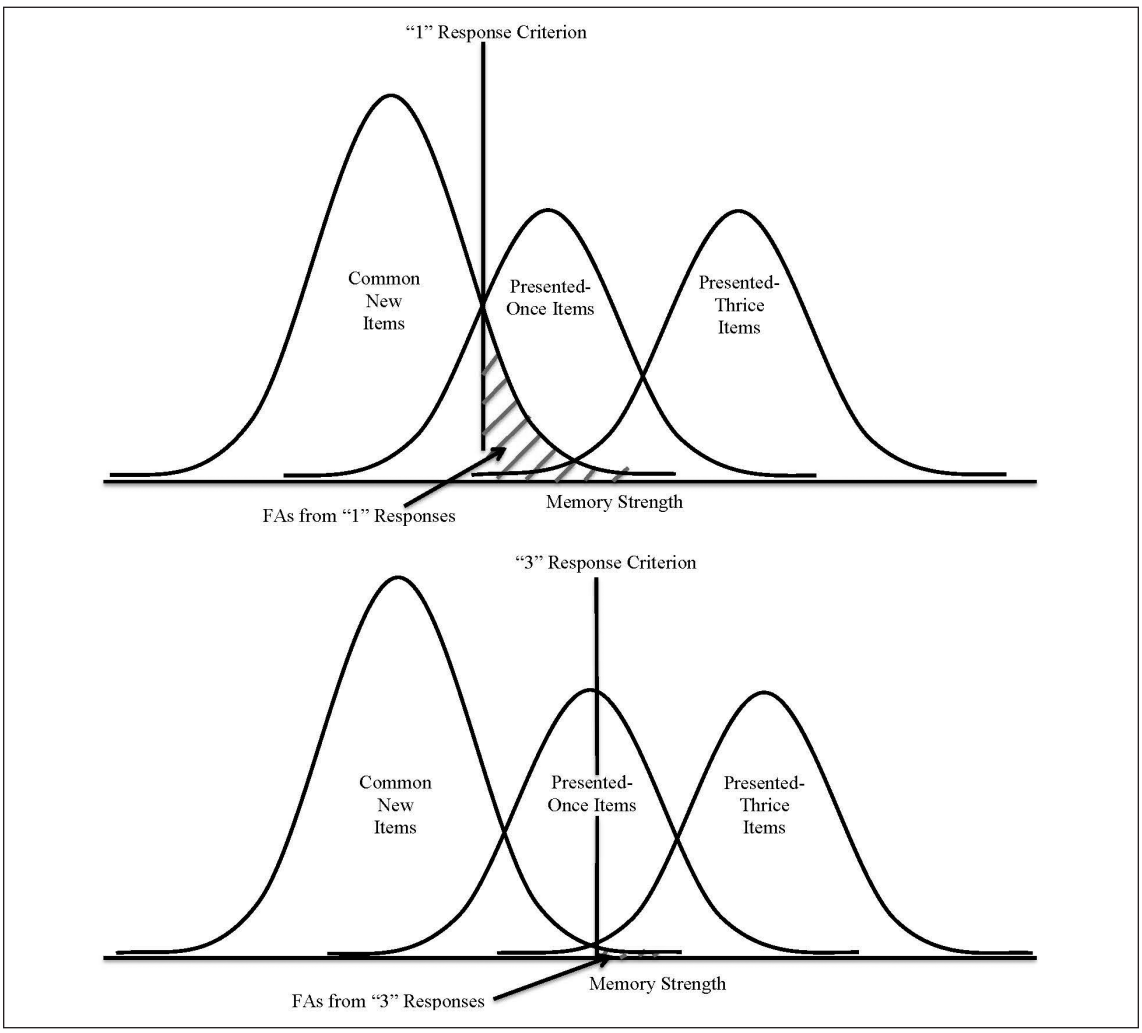

FIGURE 3. A common new (distractor) item set for both the presented-once and presented-thrice items and the decrease in false alarms as a result of a strength-based upward criterion shift

one can identify responses based on higher criteria versus lower criteria.

The sorting of the probes into the strong versus the weak classes (and the new item class) may be considered as using multiple fixed criteria instead of a single criterion that shifts from one trial to the next. However, we suggest that there may not be a simple, definitive way to define what constitutes using multiple fixed criteria versus what is using a single shifting criterion. Take Starns and Olchowski's (2015) experimental procedure as an example. The instruction told the participant explicitly that if the probe was green and displayed on the left side of the screen, it was either presented once or not presented; if the probe was red and displayed on the right side of the screen, it was either presented five times or not presented. Thus, in the cued criterion paradigm, the two criteria appear to be explicitly and directly given to the participants. It is not easy to decide whether this procedure makes participants adopt two set criteria or a single switching criterion.

In our view, the underlying mental processes in the cued and the uncued criterion paradigms are probably the same or at least similar. If there is a difference, the difference is that in the cued paradigm, participants are clearly told that the displayed probe should be judged by the strong or the weak criterion, whereas in the uncued paradigm, participants have to figure out by themselves which criterion to use for a test probe. Therefore, although both methods may

This content downloaded from
75.14.0.129 on Thu, 14 Jan 2021 19:51:39 UTC
All use subject to https://about.jstor.org/terms

possibly implement the same underlying mental process, the uncued method does away with the color, the spatial cues, and the two new item sets. It also requires participants to play a more self-initiated role in discriminating the strength classes of the probes. Also, the uncued paradigm may be intuitively more meaningful to the participants.

EXPERIMENT 1

We used a mixed-strength list paradigm in which some words of a list were studied three times and some words were studied only once. At test, participants were asked to indicate whether the test word was studied zero times (new word), one time, or three times by pressing the respective number key. The new item set was not divided into two differently cued subsets. No cues of any kind were used.

METHOD

Participants

One hundred nine University of Texas–Rio Grande Valley undergraduate psychology students, 74 women and 35 men, aged 18 years and above, participated in this experiment for course credit. These participants did not do any of the other experiments in this study.

Materials and Design

Ten lists of 28 unrelated words were compiled by selecting words from Kučera and Francis' (1967) word frequency norms in the range of 100–150 occurrences per million words. Care was taken to see that the words within a list were semantically unrelated and that no word was repeated across the lists. Seven words were randomly selected anew for each participant from the 28 words and presented 3 times. Another 7 words were randomly selected from the remaining 21 words and presented once. The remaining 14 words on the 28-word list were used as new (distractor) words in the test. Thus, the list of words used at study was composed of 14 words, seven presented 3 times and 7 presented 1 time, totaling 28 presentations.

Procedure

During the study phase, words were displayed in the center of the screen, one at a time, for 1.5 s per word, with an interword interval of 1 s, in a new random order for each participant with the constraint that the same word (in the case of the presented-thrice words) could not be presented consecutively. Participants were told that some words would be presented three times and some words would be presented once and that they had to pay attention to the number of times a word was presented because they would be tested for their memory of the number of times a word was presented. They were also told that no word was repeated across the lists, and therefore their judgment should be based on the most recent list only. After the study list, participants performed a task in which they counted backwards by 3 at a time for 20 s. In the backward counting, the program generated a three-digit random number and prompted participants to enter a number that was smaller than the generated number by 3. After typing in the less-3 number, they pressed "Enter" to receive the next prompt. Participants were urged to count at a reasonably fast speed. When they made an error in counting, they had to go back to the previous number to re-count.

At the end of the counting task, the test began. The 28 test words (14 targets and 14 distractors) were presented one at a time in a new random order for each participant. Participants were told that they should press the "0" key on the numeric keypad if they did not remember seeing the word during the study phase, the "1" key if they remembered seeing the word once, and the "3" key if they remembered seeing the word 3 times. When one of these three response keys was pressed, the test word was erased from the screen and the next test word appeared after a 1-s blank screen. Participants were told that if they pressed the keys at random the computer program would detect it and subsequently require them to redo the experiment. The purpose of the warning was to deter random responses. Nobody repeated the experiment. The same cycle of learning the 14 list words and testing with the 28 test words was repeated 10 times to complete the experiment.

RESULTS AND DISCUSSION

The proportions of items presented once and thrice that were correctly recognized as having been presented once and thrice were .723 and .785, respectively (we call these HRs the specific HRs). An analysis of variance (ANOVA) showed that the difference between these two specific HRs was significant, $F(1, 108) = 20.40$, $MSE = .010$, $p < .0001$, $\eta_p^2 = .159$. As expected, some presented-once items were incorrectly judged as presented-thrice items and some presented-thrice items were incorrectly judged as

This content downloaded from
75.14.0.129 on Thu, 14 Jan 2021 19:51:39 UTC
All use subject to https://about.jstor.org/terms

presented-once items. If both "1" and "3" responses were considered a hit (which we call the general HR) for each class of items, then the general HRs for test items categorized as presented-once and presented-thrice items were .811 and .970, respectively. This difference was also significant, $F(1, 108) = 297.98$, $MSE = .005$, $p < .0001$, $\eta_p^2 = .734$.

As noted earlier, the FAR, which is the crucial piece of information relevant to the question of whether participants made trial-by-trial criterion shifts, was calculated as the proportion of the whole new item set that was falsely recognized as either a presented-once or a presented-thrice word. We compared the two respective FARs associated with the "3" responses (*strong judgment*) and "1" responses (*weak judgment*). In this method, the total set of new items consists of three types of responses "3," "1," and "0" responses, that is, "strong" FAs, "weak" FAs, and correct rejections. We use the term *type I FAR* to refer to the FAR calculated based on this whole new item set. The proportion of the "3" (*strong*) FAs and that of the "1" (*weak*) FAs in this single new item set is taken to be the FAR for the "strong" and "weak"[5] new items, respectively. The mean "1" response FAR was .091, and that of the "3" response FAR was .006. The former was 15 times the latter, and the difference was significant, $F(1, 108) = 112.14$, $MSE = .014$, $p < .0001$, $\eta_p^2 = .509$. Again, the FAR associated with making a "3" response was significantly lower than the FAR associated with making a "1" response. This is a strength-based mirror effect. The fact that for the majority of the false acceptances, participants judged a new item to be a presented-once rather than presented-thrice word is clear evidence that they used a lower criterion for classifying a word as presented once than as presented thrice. This result was consistent with that in Hintzman's (2004a) study.

TABLE 1. Percentage Distribution of "0," "1," and "3" Responses as a Function of Presentation Frequencies, Experiment 1

	Response		
Row percentage Column percentage	"0"	"1"	"3"
Presentation frequencies			
0	90.27	9.13	0.60
	89.19	16.75	1.36
1	18.89	72.48	8.63
	9.29	66.21	9.72
3	3.06	18.53	78.41
	1.51	17.04	88.92

Based on the specific HR, the d' for the presented-thrice items was 3.15, and that for the present-once item was 2.16.[6] The difference was significant, $F(1, 108) = 228.29$, $MSE = .235$, $p < .0001$, $\eta_p^2 = .679$. The criterion c (a value of zero indicates a neutral criterion, a positive value indicates a conservatively biased criterion, and a negative value indicates a liberally biased criterion; Macmillan & Creelman, 2005) was .685 for the presented-thrice items and .427 for the presented-once items. The difference was significant, $F(1, 108) = 58.29$, $MSE = .062$, $p < .0001$, $\eta_p^2 = .351$.

In a second way of assessing these two FARs, we compared them directly with each other (type II FAR). This method gives a measure of what proportion of the total FAs was a "1" response and what proportion a "3" response. A chi-square analysis examined how the pooled FA set (which contained both the "1" response FAs and "3" response FAs) was split between these two types of FAs.[7] Within this pooled FA set, 93.81% was the "1" response FAs and only 6.19% was the "3" response FAs. The split was far from even, $\chi^2 (1, N = 742) = 570.39$, $p < .0001$.

We demonstrated that participants' classified responses (participants' judged presentation frequencies for an item) strongly influenced the FAR. But did their classified response reflect the actual presentation frequency of the items? To find that out, we looked at the association between the classified response and the actual presentation frequency. The association between these two variables was highly significant, $\chi^2 (4, N = 15{,}260) = 16{,}107$, $p < .0001$. The response percentage distribution as a function of actual presentation frequencies is presented in Table 1.

Thus, when the experimental task makes sense to the participants, they show evidence of criterion shifts. Therefore, in our view, looking at the FAR in the red-colored new item set and in the green-colored new item set, respectively, as in the cued criterion method, is looking for evidence in a place in which such evidence is at least difficult to find.

In sum, in the present task no strength cues of any kind were given to the participants. A memory task of sorting the test items into three classes based on the number of times an item appeared in the study list

This content downloaded from
75.14.0.129 on Thu, 14 Jan 2021 19:51:39 UTC
All use subject to https://about.jstor.org/terms

was readily comprehensible to the participants, and therefore both the HRs and the two ways of measuring FAR consistently obtained evidence of strength-based, trial-by-trial criterion shifts.

EXPERIMENT 2

We provide a brief report on Experiment 2 because it differed from Experiment 1 in a minor procedural difference. In Experiment 1, participants were asked to make a trinary response. However, in a typical recognition test, participants made a binary response. The main purpose of this experiment was to replicate the results of Experiment 1 by using the normal binary ("old/new") response mode followed by a second response that indicated whether the accepted word was, in their judgment, a presented-once or presented-thrice word. We assumed that participants could tell whether a particular "yes" response was based on strong evidence or weak evidence after they made the recognition decision. We predicted that when participants indicated that a "yes" response was based on the evidence of a presented-thrice item, the chance of making an FA would be lower than when it was based on a presented-once word.

METHOD

Participants

Ninety-four introductory psychology students (37 men, 57 women) participated in the experiment for course credit. Their mean age was 19.98 years, ranging from 18 to 31. The data from 6 participants were at chance-level accuracy and excluded, leaving 88 participants' data in the analysis.

Design, Materials, and Procedure

Participants responded to a test word by pressing either the "Yes" or the "No" response key. If the response was "No," the next test word appeared. If the response was "Yes," a follow-up question immediately appeared asking them whether they remembered studying the word once or three times. They responded by pressing either the "1" key or the "3" key.

RESULTS AND DISCUSSION

The .775 mean specific HR for the presented-thrice items was significantly greater than .674 for the presented-once items, $F(1, 87) = 52.53$, $MSE = .009$, $p < .0001$, $\eta_p^2 = .377$. The .950 mean general HR for the presented-thrice items was also significantly greater than the .765 for the presented-once items, $F(1, 87) = 321.46$, $MSE = .005$, $p < .0001$, $\eta_p^2 = .787$. The "strong" and "weak" type I FARs were .005 and .065, respectively. The latter was 13 times as great, $F(1, 87) = 97.54$, $MSE = .002$, $p < .0001$, $\eta_p^2 = .527$.

Based on the specific HR, the d' for the presented-thrice items was 3.15, and that for the presented-once item was 2.17. The difference was significant, $F(1, 87) = 221.25$, $MSE = .190$, $p < .0001$, $\eta_p^2 = .718$. The criterion c was .730 for the presented-thrice items and .597 for the presented-once items. The difference was significant, $F(1, 87) = 13.50$, $MSE = .057$, $p = .0004$, $\eta_p^2 = .134$.

Finally, we examined how the FA response pool was split between "1" and "3" after "yes/no" judgments (type II FAR). The percentage of "1" judgments was 92.34%, and that of "3" judgments was 7.66%. A chi-square analysis showed that they were not equal, $\chi^2 = 618.21$, $p < .0001$. We conducted another chi-square analysis using the judged frequency and actual presentation frequency as two variables. The distribution of judged item presentation frequencies is presented in Table 2 as a function of the actual presentation frequencies.

The result showed that there was a strong association between the judged frequency and the actual presentation frequency, $\chi^2 (4, \mathcal{N} = 26{,}640) = 25{,}446.86$, $p < .0001$.

TABLE 2. Percentage of 0 ("New"), Studied Once ("1"), and Studied Thrice ("3") Post-Yes/No Judgments as a Function of Presentation Frequency of the Test Words, Experiment 2

	Presentation frequency		
Row percentage Column percentage	0	1	3
Post-yes/no judgment			
"New"	86.70 93.00	10.95 23.49	2.35 5.03
"1"	13.22 6.46	68.95 67.42	17.83 17.44
"3"	1.22 0.54	10.37 9.09	88.41 77.53

This content downloaded from
75.14.0.129 on Thu, 14 Jan 2021 19:51:39 UTC
All use subject to https://about.jstor.org/terms

This is evidence of participants shifting between a laxer criterion for endorsing what they thought was a presented-once item and a more stringent one for endorsing what they thought was a presented-thrice item.

EXPERIMENT 3

Participants in this experiment studied the items in the same way as in Experiments 1 and 2. But at test, they made a "yes/no" binary response to each test item and immediately provided a confidence rating for that response. When people give a higher confidence rating to a decision or choice they make, it is analogous to declaring that the decision is based on more evidence, which is the same as adopting a stricter criterion. Conversely, a low confidence rating indicates the decision is based on less evidence, which is the same as adopting a more liberal criterion. Therefore, a confidence rating is taken as an indication of the decision criterion people adopt (Dunn, 2004; Macmillan & Creelman, 2005; McNicol, 1972; Selmeczy & Dobbins, 2013; Singer & Wixted, 2006; Stretch & Wixted, 1998b; Zawadzka et al., 2017).

One advantage of using confidence ratings is that they can be easily related to what people do in their daily lives and therefore make good sense to the participants. For example, when a professor gives a passing grade to a term paper (assuming she adopts a pass–fail grading system), the question of how confident she is that the paper deserves a pass is frequently asked and very intuitively understandable. This also applies to how the grader feels after giving any other specific grades. And it also reflects the criterion used in giving that grade. Thus, we predict that "yes" responses to the presented-thrice items will be associated with higher confidence than will "yes" responses to the presented-once items, and in addition, confidence ratings for FAs should be lower than for hits.

Analogously, a high confidence rating for a "no" response means that the response, in the participants' judgment, is based on a very low criterion (very little evidence of having studied the item) so that if an item does not even meet such a low criterion, it must not have been studied. Conversely, if a "no" response is given a low confidence rating, that means the response is probably based on a higher criterion so that the "no" response could be a miss response (the rejected item could have been studied). Our results support this analysis.

METHOD

Participants

One hundred nine introductory psychology students, 74 women and 35 men, aged 18 years and above, at University of Texas–Rio Grande Valley participated in the experiment to receive course credit. Two participants gave the highest confidence rating (5) for at least 95% of the trials, and another participant gave the lowest confidence rating (1) to all trials. These three participants' data were not included in the analysis, leaving 106 participants' data in the analysis.

Materials, Design, and Procedures

The materials and the study phase of this experiment were the same as in Experiment 1. The test was somewhat different from that in Experiment 1. For the test, participants were told that after each "yes/no" response, they should immediately provide a rating regarding how confident they were that their response was accurate, with 1 indicating *not confident at all,* 2 *moderately confident,* 3 *confident,* 4 *highly confident,* and 5 *absolutely positive.* Unlike in some studies in which the "sure new" was given the lowest numerical rating and "sure old" the highest rating, we used the same confidence instruction for "yes" and "no" responses alike. Thus, if participants were most confident that an item was surely new, they gave it a 5, just as if they were most confident that it was surely old, they gave it a 5. In our opinion, this rating mode is more in line with people's intuitive understanding that a higher number stands for a higher confidence in the accuracy of their response regardless of whether the response is "yes" or "no." The five pairs of the rating numbers and their verbal labels were displayed on five rows, with one pair per row near the middle of the screen immediately after the yes/no response was made. Participants typed in a number corresponding to their confidence and then pressed the "Enter" key. After a 1-s blank screen, the next test word appeared. Participants were asked to avoid giving a uniform confidence rating of 5 to all responses because, they were told, it was extremely unlikely for one to feel "absolutely positive" for all the responses one made.

RESULTS AND DISCUSSION

The HR for the weak items was .800, and that for the strong items was .947. The difference was signifi-

This content downloaded from
75.14.0.129 on Thu, 14 Jan 2021 19:51:39 UTC
All use subject to https://about.jstor.org/terms

cant, $F(1, 102) = 341.43$, $MSE = .003$, $p < .0001$, $\eta_p^2 = .770$. One of the crucial questions is whether "yes" responses to the presented-thrice items were associated with a higher mean confidence rating than "yes" responses to the presented-once items, and in turn whether "yes" responses to presented-once items were associated with a higher mean confidence rating than the "yes" responses to the new items. To answer these questions, we conducted an ANOVA on the mean confidence ratings among "yes" responses to presented-thrice (strong hits), presented-once (weak hits), and nonpresented test items (FAs), which were 4.82, 4.28, and 2.84, respectively. The differences were significant, $F(2, 202) = 51.79$, $MSE = 2.06$, $p < .0001$, $\eta_p^2 = .339$. A Newman–Keuls post hoc test showed that the three means were significantly different from one another. This supports the claim that there was a trial-to-trial shift in the response criterion because the confidence ratings varied from trial to trial.

We then examined how the confidence level of "yes" responses was related to FARs. We split the 5-point confidence scale into the high and low categories and used the single, whole new item set to evaluate the FARs. Ratings of 4 and 5 were combined to form the high confidence category, and ratings of 1, 2, and 3 combined to form the low confidence category. The mean type I FARs for low- and high-confidence "yes" responses were .151 and .068, respectively. This difference was significant, $F(1, 102) = 41.24$, $MSE = .009$, $p < .0001$, $\eta_p^2 = .288$.

We showed that higher confidence was associated with higher HR and lower FAR. But was there an association between the confidence rating and the strength of encoding? To answer this, we conducted a chi-square analysis with confidence level and presentation frequency of words (0, 1, 3) as two dimensions. The percentage distribution of confidence ratings as a function of presentation frequencies is presented in Table 3.

The association between the confidence rating and the encoding strength was significant, $\chi^2 (2, N = 14,237) = 2,177.28$, $p < .0001$. The response distribution showed that the majority of the FAs (64.77%) were made in the low confidence category, and the pattern for hits for the presented-once items (77.81% in the high confidence category) was reversed from that of the FAs. The response frequency distribution pattern for the presented-thrice items showed that an overwhelming majority of hits (90.11%) was made in the high confidence category.

Another chi-square test was conducted to determine specifically whether there was an association between the confidence level and the presentation frequencies of 1 and 3 (the data for the new items with frequency 0 were excluded), which was crucial for answering the question of whether participants used a stricter criterion for the strongly encoded than for the weakly encoded items. The response frequency distribution is presented in Table 4.

The association between confidence level and presentation frequency was still significant, $\chi^2 (1, N = 12,846) = 368.06$, $p < .0001$. The response distribution pattern showed that a higher criterion (as reflected in a higher confidence rating) was used more frequently for the presented-thrice than for the presented-once items. In sum, with confidence

TABLE 3. Percentage of "Yes" Responses as a Function of Confidence Rating and Presentation Frequency of the Test Words, Experiment 3

	Presentation frequency		
Row percentage Column percentage	0	1	3
Low confidence rating	31.12 64.77	45.08 22.19	23.80 9.89
High confidence rating	4.32 35.23	40.34 77.81	55.34 90.11

TABLE 4. Percentage of "Yes" Responses as a Function of Confidence Rating and Presentation Frequency of the Test Words (with New Words Excluded), Experiment 3

	Presentation frequency	
Row percentage Column percentage	1	3
Low confidence rating	65.45 22.19	34.55 9.89
High confidence rating	42.16 77.81	57.84 90.11

This content downloaded from
75.14.0.129 on Thu, 14 Jan 2021 19:51:39 UTC
All use subject to https://about.jstor.org/terms

rating as a measure of participants' adopted response criteria, the data suggest trial-by-trial criterion shifts as a function of memory strength.

To gather more evidence for trial-by-trial criterion shifts, we examined the data in the rejected ("no" response) category in the same way as we did the data in the accepted ("yes" response) category. When participants report high confidence in rejecting a test item, they are saying that they cannot find much evidence of studying the item to make it meet even a low criterion. On the other hand, when they reject a test item with a lower confidence, they are signifying that they use a higher criterion (and therefore the probability of making a miss error is higher). An ANOVA on the "no" response data with presentation frequency (0, 1, and 3) as a factor and confidence rating as a dependent measure showed that the mean confidence rating of 3.76 for the rejected new items (a correct rejection), 2.80 for the presented-once items (a miss), and 2.47 for the presented-thrice items (a miss), respectively, were significantly different, $F(2, 183) = 54.57$, $MSE = .791$, $p < .0001$, $\eta_p^2 = .374$. A Newman–Keuls test indicated that the three means were significantly different from one another. For the "no" responses, the new items were rejected at the lowest criterion but with the highest confidence, and the presented-thrice items were rejected at the highest criterion but with the lowest confidence, and the presented-once items were rejected at a criterion in between the two. This suggests that not all rejections, just as not all acceptances, were made based on a single criterion.

As we did for the "yes" response data, we conducted a chi-square analysis for the "no" response data. The "no" response data can provide evidence reciprocal to but consistent with the evidence provided by the "yes" response data. The percentage distribution of "no" response frequencies as a function of presentation frequency and confidence level is presented in Table 5.

The association between the two variables was significant, $\chi^2 (2, N = 15{,}154) = 418.84$, $p < .0001$. As shown in Table 5, more than half of the correct rejections (55.91%) were made in the high confidence category. In contrast, the majority of the misses was made in the low confidence category (68.10% for the presented-once and 74.02% for the presented-thrice items). When the new items were removed, the response distribution shift that occurred between the presented-once and presented-thrice items as a function of confidence level was still significant, $\chi^2 (1, N = 1{,}848) = 5.00$, $p = .026$. The response frequency distribution of the reduced data is presented in Table 6.

Thus, just as for the acceptance responses, rejections associated with items of differing memory strengths were also based on different criteria. The most important point that can be gleaned from the results of Experiment 3 is that participants were able to switch between different internal criteria for test items of different memory strengths on a trial-by-trial basis.

EXPERIMENT 4

We give a brief report on this experiment because the logic of this experiment is the same as for Experiment 3. In this experiment, after participants made a "yes"

TABLE 5. Percentage of "No" Responses as a Function of Confidence Rating and Presentation Frequency of the Test Words, Experiment 3

	Presentation frequency		
Row percentage Column percentage	0	1	3
Low confidence rating	82.08 44.09	13.98 68.10	3.95 74.02
High confidence rating	92.92 55.91	5.84 31.90	1.24 25.98

TABLE 6. Percentage of "No" Responses as a Function of Confidence Rating and Presentation Frequency of the Test Words (with New Words Excluded), Experiment 3

	Presentation frequency	
Row percentage Column percentage	1	3
Low confidence rating	77.99 68.10	22.01 74.02
High confidence rating	82.54 31.90	17.46 25.98

This content downloaded from
75.14.0.129 on Thu, 14 Jan 2021 19:51:39 UTC
All use subject to https://about.jstor.org/terms

response, they were asked to indicate the quality of their memory experience by responding with either a "remember" (indicating a recollective memory) or a "know" judgment (indicating a vaguely familiar memory) (Gardiner, 1988; Rajaram, 1993; Tulving, 1985). Some researchers suggest that "remember" and "know" judgments represent two recognition criteria, with "remember" as a stricter criterion than "know" (Donaldson, 1996; Dunn, 2004; Wixted & Stretch, 2004). We predicted that a "yes" response to a presented-thrice item would be more frequently based on a "remember" experience than one to a presented-once item and that when participants indicated that a "yes" response was a "remember" experience, the chance that the "yes" response was a FA would be lower than when they indicated that the response was a "know" experience (Dewhurst, Holmes, Brandt, & Dean, 2006; Reder, Angstadt, Cary, Erickson, & Ayers, 2002).

METHOD

Participants

Ninety-three introductory psychology students (36 men, and 57 women) participated in the experiment for course credit. Their mean age was 20.46 years, ranging from 18 to 43. The data of 5 participants showed chance-level accuracy and were excluded, leaving 88 participants' data in the analysis.

Design, Materials, and Procedure

After participants pressed the "Yes" key, they were asked to indicate whether their memory experience for the test word was a "remember" or a "know" experience by pressing either the "R" key or the "K" key. The instruction given to participants on what constituted a "remember" or the "know" experience was closely modeled after Rajaram's (1993) instructions.

RESULTS AND DISCUSSION

The hit rate was .961 for the presented-thrice items and .798 for the presented-once items. The difference was significant, $F(1, 87) = 243.65$, $MSE = .005$, $p < .0001$, $\eta_p^2 = .737$. The proportion of new items falsely recognized as a "remember" old item was .019, and the corresponding proportion falsely recognized as a "know" old item was .051. The difference was significant, $F(1, 87) = 35.93$, $MSE = .001$, $p < .0001$, $\eta_p^2 = .293$. Within the FA response category, the "know" FAs made up 72.35% and the "remember" 27.65%. A chi-square analysis for equal split of FAs showed that they were not equal, χ^2 (1, N =868) = 173.44, $p < .0001$. Another chi-square analysis showed that the association between the judgment and the presentation frequency was highly significant, χ^2 (4, N = 24,640) = 17,662.49, $p < .0001$.

The percentages of "new," "know," and "remember" responses as a function of the presentation frequency are presented in Table 7.

GENERAL DISCUSSION

In recognition memory tests, people seem disinclined to change their recognition criterion for items encoded strongly and weakly within the same list and tested within the same test even when many cues are provided to facilitate the discrimination between the two classes of items (Higham et al., 2009; Morrell et al., 2002, Experiment 1; Singer, 2009, under a rote learning condition; Singer & Wixted, 2006, Experiments 1 and 2; Stretch & Wixted, 1998a, Experiments 2–5; Verde & Rotello, 2007, Experiments 1–4; Starns et al., 2006; for reviews, see Bruno et al., 2009; Starns & Olchowski, 2015). Different hypotheses were proposed to explain this lack of a consistent, strength-based, trial-by-trial criterion shift. In the following, we summarize the new conceptual and methodological contributions made by the present study toward

TABLE 7. Percentage of "New," "Know," and "Remember" Judgments as a Function of Presentation Frequency of the Test Words, Experiment 4

	Presentation frequency		
Row percentage Column percentage	0	1	3
Post-yes/no judgment			
"New"	88.54 92.95	9.60 20.16	1.86 3.90
"Know"	25.54 5.10	50.31 20.08	24.16 9.64
"Remember"	2.60 1.95	39.81 59.76	57.60 86.46

This content downloaded from
75.14.0.129 on Thu, 14 Jan 2021 19:51:39 UTC
All use subject to https://about.jstor.org/terms

furthering the understanding of strength-based, trial-by-trial criterion shifts.

The central point of the present study is to bring to the fore the fact that strength-based, trial-by-trial criterion shifts are a commonly occurring phenomenon, both in our everyday life and in lab settings, and that the reason why in Stretch and Wixted's (1998a) cued criterion paradigm it happens only under such extraordinarily elaborate cuing procedures as in Starns and Olchowski (2015) is the abstract nature of the cued criterion test and the way FAs are measured in that paradigm. We showed that strength-based, trial-by-trial criterion shifts can actually occur naturally in such tasks as frequency judgments and source monitoring (Experiments 1 and 2; Hintzman, 2004a, 2004b, 2005 for frequency judgments; and Hicks & Starns, 2006a, 2006b for source monitoring) without the aid of any cuing. Also, we demonstrated that the commonly used confidence rating and remember/know judgments can reflect strength-based, trial-by-trial criterion shifts, although to our knowledge nobody has brought these common methods to bear on resolving this controversial subject. Instead, people working on this issue are trying to achieve the goal of attaining trial-by-trial criterion shifts by developing more elaborate cuing systems. We used the commonly used confidence rating and memory quality judgments to show that each of these simple methods actually reflects people's capability of making strength-based, trial-by-trial criterion shifts. However, researchers doing memory quality rating studies have not applied these methods to addressing the strength-based, trial-by-trial criterion shift issue. To our knowledge, we are the first to expressly bring these commonly used memory judgment methods to bear on the specific issue of whether people can make strength-based, trial-by-trial criterion shifts.

A methodological point we made is that dividing the new items arbitrarily into two subsets, one marked as "strong" and one as "weak" with physical features may be a major source of the difficulty in the cued criterion paradigm. In that paradigm, the ability to make strength-based, trial-by-trial criterion shifts becomes the ability to comprehend the experimenter-intended meaning of these cuing labels and act on them accordingly, which in our opinion is a different thing than making a trial-by-trial criterion shift. Instead of "forcing" participants to recognize that one distractor set is strong and the other is weak by using different color cues and spatial location cues, we had participants sort the probes or give a post-recognition rating and directly calculated the FAR by using a single distractor set. We removed the color and location cues that are the standard practices in many studies on this subject matter (e.g., Starns & Olchowski, 2015). Thus, in the uncued method, there are no "strong" and "weak" new item sets. We believe that the lack of trial-by-trial criterion shifts often found in the typically used cued criterion paradigm results mainly from the difficulty of the method rather than the difficulty inherent in the mental process of making a strength-based, trial-by-trial criterion shift.

Finally, as discussed earlier, our findings may be construed as evidence of using multiple fixed criteria rather than of using a single criterion that shifts from trial to trial. There may be no substantive or definitive way to distinguish between these two underlying processes. We suggest that the criterion shifting in the cued criterion paradigm may be conceptualized as a form of switching between two fixed criteria. After participants have encountered several instances of strong and weak items in the test, they may be able to deduce two fixed criteria, one for items in the green (weak criterion) and one for items in red (strong criterion), and simply maintain the two fixed criteria and use the one or the other for a recognition decision. Thus, it seems to us that the distinction is fuzzy between the concept of shifting a single criterion and that of using multiple criteria. We consider it important to recognize the need for a clearer theoretical and empirical definition of trial-by-trial criterion shifts. At any rate, however, the results we obtained can be safely regarded as evidence of observers' capability of using multiple recognition criteria from one trial to another, if not that of switching a criterion.

We end the discussion with these statements: Just because no difference in FAR is found between two arbitrarily defined subsets of new items does not necessarily mean that participants do not shift their decision criteria according to the memory strength of items on a case-by-case basis in a recognition tests. Most likely, in the cued criterion paradigm, participants may not have understood the purpose of the distinguishing cues as intended by the experimenter. We provided evidence for strength-based, mixed-list, trial-by-trial criterion shifts using no cuing

This content downloaded from
75.14.0.129 on Thu, 14 Jan 2021 19:51:39 UTC
All use subject to https://about.jstor.org/terms

techniques at all. It is possible that in the previous studies that failed to find evidence of strength-based criterion shifts, the criterion shifts might have occurred, but they might not have been captured in the measures researchers used to index them. Therefore, we conclude that it is not the trial-by-trial criterion shifts that are hard as a mental process; it is the typical cued criterion paradigm that makes it hard to observe.

NOTES

We thank Jim Neely sincerely for the very long discussions with him, his countless feedbacks, and suggestions on earlier versions of this article. We also thank Richard Zuniga, Martin Perez, and Paola Salazar for helping with data collection, the three anonymous reviewers for their critiques, and Robert Proctor for his detailed editorial feedback from which this article benefited greatly.

Address correspondence about this article to Jerwen Jou, Department of Psychological Science, University of Texas–Rio Grande Valley, 1201 West University Drive, Edinburg, TX 78539-2999 (e-mail: jerwen.jou@utrgv.edu).

1. An alternative interpretation of the mirror effect is the differentiation theory (Criss, 2009; Koop & Criss, 2016; Shiffrin et al., 1990; for a summary see Criss & Koop, 2015). According to that theory, the strength-based mirror effect is caused not by a criterion shift but by an amplified differentiation between the strong and the weak items caused by enhanced encoding, relative to weak encoding. According to that idea, the more one learns about a stimulus, the more different it becomes in perception and memory from the unstudied ones.

2. As long as the strong and weak items are mixed in the same test, participants tend not to make strength-based, trial-by-trial criterion shifts regardless of whether the two classes of items are learned in a pure or mixed list condition (Singer, 2009).

3. Hintzman (2004a) investigated whether the duration of study time or the frequency of presentations was the primary determinant of the frequency judgment and recognition confidence (or whether there was one common determinant or multiple determinants for the two judgments). Different FA responses ("1," "2," and "3") for items not presented at study were listed in a table in the appendix, which showed that "3" FA response was the lowest. Hintzman's other studies (2004b, 2005) are also relevant, although he did not present FARs.

4. The source monitoring approach of Hicks and Starns (2006a, 2006b) is applicable to the present topic although they did not discuss the issue of strength-based, trial-by-trial criterion shifts in their articles.

5. We put quotation marks around the words *strong* and *weak* when they refer to a new item because the word does not literally mean *strong* and *weak* because the new item had never been studied.

6. When the HR was 1.0, we changed it to .99; when the FAR was 0, we changed it to .01, following the protocol of Hirshman (1995), Jou et al. (2004), and Kilic and Oztekin (2014).

7. In our view, there is no compelling reason to assume a correlation between one response and another in this experiment; therefore, we performed a chi-square test to assess the association between the frequency judgment and the word presentation frequency. Treating these responses as independent makes the test more conservative (Kachigan, 1986).

REFERENCES

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

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17,* 391–416.

Cheng, P. W., Holyoak, K. J., Nisbett, R. E., & Oliver, L. M. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology, 18,* 293–328.

Cho, K. W., & Neely, J. H. (2013). Null category-length and target–lure relatedness effects in episodic recognition: A constraint on item–noise interference models. *Quarterly Journal of Experimental Psychology, 66,* 1331–1355.

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

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

Dewhurst, S. A., Holmes, S. T., Brandt, K. R., & Dean, G. M. (2006). Measuring the speed of the conscious components of recognition memory: Remembering is faster than knowing. *Consciousness & Cognition, 15,* 147–162.

Dobbins, I. G., & Kroll, N. E. A. (2005). Distinctiveness and the recognition mirror effect: Evidence for an item-based criterion placement heuristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1186–1198.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition, 24,* 523–533.

Dunn, J. C. (2004). Remember-know: A matter of confidence. *Psychological Review, 111,* 524–542.

Gardiner, J. M. (1988). Functional aspects of recollective experiences. *Memory & Cognition, 16,* 309–313.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67.

This content downloaded from
75.14.0.129 on Thu, 14 Jan 2021 19:51:39 UTC
All use subject to https://about.jstor.org/terms

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

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

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic materials effects in Wason's selection task. *British Journal of Psychology, 73,* 407–420.

Healy, A. F., & Jones, C. (1973). Criterion shifts in recall. *Psychological Bulletin, 79,* 335–340.

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

Hicks, J. L., & Marsh, R. L. (1998). A decrement-to-familiarity interpretation of the revelation effect from forced-choice tests of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1105–1120.

Hicks, J. L., & Starns, J. J. (2006a). Remembering source evidence from associatively related items: Explanations from a global matching model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 1164–1173.

Hicks, J. L., & Starns, J. J. (2006b). The role of associative strength and source memorability in contextualization of false memory. *Journal of Memory and Language, 54,* 39–53.

Hicks, J. L., & Starns, J. J. (2014). Strength cues and blocking at test promote reliable within-list criterion shifts in recognition memory. *Memory & Cognition, 42,* 742–754.

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

Hintzman, D. L. (2004a). Judgment of frequency versus recognition confidence: Repetition and recursive reminding. *Memory & Cognition, 32,* 336–350.

Hintzman, D. L. (2004b). Time versus items in judgment of recency. *Memory & Cognition, 32,* 1298–1304.

Hintzman, D. L. (2005). Memory strength and recency judgment. *Psychonomic Bulletin & Review, 12,* 858–864.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list length effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 302–313.

Hockley, W. E., Hemsworth, D. H., & Consoli, A. (1999). Shades of the mirror effect: Recognition of faces with and without sunglasses. *Memory & Cognition, 27,* 128–138.

Hockley, W. E., & Niewiadomski, M. W. (2001). Interrupting recognition memory: Tests of a criterion-change account of the revelation effect. *Memory & Cognition, 29,* 1176–1184.

Johnson-Laird, P. N., Legrenzi, P., & Legrenzi, M. (1972). Reasoning and a sense of reality. *British Journal of Psychology, 63,* 395–400.

Jou, J., Escamilla, E. E., Arredondo, M. L., Pena, L., Zuniga, R., Perez, M. Jr., & Garcia, C. (2018). The role of decision criterion in the DRM false recognition memory: False memory falls and rises as a function of restriction on criterion setting. *Quarterly Journal of Experimental Psychology, 71,* 499–521.

Jou, J., Flores, S., Cortes, H. M., & Leka, B. G. (2016). The effects of weak versus strong relational judgments on response bias in two-alternative-forced-choice recognition: Is the test criterion-free? *Acta Psychologica, 167,* 30–44.

Jou, J., Matus, Y. E., Aldridge, J. W., Rogers, D. M., & Zimmerman, R. L. (2004). How similar is false recognition to veridical recognition objectively and subjectively? *Memory & Cognition, 32,* 824–840.

Jou, J., Shanteau, J., & Harris, R. J. (1996). An information processing view of framing effects: The role of causal-schemas in decision making. *Memory & Cognition, 24,* 1–15.

Kachigan, S. K. (1986). *Statistical analysis: An interdisciplinary introduction to univariate and multivariate methods.* Radius Press.

Kent, C., Lamberts, K., & Patton, R. (2018). Cue quality and criterion setting in recognition memory. *Memory & Cognition, 46,* 757–769.

Kilic, A., & Oztekin, I. (2014). Retrieval dynamics of strength based mirror effect in recognition memory. *Journal of Memory and Language, 76,* 158–173.

Kintsch, W., Welsh, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language, 29,* 133–159.

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

Koop, G. J., Criss, A. H., & Pardini, A. M. (2019). A strength-based mirror effect even when criterion shifts are unlikely. *Memory & Cognition, 47,* 842–854.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in strategic regulation of memory accuracy. *Psychological Review, 103,* 490–517.

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Erlbaum.

McNicol, D. (1972). *A primer of signal detection theory*. George Allen & Unwin.

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

This content downloaded from
75.14.0.129 on Thu, 14 Jan 2021 19:51:39 UTC
All use subject to https://about.jstor.org/terms

North, L. L., Olfman, D., Caldera, D. R., Munoz, E., & Light, L. L. (2018). Age, criterion flexibility, and item recognition. *Aging, Neuropsychology, and Cognition, 25,* 390–405.

Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory & Cognition, 21,* 89–102.

Reder, L. M., Angstadt, P., Cary, M., Erickson, M. A., & Ayers, M. S. (2002). A reexamination of stimulus-frequency effects in recognition: Two mirrors for low- and high-frequency pseudowords. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 138–152.

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.

Selmeczy, D., & Dobbins, I. G. (2013). Relating the content and confidence of recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 66–85.

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

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

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

Starns, J. J., Hicks, J. L., & Marsh, R. L. (2006). Repetition effects in associative false recognition: Theme-based criterion shifts are the exception, not the rule. *Memory, 14,* 743–761.

Starns, J. J., & Olchowski, J. E. (2015). Shifting the criterion is not the difficult part of the trial-by-trial criterion shifts in recognition memory. *Memory & Cognition, 43,* 49–59.

Stretch, V., & Wixted, J. T. (1998a). 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.

Stretch, V., & Wixted, J. T. (1998b). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1397–1410.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychologist, 26,* 1–12.

Tversky, A., & Kahneman, D. (1981, January 30). The framing of decision and the psychology of choice. *Science, 211,* 453–458.

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

Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology* (Vol. 1). Penguin.

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

Zawadzka, K., Higham, P. A., & Hanczakowski, M. (2017). Confidence in forced-choice recognition: What underlies the ratings? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43,* 552–564.

This content downloaded from
75.14.0.129 on Thu, 14 Jan 2021 19:51:39 UTC
All use subject to https://about.jstor.org/terms