

# Who Gives a Criterion Shift? A Uniquely Individualistic Cognitive Trait

Evan Layher  
University of California, Santa Barbara

Anjali Dixit  
University of California, Irvine

Michael B. Miller  
University of California, Santa Barbara




Individuals should *strategically* shift decision criteria when there are disproportionate likelihoods or consequences for falsely identifying versus missing target items. Despite being *explicitly* aware of the advantages for criterion shifting, people *on average* do not shift extremely, leading many theories to conclude that people are generally suboptimal at placing decision criteria. However, assessments of individual differences reveal that some people actually do criterion shift quite well while others fail to shift entirely. These individual differences may carry meaningful information about the nature and consistency of a person's decision-making strategies, but no studies have systematically assessed the stability of strategic criterion shifting within individuals over time. We assessed criterion shifting stability by administering test–retest recognition memory and visual detection tests where we induced decision biases through instruction, payoff, and base rate manipulations. Criterion shifting tendencies proved to be stable within and across decision domains regardless of the inducement. Individual differences in criterion shifting could not be explained by personality characteristics, metacognitive sensitivity, motivation, or performance on other cognitive tasks. Reports of confidence ratings, which are used to assess various criterion placements, showed no relationship to the extent of criterion shifting unless participants received instructions to make certain response types with high confidence only. Participants who inadequately shifted criteria still tended to set extreme criteria for reporting high confidence, suggesting that these individuals are *capable* of shifting to greater extents, but appear *unwilling* to do so. These findings demonstrate that strategic criterion shifting tendencies are a stable and uniquely individualistic cognitive trait.

**Keywords:** confidence ratings, criterion shifting, individual differences, recognition memory, test–retest reliability

**Supplemental materials:** <http://dx.doi.org/10.1037/xlm0000951.supp>

*Strategic* criterion shifting occurs when a person *knowingly* alters a decision strategy when the known prevalence of a target item changes or when the relative rewards or consequences of different response types change. Shifting decision criteria can improve decisional outcomes, particularly when there is uncertainty in the detected signal. A common example of this is when you see a person who looks familiar (the signal), but are unsure

whether you know them (a target) or not (a nontarget). The ideal goal is to greet a known acquaintance (a hit) and ignore a stranger (a correct rejection), but the uncertainty in your memory prevents you from knowing the correct course of action. Fortunately, there usually is other information at your disposal that can help minimize the chances of either potentially greeting a stranger (a false alarm) or failing to greet a known acquaintance (a miss). For

 Evan Layher, Department of Psychological and Brain Sciences, University of California, Santa Barbara;  Anjali Dixit, Department of Pharmaceutical Sciences, University of California, Irvine;  Michael B. Miller, Department of Psychological and Brain Sciences, University of California, Santa Barbara.

Datasets for this study can be accessed through the Open Science Framework: <https://osf.io/4k2hb/> (Layher & Miller, 2019). OSF Pre-registration links: Experiment 3, <https://osf.io/jkfp6/>; Experiment 4(1), <https://osf.io/4wnjm/>; Experiment 4(2), <https://osf.io/ae2rp/>; Experiment 5, <https://osf.io/tqc42/>.

All authors contributed to the design of the studies. Evan Layher performed data analyses and wrote the manuscript. Anjali Dixit helped collect data and edited the manuscript. Michael B. Miller edited the manuscript and oversaw all research activities.

The Institute of Collaborative Biotechnologies supported this research through Grant W911NF-19-0026 from the U.S. Army Research Office. The authors thank Justin Kantner and Tyler Santander for providing helpful feedback on the manuscript. The authors also thank Tyler Santander and Allison Shapiro for guidance with the statistics and figures as well as undergraduate research assistants who assisted with data collection: Tanya Bhatia, Ziyuan Chen, Shelsea De La O Sanchez, Jason Dong, Nickita Gupta, Luke Hamilton, Matejas Mackin, Saleem Omary, Kaylie Raymer, Nohemi Reyes Meraz, Isaiah Rodriguez-Anguiano, Hana Simon, Omeed Soltanalipour, Thien Truong, Amy Tsang, and Nicole Zimmerman.

Correspondence concerning this article should be addressed to Evan Layher, Department of Psychological and Brain Sciences, University of California, Santa Barbara, Building 251, Santa Barbara, CA 93106. E-mail: [layher@psych.ucsb.edu](mailto:layher@psych.ucsb.edu)

instance, if you believe the person is a coworker and you are in the workplace, then you should establish a *liberal* criterion by greeting the person even when your memory is vague, because the chances of such an encounter are high (a strategy to avoid *misses*). However, if you are on vacation in Tahiti you should establish a conservative criterion by only greeting a potential coworker when your memory is strong because this encounter is much less likely to occur (a strategy to avoid *false alarms*). Remarkably, in situations where criterion shifting is *clearly* advantageous, some individuals will readily shift decision criteria while others fail to shift entirely, which can detrimentally impact decisional outcomes (Aminoff et al., 2012, 2015; Frithsen, Kantner, Lopez, & Miller, 2018; Kantner, Vettel, & Miller, 2015; Layher, Santander, Volz, & Miller, 2018; Miller & Kantner, 2020). Extreme variability in criterion shifting across participants is well-documented, but currently no studies have systematically characterized the stability of criterion shifting tendencies within individuals over time. Yet, stable differences in criterion shifting tendencies across individuals may represent a fundamental aspect of those individual's decision-making strategies and carry theoretical implications for signal detection models of recognition memory.

Although the within-subject stability of *criterion shifting* tendencies is poorly understood, test-retest recognition memory studies suggest that *criterion placement* tendencies are stable over time (Kantner & Lindsay, 2012, 2014). In recognition memory, criterion placement is the threshold of familiarity strength that must be exceeded to recognize items. Criterion placement, like criterion shifting, is quite variable across individuals (Aminoff et al., 2012, 2015; Frithsen et al., 2018; Kantner et al., 2015; Kantner & Lindsay, 2012, 2014; Layher et al., 2018; Miller & Kantner, 2020). Despite large between-subjects variability, Kantner and Lindsay (2012, 2014) propose that the within-subject consistency of criterion placement over time makes it a stable cognitive trait. Some individuals regularly recognize stimuli based on weak familiarity evidence while others routinely require strong memory evidence before recognizing items. Criterion shifting, on the other hand, is a *shift* in the placement of a decision threshold to require more or less evidence before identifying a target when the circumstances surrounding a decision change (e.g., when recognizing a coworker in the workplace vs. a foreign vacation spot). The consistency and extent to which a person shifts a criterion is likely unrelated to an individual's criterion placement tendencies (though empirical reports of this relationship are lacking) because placing and shifting a criterion are separate behaviors. For example, two individuals might be quite adept at regularly establishing a *neutral* criterion by missing and falsely identifying items at equal rates. However, one individual might adaptively shift between conservative and liberal criteria when the situation calls for it, while the other may continuously maintain a neutral criterion even when criterion shifting becomes advantageous.

Criterion shifting can be categorized into two putative classes where an individual may either (a) knowingly shift a decision criterion based on known changes in the circumstances surrounding a decision (Aminoff et al., 2012; Egan, 1958; Banks, 1970; Healy & Kubovy, 1978; Rotello, Macmillan, Reeder, & Wong, 2005) or (b) unknowingly shift a decision criterion, such as through reinforcement learning<sup>1</sup> (Han & Dobbins, 2008, 2009; Wixted & Gaitan, 2002). An example of criterion shifting through reinforcement learning comes from Han and Dobbins (2008), who

covertly altered feedback conditions by rewarding one error type (either false alarms or misses) but not the other. Over time, participants *unknowingly* shifted toward a more liberal criterion when feedback encouraged false alarms and established a more conservative criterion when misses resulted in positive feedback. Strategic criterion shifting, on the other hand, occurs immediately and does not require feedback, but participants must be explicitly aware of the advantages for shifting criteria. For example, making participants aware that target items are more likely to appear than nontarget items during a test block (known as a base rate manipulation) will immediately cause many participants to shift to a liberal criterion, whereas participants who are unaware of such information will tend not to shift (Rhodes & Jacoby, 2007). We specifically investigated the stability of *strategic* criterion shifting tendencies where participants always received explicit instructions that informed them of the advantages for switching between conservative and liberal criteria.

Our analyses of criterion shifting focus on individual differences, which can reveal aspects of data that may contradict previous hypotheses that draw conclusions from group averages (Miller & Kantner, 2020). For example, a longstanding observation of group-averaged data shows that people are generally sub-optimal<sup>2</sup> at placing a criterion (i.e., people do not shift criteria extreme enough given the circumstances), leading to several hypotheses that attempt to explain this phenomenon (Benjamin, Diaz, & Wee, 2009; Hirshman, 1995; Kubovy, 1977; Lynn & Barrett, 2014; Maddox & Bohil, 2005; Parks, 1966; Thomas & Legge, 1970; Uehla, 1966). One hypothesis advocates that participants will probability match during test blocks that include a base rate manipulation (Parks, 1966; Thomas & Legge, 1970). That is, if 70% of items are targets, participants will respond "target" 70% of the time, even though the best strategy for maximizing accuracy is to always respond "target" unless there is strong evidence that an item is a nontarget. Aminoff and colleagues (2012) employed a base rate manipulation during recognition memory tests where participants received explicit instructions informing them that target (previously studied) items would appear either 70% (liberal condition) or 30% (conservative condition) of the time. Group-averaged results from this study suggested that probability matching is indeed a plausible explanation. However, when examining the data at an individual level, this hypothesis seems less plausible because some individuals actually do shift criteria quite well (i.e., almost always respond "target" in the liberal condition and almost

<sup>1</sup> There are many ways a criterion could be influenced unknowingly, which may include semantic similarity, sequential effects of test items, multidimensional representations, emotion-laden stimuli, word frequency, etc.

<sup>2</sup> We use the term "optimal" performance to describe a criterion that maximizes payoffs or the proportion correct at any level of discriminability, given the assumptions of signal detection theory (see Macmillan & Creelman, 2005). Although the classification of an "optimal" criterion will vary depending on the theoretical model (see Lynn & Barrett, 2014), our underlying claim that strategic criterion shifting is a stable cognitive trait is unrelated to whether certain individuals actually implement a model's definition of "optimal" criteria. We implement this simple device to demonstrate the inherent disadvantages of not shifting a criterion in response to changes in payoffs or base rates (i.e. when discriminability is held constant, individuals who appropriately shift criteria will achieve better outcomes, in regards to the intended goals of the task, than those who do not shift criteria).

never respond “target” in the conservative condition), whereas others fail to shift entirely (i.e., respond “target” at equal rates across both criterion conditions). We believe it is necessary to take into account these individual differences to gain a full understanding of the nature of criterion shifting tendencies.

Previous studies demonstrate that individual tendencies in strategic criterion shifting are largely consistent across different task types (Aminoff et al., 2012; Frithsen et al., 2018; Kantner et al., 2015). Aminoff and colleagues (2012) found a strong relationship in the degree to which individuals shift criteria during recognition memory tests for word versus face stimuli,  $r(93) = .58$ , which greatly exceeded the relationships in discriminability between the two tasks,  $r(93) = .07$  in the conservative criterion condition and  $r(93) = .22$  in the liberal criterion condition. Criterion shifting tendencies also appear stable across bias manipulations, regardless of whether a base rate manipulation is employed or individuals are incentivized to shift via a payoff manipulation (e.g., participants earn money for correct responses, but lose money for either false alarms or misses; Frithsen et al., 2018; Kantner et al., 2015). Frithsen and colleagues (2018) additionally found that criterion shifting tendencies are generally consistent across decision domains regardless of whether participants make recognition memory, visual detection, or visual discrimination judgments. However, the strength in the relationship of criterion shifting tendencies across tasks is sometimes mixed. For instance, Frithsen and colleagues (2018) found strong correlations between the extent of criterion shifting during a recognition memory test for *words* versus a visual detection test for identifying the presence of a white blob on a noisy background,  $r(47) = .53$ , and a visual discrimination test for determining the orientation of a Gabor patch on a noisy background,  $r(49) = .64$ . However, a weak relationship occurred between a recognition memory test for *faces* and a visual detection test for spotting white blobs on noisy backgrounds,  $r(49) = .17$ . Franks and Hicks (2016) found no significant relationship in the extent of criterion shifting between recognition memory tests that employed a base rate manipulation versus a manipulation that varied the known memory strength of studied items. However, manipulating the memory strength of items produces the strength-based mirror effect where an increase in discriminability results in both an increased hit rate and decreased false alarm rate (Hirshman, 1995; Starns & Olchowski, 2015). The underlying cause of the strength-based mirror effect is strongly debated where some argue that the familiarity strength of novel items remains constant and an increase in discriminability results in a criterion shift toward being more conservative (i.e., require stronger memory evidence), which decreases the false alarm rate (Starns, White, & Ratcliff, 2010; Stretch & Wixted, 1998b). Others argue that the familiarity strength of novel items *can* change and observed decreases in false alarm rates when discriminability increases can be attributed to memory processes instead of strategic criterion shifts (Criss, 2006; Shiffrin & Steyvers, 1997). Therefore, it is inconclusive as to whether using the strength-based mirror effect as a criterion manipulation is even valid because observed changes in false alarm rates may not actually be a result of strategic criterion shifting. Nevertheless, it helps raise the question as to why there are occasional inconsistent findings in the cross-task stability of criterion shifting; is the stability of criterion shifting a task specific phenomenon (Franks & Hicks, 2016), or are there alternative explanations for these mixed results?

One potential explanation for the observed inconsistencies in the cross-task stability of criterion shifting might be attributed to differences in demand characteristics between the two tasks (Kantner et al., 2015). Kantner and colleagues (2015) found that completely removing the study phase from recognition memory tests (i.e., participants received instructions that the study phase “malfunctioned” and thus could not encode any of the study images, but were still asked to perform the test phase anyways with an instruction induced criterion manipulation) dramatically affected the extent of criterion shifting (as it should), but only for a subset of participants. Because participants could not reliably use memory evidence to inform their decisions during the test phase (because they did not actually study any images), the best strategy is to maximally shift criteria by always responding “old” or always responding “new” depending on which response is more advantageous. However, many participants seemed unwilling to adopt this extreme strategy when task instructions required making a memory-based decision. Some individuals may have attempted to use other irrelevant perceptual cues to make recognition judgments or may have felt compelled to vary response types due to demand characteristics. This suggests that systematic design differences across tasks could alter response strategies and differentially affect how people integrate decision evidence with a criterion. For instance, Franks and Hicks (2016) observed no relationship in the extent of criterion shifting when comparing across recognition memory tests that implemented a base rate manipulation with a *blocked* design versus a strength-based manipulation with an *unblocked* design. Even if altering the memory strength of items is considered a valid criterion manipulation, a blocked design allows participants to shift and *maintain* a decision criterion throughout a test block whereas an unblocked design requires individuals to shift criteria on a trial-by-trial basis. Some individuals may be less willing to change decision strategies on a trial-by-trial basis compared with changing strategies once per test block (see Stretch & Wixted, 1998b). These differing task designs may have disrupted the stability of criterion shifting between the two bias manipulations that may otherwise be quite strong if both tasks incorporated the same design structure. Task design inequalities that could also differentially affect individual criterion shifting tendencies may result from other disparate design features such as differences in stimulus complexity or presentation times. For example, Frithsen and colleagues (2018) observed the weakest correlation in the extent of criterion shifting when comparing between a recognition memory test for faces and a visual detection test for the presence of a white blob on a noisy background. The critical difference in the design of the two tasks is that the face stimuli appeared for 1,500 ms whereas the white blob stimuli appeared for less than 350 ms. If task design differences are the culprit for this weak relationship, then homogenizing the presentation times across tests should improve the stability of criterion shifting between the two decision domains. For example, Aminoff and colleagues (2012) implemented identical recognition memory task designs that either used word or face stimuli and found a strong relationship in the extent of criterion shifting across the two tasks,  $r(93) = .58$ . The effect of differing task designs on criterion shifting tendencies is poorly understood but should be considered when assessing criterion shifting stability across tasks. Nevertheless, strong relationships in criterion shifting should be observed across all test types



and decision domains when demand characteristics are equivalent, if criterion shifting tendencies are truly a stable cognitive trait.

To provide evidence that strategic criterion shifting tendencies are a stable cognitive trait, we sought to empirically demonstrate that individual criterion shifting tendencies (a) are stable over multiple testing sessions, (b) generalize across decision domains when demand characteristics are held constant, and (c) are not epiphenomenal of another trait or simply reflect a lack of motivation to perform well on the tasks. Most studies comparing the cross-task stability of criterion shifting occur within a single research visit (Aminoff et al., 2012; Frithsen et al., 2018; Kantner et al., 2015). However, individual criterion shifting tendencies may change over time. Franks and Hicks (2016), to our knowledge, provide the only report that criterion shifting in recognition memory is stable over time, but these results only included two time points across two days. We assessed the stability of criterion shifting over longer time periods in Experiments 1 and 2, where participants conducted test–retest recognition memory tasks on 10 separate days across six weeks. In Experiment 3, we examined the test–retest reliability of criterion shifting tendencies across decision domains by comparing performance on recognition memory and visual detection tests (with equivalent demand characteristics) across two separate testing sessions. In both sessions, participants also conducted a test–retest battery of other cognitive tasks and questionnaires to determine whether other factors are related to individual criterion shifting tendencies. Aminoff and colleagues (2012) attempted to search for factors, such as cognitive style, personality traits, and executive functioning skills, that may explain individual differences in the extent of criterion shifting. Despite administering many questionnaires that assess a wide variety of cognitive and personality characteristics, only a few measures showed significant relationships with the extent of criterion shifting. These included positive relationships with a fun-seeking personality and verbal cognitive style as well as a negative relationship with characteristics of a negative affect. However, no published studies have attempted to replicate these findings. In Experiment 3, we attempted to both replicate some of these previously reported relationships and probe several novel factors, such as performance on tasks assessing risk aversion, response inhibition, working memory, and task-switching ability. We also examined whether a motivation to perform well on these tasks related to criterion shifting tendencies.

Another intriguing possibility is that individual differences in strategic criterion shifting tendencies are related to individual decision strategies for reporting confidence in recognition memory judgments. In lieu of criterion manipulations, many recognition memory studies implement confidence ratings to measure criterion shifts because confidence judgments require individuals to establish multiple decision criteria to differentiate between various levels of memory strength (Macmillan & Creelman, 2005; Yonelinas & Parks, 2007). Because many recognition studies implement confidence ratings to assess criterion shifts, it is reasonable to predict that both encompass similar decision processes. The ability to provide a confidence rating to a recognition memory judgment is believed to represent an individual's meta awareness for the amount of familiarity strength that an item elicits (Koriat, 2007). Confidence ratings are oftentimes used to assess measures of meta awareness, such as metacognitive sensitivity (how accurately one can distinguish between correct and incorrect responses) and metacognitive bias (the propensity to make judgments

with the highest levels of confidence regardless of performance; Fleming & Lau, 2014). Previous studies demonstrated that individuals typically have high metacognitive sensitivity, where judgments made with high confidence are generally more accurate than low confidence judgments (Mickes, Hwe, Wais, & Wixted, 2011; Mickes, Wixted, & Wais, 2007). However, Mickes and colleagues (2011) found individual differences in metacognitive bias—some reserve the most confident responses for the strongest (or weakest) of memories while others make high confidence responses more frequently, even when discriminability is held constant. Similar to strategic criterion shifting, obtaining appropriate confidence ratings from participants requires very little training and occurs immediately. Given the commonalities between rating confidence and strategic criterion shifting, it is possible that criterion shifting tendencies are driven by an individual's meta awareness of familiarity strength in test items. If this is the case, then individual tendencies to shift criteria should be directly related to measures of metacognitive sensitivity or metacognitive bias. For example, people with high metacognitive sensitivity might be more *capable* of shifting criteria to large extents relative to those who struggle with distinguishing between strong versus weak memory evidence. Individuals with high metacognitive bias have a lax standard for the amount of memory evidence needed (or lack thereof) to make recognition judgments with high confidence, which could result in smaller criterion shifts if participants shift criteria based on the level of confidence in “old” or “new” responses. Alternatively, strategic criterion shifting tendencies and tendencies to report meta awareness in test items via confidence ratings may represent completely different decision process. Miller and Kantner (2020) conducted post hoc analyses on previously reported data in which participants both shifted criteria and rated confidence, but found no relationship between the extent of criterion shifting and metacognitive bias. However, no studies have systematically assessed this relationship a priori. In Experiments 4 and 5, we examined whether individual differences in strategic criterion shifting tendencies related to metacognitive sensitivity or metacognitive bias.

Taken together, we predicted that individual tendencies in the extent of strategic criterion shifting would be stable across time and decision domains while being unrelated to other factors such as performance on other cognitive tasks, personality traits, motivation to perform well, metacognitive sensitivity, or metacognitive bias. We believe that stable differences in criterion shifting tendencies reflect individual differences in people's *willingness* to ignore uncertain evidence in favor of a decision strategy that optimizes decisional outcomes (Aminoff et al., 2012; Green & Swets, 1966; Kantner et al., 2015; Miller & Kantner, 2020) as opposed to individual differences in people's *ability* to shift criteria. Because we believe all individuals should be *capable* of shifting criteria to great extents, we do not expect strategic criterion shifting tendencies to be related to other abilities or characteristics.

## General Method

### Participant Recruitment

Participants enrolled in the five experiments via the University of California Santa Barbara (UCSB) paid research participation website. The experiments received approval from the UCSB Human Subjects Committee Institutional Review Board and all participants provided written informed consent.

## Signal Detection Theory

Across all five experiments we used an equal-variance signal detection theory (SDT) model to calculate discriminability ( $d'$ ), criterion placement ( $c$ ), and criterion shifting ( $C$ ; Macmillan & Creelman, 2005). For each test condition, we summed the number of hit (H), miss (M), correct rejection (CR), and false alarm (FA) trials to compute hit rate ( $HR$ ), false alarm rate ( $FAR$ ), percent correct ( $PC$ ), and SDT measures through the following equations:

$$\begin{aligned} HR &= H/(H + M) \\ FAR &= FA/(CR + FA) \\ d' &= z(HR) - z(FAR) \\ c &= -0.5 \times [z(HR) + z(FAR)] \\ C &= c(\text{conservative}) - c(\text{liberal}) \\ PC &= (H + CR)/(H + M + CR + FA), \end{aligned}$$

where  $z$  represents the density of the standard normal distribution (Macmillan & Creelman, 2005; Stanislaw & Todorov, 1999). To prevent infinite normalized values, we adjusted rare occurrences of  $HR$ s and  $FAR$ s of 0% and 100% by adding or subtracting, respectively, 1 divided by the total number of trials within a test condition (see Macmillan & Kaplan, 1985). Of the 5,510  $HR$ s computed across all five experiments, 88 required correction whereas a total of 155 of 5,510  $FAR$ s underwent correction.

Equal-variance SDT models assume that the variance of the target and lure distributions are equal. However, recognition memory experiments reveal that the target distribution typically has greater variance than the lure distribution indicating that unequal-variance SDT models provide more accurate measures of discriminability and criterion placement (Egan, 1958; Mickes et al., 2007). The challenge with implementing an unequal-variance SDT model is that it requires many criterion manipulations or confidence ratings to accurately assess the degree to which the variance of the target and lure distributions are unequal (Macmillan & Creelman, 2005). We therefore implement an equal-variance SDT model, because all five experiments include criterion shift tasks that only have two or three criterion manipulations. In the [online supplemental materials](#) we report results from an unequal-variance SDT model, which are generally consistent with the findings from the equal-variance SDT model in regards to the stability of individual criterion shifting tendencies.

Criterion placement and discriminability are behaviorally independent processes; however, a statistical relationship exists in SDT between the optimal criterion placement of an ideal observer<sup>3</sup> and the extent of discriminability when a biased decision criterion is advantageous (i.e., the more uncertain the discrimination, the more extreme the criterion should be; Macmillan & Creelman, 2005). Therefore, we must control for potential changes in criterion placement that simply arise from changes in discriminability. To do this, we residualize  $c$  against  $d'$  across all participants within each test condition and session to obtain normalized  $c$  ( $c_n$ ) values, which ensures statistical independence (see Aminoff et al., 2012). This computation consists of correlating  $c$  with  $d'$  and adding the residuals of  $c$  to the grand mean of  $c$  to obtain  $c_n$  values. This ensures no linear relationship between  $c_n$  and  $d'$  values (i.e.,  $r = 0$ ) across participants within the specific test conditions of each session (e.g., conservative criterion condition in Session 1). This correction is advantageous because it removes the correlation

between  $c$  and  $d'$  while maintaining the same group average for  $c$  (i.e., mean  $c_n = \text{mean } c$ ). We obtained normalized  $C$  ( $C_n$ ) values by taking  $c_n$  in the conservative condition and subtracting  $c_n$  in the liberal condition. In the [online supplemental materials](#) we report results using  $c$  values without correcting for changes in  $d'$ , which produce similar results as the  $c_n$  values with regard to criterion shifting stability.

Although we believe that some form of the standard measure of  $c$  best represents a strategic threshold of memory strength for responding “old” on a recognition test, there are limitations to this measure. For instance,  $c$  cannot determine the source of the biases, which may not necessarily arise from decision processes alone (Witt, Taylor, Sugovic, & Wixted, 2015). Our experiments induced low discriminability and included extreme criterion manipulations in an attempt to induce large strategic criterion shifts, which should overpower the contributions of nondecisional biases. Additionally, we only changed the instructions for each criterion manipulation (not the task parameters themselves) and randomized conditions both between and within subjects, so biases arising from nondecisional sources should be relatively equivalent across criterion conditions. Other common measures of decision bias, such as relative criterion ( $c' = c/d'$ ) and log likelihood ratios ( $\ln[\beta] = c \times d'$ ) become nonmonotonic at very low levels of  $d'$  and reach infinity or zero, respectively, when  $d' = 0$ . Because several of our testing conditions resulted in very low mean  $d'$  values, we avoided using such measures. Nevertheless, in the [online supplemental materials](#) we include a table of many decision bias measures, which demonstrate that the consistency of criterion shifting tendencies across sessions and tasks are fairly similar regardless of the criterion shift measure.

## Statistical Analysis

Effect size measures of Cohen’s  $d$  and Pearson  $r$  correlation coefficients are reported with 95% CIs. Any CI spanning zero is considered nonsignificant. When assessing whether multiple correlation coefficients are statistically significant, we controlled for the false discovery rate (FDR) as described by Benjamini and Hochberg (1995). Averaged group results are presented with  $SD$  values that are adjusted for within-subject variables using the method described in Morey (2008). For nonsignificant Pearson correlation coefficients, we implemented the BayesFactor package (Morey et al., 2018) in R to compute Bayes factors that assess the strength of evidence for the null versus alternative hypotheses ( $BF_{01}$ ) using uninformed uniform priors.  $BF_{01}$  values greater than three are considered strong evidence for the null hypothesis (see Jeffreys, 1961).

## Materials

Experiments 1, 2, 4, and 5 contained face stimuli drawn from the 10k U.S. Adult Faces database (Bainbridge, Isola, & Oliva, 2013).

<sup>3</sup> We use the term “ideal observer” to reference an individual who responds in a way that maximizes accuracy or payoffs (depending on the intended goals of the task) in an SDT framework. However, “ideal” performance will differ depending on the theoretical model and the specific goals of the individual, which may differ from the intended goals of the task (see Malmberg, 2008).

The stimulus set of Experiment 3 contained two versions of 1,024 scene images. One version contained a single person whereas an edited version did not include a person. All scene stimuli derived from a cropped  $500 \times 500$  pixel portion of images found on several open source online databases. Participants conducted all tasks at a computer using MATLAB version R2016B that incorporated open source code from Psychophysics Toolbox, v3 (Brainard, 1997).

## Experiments 1 and 2

In determining whether the tendency to strategically criterion shift is a stable cognitive trait, we first assessed the test–retest reliability of criterion shifting during recognition memory tests for faces across 10 sessions in the span of six weeks. In Experiment 1 we used a payoff manipulation to incentivize criterion shifting, where participants received payment at the end of each session based *entirely* on an individual’s performance. Participants earned five cents for each correct response, lost 10 cents for critical errors (either false alarms or misses), but received no penalty for non-critical errors. The likelihood of encountering old and new images remained equal in Experiment 1, but criterion shifting in this paradigm is advantageous for avoiding costly critical errors. When the critical error is a false alarm, participants should maintain a conservative criterion, but a liberal criterion becomes advantageous when the critical error is a miss. Perfect accuracy during an Experiment 1 session would result in a payment of \$30, but participants could easily earn \$15 by simply maximizing responses (i.e., always choosing the response that went unpenalized if incorrect). Experiment 2 followed similar procedures as Experiment 1 except a base rate manipulation induced criterion shifts and everyone earned \$10 per session regardless of performance. Franks and Hicks (2016) showed that a base rate manipulation during recognition tests for words produced a modest test–retest relationship in the extent of criterion shifting across two testing sessions separated by 48 hr,  $r(109) = .38$ . We investigated whether test–retest relationships in criterion shifting tendencies during recognition memory is sustained across many testing sessions.

Although criterion shifting tendencies are consistent across payoff and base rate manipulations when conducted within the same testing session (Frithsen et al., 2018; Kantner et al., 2015), we examined whether the extra monetary incentive to shift in Experiment 1 affected the stability of criterion shifting over *time*. For instance, participants in Experiment 1 who initially inadequately shift criteria may learn to shift to greater extents in subsequent sessions because doing so results in greater payouts. Participants in Experiment 2 lacked this additional monetary incentive to shift criteria, which could affect the long-term stability of criterion shifting. For example, some participants may shift criteria more extremely across sessions to improve accuracy whereas others may become less concerned about shifting criteria and more focused on completing the task as quickly as possible. If strategic criterion shifting tendencies are a stable cognitive trait, then there should be strong test–retest relationships in the extent of criterion shifting over time regardless of the type of criterion manipulation.

For strategic criterion shifting tendencies to be considered a stable trait, the test–retest reliability of criterion shifting should be as strong as other performance measures that are believed to reflect cognitive traits. In both experiments we included a neutral criterion

condition where participants either received no penalty for any errors (Experiment 1) or the likelihood of encountering a target or nontarget remained equal at test (Experiment 2). This allowed us to assess whether the test–retest reliability of criterion shifting is as stable as criterion placement in situations where criterion shifting yields no advantage, which Kantner and Lindsay (2012, 2014) identified as a stable cognitive trait (though this is by no means a standard for what should constitute a “stable cognitive trait”). If criterion shifting proves to be as stable as criterion placement in the neutral criterion condition, then we believe the tendency to strategically shift criteria should also be considered a stable cognitive trait.

An additional factor that may affect the stability of criterion shifting tendencies is the strength of discriminability (i.e., how well participants can distinguish between studied and novel test images). According to SDT, as the strength of discriminability increases the need to criterion shift decreases (Macmillan & Creelman, 2005). Though most people fail to shift criteria to an extent that maximizes accuracy or payoffs during recognition memory tests (Aminoff et al., 2012, 2015; Frithsen et al., 2018; Kantner et al., 2015), studies of the strength-based mirror effect reveal that changes in the level of discriminability can affect the extent of criterion shifting (Franks & Hicks, 2016; Glanzer & Adams, 1985), regardless of whether these changes are believed to result from strategic criterion shifts (Stretch & Wixted, 1998b) or not (Criss, 2006). In both experiments we altered the strength of discriminability after the first five sessions to assess whether the extent of criterion shifting becomes less (or more) stable as discriminability improves. In situations where a neutral criterion is most advantageous, SDT predicts that changes in discriminability should not impact the stability of criterion placement because an ideal observer should always maintain an unbiased criterion regardless of whether memory strength is strong or weak (Macmillan & Creelman, 2005). We predicted criterion shifting tendencies to be as stable as neutral criterion placement tendencies regardless of the criterion manipulation or level of discriminability.

## Method

**Participants.** Thirty-nine participants successfully completed all 10 test–retest sessions on separate days across six weeks in Experiment 1 (10 males;  $M = 19.7$  years, range = 18–28 years,  $SD = 1.7$ ). A separate sample of 39 participants completed Experiment 2 within a 6-week span (11 males;  $M = 19.8$  years, range = 18–37 years,  $SD = 3.1$ ). Additional participants failed to complete all 10 sessions in Experiment 1 (five) and Experiment 2 (seven) and are excluded from all analyses. Participants in Experiment 1 earned anywhere from \$5–\$30 per session depending on performance (see the Procedure section) whereas participants in Experiment 2 received \$10 per session. Participants in both experiments earned an additional \$50 bonus for completing all 10 sessions.

**Procedure.** The recognition memory task consisted of three blocks where each of the 10 self-paced sessions lasted for 20–60 min. A block consisted of a 100-image study phase followed by a 200-image test phase. Each session contained three different test phase conditions (conservative, liberal, and neutral) where instructions prior to each test phase (unless otherwise specified) *explicitly* informed participants of an advantage for establishing a conserva-

tive, liberal, or neutral decision criterion, respectively. A discriminability manipulation occurred after session five where the number of times each image appeared during the study phase changed. In the low discriminability condition, each study image appeared once whereas study images in the moderate discriminability condition appeared five times.

During each study phase, participants passively viewed a sequence of 100 unique face images on a black background in the center of a computer screen for 300 ms followed by a 200-ms blank screen presentation. The quick presentation time induced low discriminability making it more advantageous to shift criteria. During the test phases, each image appeared in the center of the screen with text displayed above the image to remind participants of the criterion condition. A number of “0” or “1” appeared at the bottom of the screen to indicate the keyboard button corresponding to an “old” (studied) or “new” (unstudied) response, which randomly changed on a trial-by-trial basis. Images remained on screen until the participant made a response. Participants received feedback at the end of each session indicating the amount of money earned (Experiment 1) or the percentage of correct trials obtained (Experiment 2) for the entire session.

In each test phase of Experiment 1, participants received five cents for correctly responding “old” to a studied image (a hit) and “new” to an unstudied image (a correct rejection). Incorrect re-

sponses consisted of penalized critical errors and penalty-free noncritical errors. In the conservative condition, participants lost 10 cents for responding “old” to an unstudied image (a false alarm) but did not lose money for responding “new” to a studied image (a miss). In the liberal condition, participants lost 10 cents for a miss, but received no penalty for a false alarm. In the neutral condition, participants did not lose any money for false alarms or misses. Participants conducted the conservative, liberal, and neutral test phases in three separate blocks, each of which included 100 studied and 100 novel images. We presented the conservative, liberal, and neutral test blocks in a pseudorandom order across sessions and subjects. All participants conducted the low discriminability condition for Sessions 1–5 and the moderate discriminability condition for Sessions 6–10 (Figure 1).

In each test phase of Experiment 2, a studied item appeared 25% (conservative), 75% (liberal), or 50% (neutral) of the time during a test block. Importantly, the probability manipulations in Experiment 2 and the payoff manipulations in Experiment 1 required the same conservative, liberal, and neutral criterion placements for *optimal* performance according to the equal-variance SDT model (see Macmillan & Creelman, 2005). Unlike Experiment 1, participants did NOT receive information about the 50% likelihood of encountering a studied item in the neutral condition to ensure that explicit instructions did not affect an individual’s criterion place-

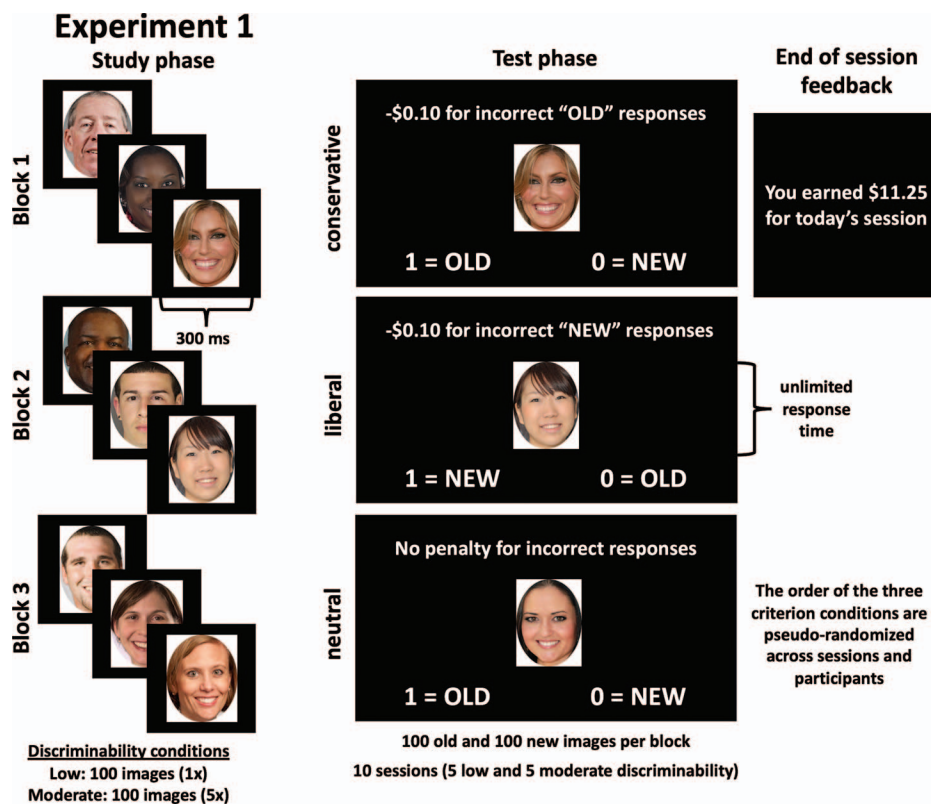


Figure 1. Experiment 1 recognition memory task. A 100-image study block preceded a 200-image test block for each criterion condition (conservative, liberal, and neutral). A payoff manipulation induced criterion shifts where participants earned 5 cents for correct responses, lost 10 cents for critical errors, but did not lose money for noncritical errors. At the end of each session, participants received feedback on total money earned for that session. See the online article for the color version of this figure.



ment tendencies. The neutral criterion block always occurred first, whereas test blocks 2 and 3 consisted of two parts: a 100-image conservative (25 studied, 75 novel images) and a 100-image liberal (75 studied, 25 novel images) test mini block. Instructions appeared before each mini block to indicate the likelihood of encountering a studied item. The mini block presentation order appeared pseudorandomly so that each session consisted of the two possible order types (conservative before liberal, or vice versa) and the block orders switched every other session (conservative in test block two or three). Twenty participants conducted the low discriminability condition in Sessions 1–5 and moderate discriminability condition in Sessions 6–10, whereas 19 participants conducted the low and moderate discriminability conditions in the reverse order (Figure 2). Experiments 1 and 2 included the same 6,000 unique face images (600 per session) and each participant received a completely randomized assignment and presentation order for the target and nontarget images.

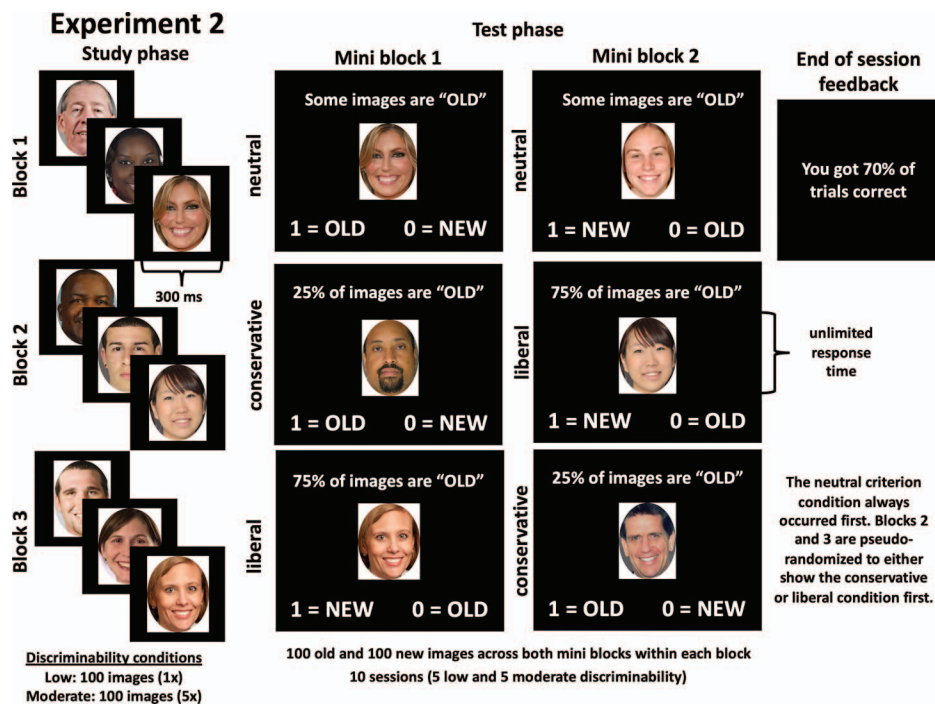
**Statistical analysis.** Though we did not preregister our hypotheses for Experiments 1 and 2, we predicted  $C_n$  to be as stable as  $c_n$  in the neutral criterion condition (neutral  $c_n$ ), regardless of the criterion manipulation or extent of  $d'$ . Here we report Pearson  $r$  correlation coefficients for the nine session-to-session comparisons as well as all 45 pairwise comparisons across the 10 sessions for values of  $C_n$  and neutral  $c_n$ . We also compare the relationship between  $C_n$  and neutral  $c_n$  across all 10 sessions as well as  $C_n$  and the absolute value of neutral  $c_n$  (as a measure of criterion extremeness). In the [online supplemental materials](#) we report session-to-

session correlation coefficients for criterion shifting and neutral criterion placement using a variety of different SDT measures.

## Results and Discussion

**Experiment 1.** Mean  $d'$  across all low discriminability sessions ( $M = 0.41$ ,  $SD = 0.43$ ) remained lower compared with the mean  $d'$  of the moderate discriminability sessions ( $M = 0.87$ ,  $SD = 0.59$ ),  $d = 1.00$ , 95% CI [0.88, 1.12], confirming that viewing stimuli once versus five times during the study phase effectively modulated discriminability. On average, participants in the low discriminability condition shifted between  $c_n$  in the conservative ( $M = 1.03$ ,  $SD = 0.63$ ) and liberal criterion conditions ( $M = -1.00$ ,  $SD = 0.71$ ),  $d = 2.94$ , 95% CI [2.65, 3.23], as well in the moderate discriminability condition between the conservative ( $M = 1.00$ ,  $SD = 0.56$ ) and liberal criterion conditions ( $M = -0.98$ ,  $SD = 0.70$ ),  $d = 3.05$ , 95% CI [2.76, 3.35]. Surprisingly, mean  $C_n$  did not significantly differ between the low ( $M = 2.03$ ,  $SD = 1.00$ ) and moderate ( $M = 1.98$ ,  $SD = 0.77$ ) discriminability conditions,  $d = 0.05$ , 95% CI [-0.15, 0.25], even though SDT predicts that lower discriminability will lead to greater criterion shifts (see Macmillan & Creelman, 2005). Table 1 shows a complete list of mean  $c_n$ ,  $C_n$ ,  $d'$ , and  $PC$  values for each session as well as for the low and moderate discriminability conditions combined across sessions.

In the neutral criterion condition,  $c_n$  remained fairly consistent across the 10 sessions,  $r(37)$  session-to-session range: .45–.74,



*Figure 2.* Experiment 2 recognition memory task. A 100-image study block preceded two 100-image test mini blocks for each criterion condition (conservative, liberal, and neutral). A base rate manipulation induced criterion shifts where "old" images appeared either 25% (conservative), 50% (neutral), or 75% (liberal) of the time. At the end of each session, participants received feedback on the percentage of correct trials for that entire session. See the online article for the color version of this figure.



Table 1  
 Experiment 1 Mean and Standard Deviation Values (in Parentheses) for  $c_n$  (Three Criterion Conditions),  $C_n$ ,  $d'$ , and PC for Each of the 10 Sessions and Collapsed Across the Low and Moderate (Mod) Discriminability Conditions

Session	$c_n$			$C_n$	$d'$	PC
	Conservative	Neutral	Liberal			
1	0.75 (0.59)	0.22 (0.41)	-0.76 (0.67)	1.51 (0.89)	0.36 (0.28)	55.28% (4.22)
2	1.06 (0.53)	0.00 (0.44)	-1.05 (0.66)	2.11 (0.81)	0.39 (0.30)	55.03% (4.30)
3	1.06 (0.60)	0.00 (0.61)	-1.04 (0.59)	2.10 (0.53)	0.45 (0.33)	55.59% (4.01)
4	1.10 (0.60)	0.13 (0.45)	-1.04 (0.60)	2.14 (0.63)	0.38 (0.34)	54.92% (3.83)
5	1.18 (0.53)	0.13 (0.50)	-1.10 (0.71)	2.28 (0.56)	0.46 (0.33)	55.36% (4.04)
6	0.93 (0.57)	0.07 (0.54)	-0.86 (0.65)	1.78 (0.64)	0.88 (0.44)	61.74% (5.45)
7	1.05 (0.48)	0.16 (0.38)	-1.08 (0.73)	2.13 (0.47)	0.89 (0.43)	61.15% (5.62)
8	1.00 (0.46)	-0.02 (0.39)	-0.97 (0.60)	1.96 (0.50)	0.87 (0.45)	61.05% (5.43)
9	0.99 (0.54)	0.09 (0.41)	-0.99 (0.66)	1.98 (0.59)	0.86 (0.41)	61.00% (5.99)
10	1.04 (0.57)	0.08 (0.45)	-0.98 (0.59)	2.02 (0.62)	0.84 (0.47)	61.01% (6.28)
Low	1.03 (0.63)	0.10 (0.53)	-1.00 (0.71)	2.03 (1.00)	0.41 (0.43)	55.24% (5.47)
Mod	1.00 (0.56)	0.08 (0.47)	-0.98 (0.70)	1.98 (0.77)	0.87 (0.59)	61.19% (7.72)

Note.  $c_n$  = normalized criterion placement;  $C_n$  = normalized criterion shifting;  $d'$  = discriminability; PC = percent correct.

$Mdn = .54$ ; all pairwise comparisons range:  $.18-.74$ ,  $Mdn = .51$ , which is comparable with all of the test-retest relationships of  $c$  reported by Kantner and Lindsay (2012, 2014; range:  $.31-.81$ ,  $Mdn = .64$ ). Correlations of  $C_n$ ,  $r(37)$  session-to-session range:  $.58-.85$ ,  $Mdn = .75$ ; all pairwise comparisons range:  $.38-.85$ ,  $Mdn = .68$ , remained high despite the discriminability manipulation that occurred after the first five sessions. Figure 3 shows matrices of Pearson correlations for all 45 pairwise comparisons of neutral  $c_n$  and  $C_n$ . No significant relationships existed between  $C_n$  and neutral  $c_n$  across any of the 10 sessions after FDR correction,  $r(37)$  range:  $-.14-.34$ ,  $Mdn = .01$ ;  $BF_{01}$  range:  $0.56-4.96$ ,  $Mdn = 4.20$ , providing support to the assumption that criterion shifting and placement are independent behaviors. We also tested whether individual tendencies to establish extreme criteria in the

neutral condition (regardless of whether individuals established more conservative or liberal criteria) related to the extent of criterion shifting, but found no consistent relationship between  $C_n$  and the absolute value of neutral  $c_n$  after FDR correction,  $r(37)$  range:  $-.14-.43$ ,  $Mdn = .09$ ;  $BF_{01}$  range:  $0.13-4.94$ ,  $Mdn = 4.10$ . Figure 4 displays  $c_n$  values for each criterion condition across all 10 sessions, ordered from left to right based on who shifted criteria to the greatest extent during Session 10.

**Experiment 2.** Although some participants conducted the low discriminability condition in Sessions 1-5 and others in Sessions 6-10, we report statistics with all 39 subjects together (see the online supplemental materials for analyses of the two groups separately). Mean differences are computed within the discriminability conditions regardless of session number and correlation

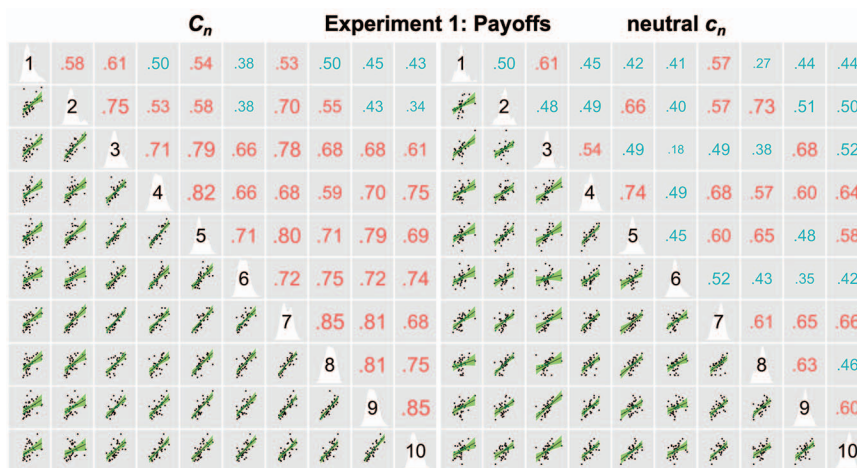
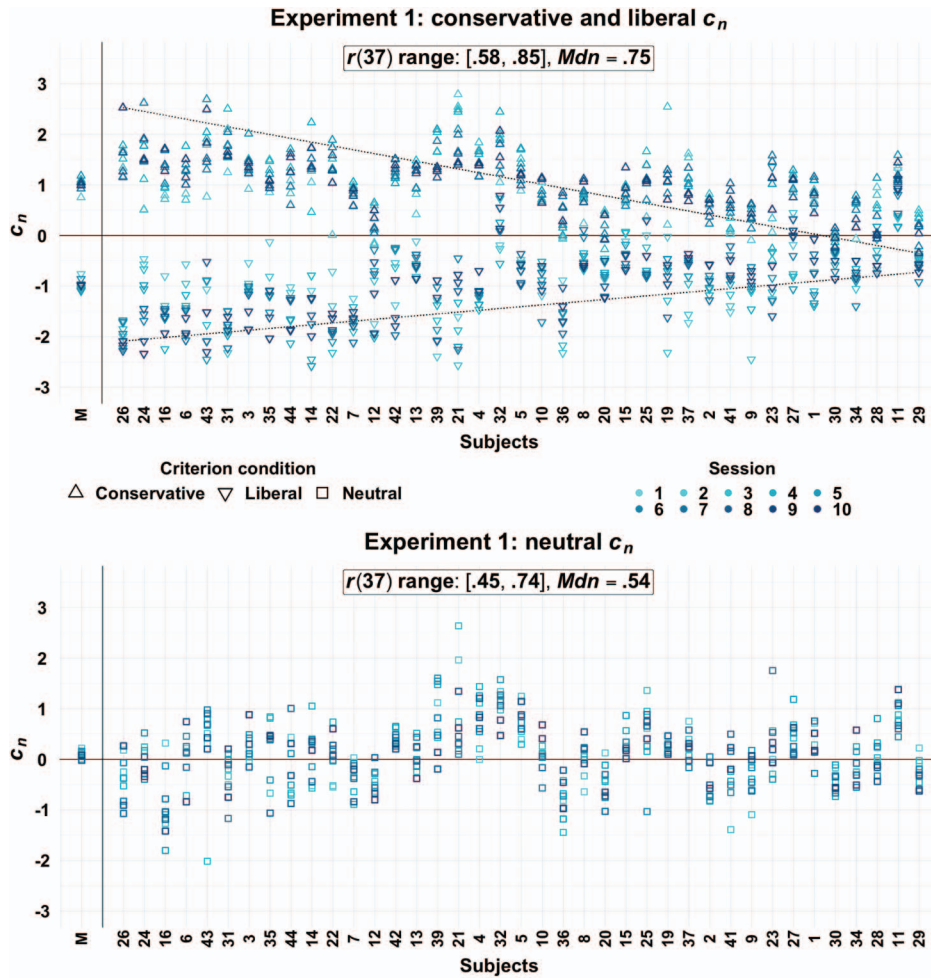


Figure 3. Experiment 1 Pearson correlation matrices comparing all 45 pairwise comparisons of normalized criterion shifting ( $C_n$ ; left) and neutral normalized criterion placement ( $c_n$ ; right). The left side of each matrix displays the regression line for each comparison whereas the right side shows Pearson  $r$  values (red values (values  $\geq .53$ ) are  $p < .001$ , FDR-corrected). The diagonal displays the session number along with the distribution of values (in white) for each session. See the online article for the color version of this figure.



*Figure 4.* Experiment 1 normalized criterion placement ( $c_n$ ) values for each participant and the mean ( $M$ ) in the conservative and liberal criterion conditions (top) as well as the neutral criterion condition (bottom) across the 10 sessions. The extent of criterion shifting is depicted by the distance between the triangles representing the conservative and liberal  $c_n$  values. Participants are ordered from left to right based on who shifted criteria the most during Session 10. Participants on the left have a large spread between conservative and liberal  $c_n$  values, but the magnitude of the spread steadily decreases as you view subjects from left to right. The dotted lines emphasize this criterion shifting trend by connecting the Session 10 conservative and liberal  $c_n$  values of the leftmost and rightmost subjects. The range and median session-to-session Pearson correlation coefficients are shown for normalized criterion shifting ( $C_n$ ; top) and neutral  $c_n$  (bottom) below the graph titles. See the online article for the color version of this figure.

coefficients are computed across session number regardless of the order of the discriminability conditions.

Mean  $d'$  remained lower in the low discriminability sessions ( $M = 0.27$ ,  $SD = 0.46$ ) relative to the moderate discriminability sessions ( $M = 0.85$ ,  $SD = 0.66$ ),  $d = 1.16$ , 95% CI [1.04, 1.28]. On average, participants in the low discriminability condition shifted between  $c_n$  in the conservative ( $M = 0.86$ ,  $SD = 0.59$ ) and liberal criterion conditions ( $M = -0.34$ ,  $SD = 0.73$ ),  $d = 1.83$ , 95% CI [1.59, 2.06], as well as in the moderate discriminability condition between the conservative ( $M = 0.82$ ,  $SD = 0.56$ ) and liberal criterion conditions ( $M = -0.53$ ,  $SD = 0.53$ ),  $d = 2.61$ , 95% CI [2.34, 2.88]. As in Experiment 1, mean  $C_n$  did not significantly differ in the low ( $M = 1.20$ ,  $SD = 0.85$ ) versus moderate ( $M = 1.35$ ,  $SD = 0.89$ ) discriminability conditions,  $d =$

0.15, 95% CI [-0.05, 0.35]. Table 2 shows a complete list of mean  $c_n$ ,  $C_n$ ,  $d'$ , and  $PC$  values for each session as well as for the low and moderate discriminability conditions combined across sessions.

Similar to Experiment 1, correlation coefficients for neutral  $c_n$ ,  $r(37)$  session-to-session range: .68–.82,  $Mdn = .76$ ; all pairwise comparisons range: .10–.82,  $Mdn = .59$ , are comparable with the test–retest relationships of  $c$  reported by Kantner and Lindsay (2012, 2014; range: .31–.81,  $Mdn = .64$ ). Strong correlation coefficients of  $C_n$  persisted across sessions,  $r(37)$  session-to-session range: .71–.89,  $Mdn = .83$ ; all pairwise comparisons range: .11–.90,  $Mdn = .67$ , despite counterbalancing the order in which participants conducted the low and moderate discriminability conditions. Figure 5 shows matrices of Pearson correlations for all 45 pairwise comparisons of neutral  $c_n$  and  $C_n$ . Across the 10 sessions,

Table 2  
 Experiment 2 Mean and Standard Deviation Values (in Parentheses) for  $c_n$  (Three Criterion Conditions),  $C_n$ ,  $d'$ , and PC for Each of the 10 Sessions and Collapsed Across the Low and Moderate (Mod) Discriminability Conditions

Session	$c_n$			$C_n$	$d'$	PC
	Conservative	Neutral	Liberal			
1	0.60 (0.51)	0.16 (0.33)	-0.36 (0.39)	0.96 (0.85)	0.67 (0.51)	66.62% (8.40)
2	0.70 (0.42)	0.18 (0.30)	-0.27 (0.42)	0.97 (0.51)	0.63 (0.50)	65.91% (9.55)
3	0.83 (0.49)	0.23 (0.34)	-0.38 (0.54)	1.20 (0.57)	0.62 (0.51)	66.29% (9.33)
4	0.81 (0.43)	0.34 (0.30)	-0.34 (0.52)	1.15 (0.52)	0.62 (0.50)	66.06% (9.49)
5	0.81 (0.41)	0.37 (0.24)	-0.30 (0.60)	1.11 (0.50)	0.51 (0.48)	64.06% (9.11)
6	0.85 (0.45)	0.25 (0.25)	-0.41 (0.56)	1.26 (0.47)	0.58 (0.59)	66.34% (10.34)
7	0.91 (0.63)	0.34 (0.25)	-0.49 (0.72)	1.40 (0.68)	0.55 (0.55)	65.79% (9.74)
8	1.04 (0.57)	0.37 (0.26)	-0.56 (0.70)	1.59 (0.69)	0.54 (0.56)	66.25% (10.13)
9	0.91 (0.57)	0.32 (0.46)	-0.61 (0.66)	1.53 (0.56)	0.49 (0.51)	65.60% (9.14)
10	0.95 (0.66)	0.33 (0.37)	-0.61 (0.70)	1.57 (0.71)	0.40 (0.47)	64.18% (9.83)
Low	0.86 (0.59)	0.38 (0.32)	-0.34 (0.73)	1.20 (0.85)	0.27 (0.46)	61.50% (11.90)
Mod	0.82 (0.56)	0.20 (0.34)	-0.53 (0.53)	1.35 (0.89)	0.85 (0.66)	69.92% (10.68)

Note.  $c_n$  = normalized criterion placement;  $C_n$  = normalized criterion shifting;  $d'$  = discriminability; PC = percent correct.

only 1 significant relationship existed between  $C_n$  and neutral  $c_n$  after FDR correction (Session 8:  $r[37] = -.45$ , 95% CI  $[-.67, -.15]$ ). However, no obvious relationships existed when considering all 10 sessions together,  $r(37)$  range:  $-.45-.19$ ,  $Mdn = -.14$ ;  $BF_{01}$  range: 0.10-4.99,  $Mdn = 2.33$ . We also found no consistent relationship between  $C_n$  and the absolute value of neutral  $c_n$  across the 10 sessions,  $r(37)$  range:  $-.36-.21$ ,  $Mdn = -.04$ ;  $BF_{01}$  range: 0.44-5.00,  $Mdn = 3.63$ . Figure 6 displays  $c_n$  values for each criterion condition across all 10 sessions, ordered from left to right based on who shifted criteria to the greatest extent during Session 10.

Comparing across Experiments 1 and 2, participants on average shifted criteria to a greater extent in Experiment 1 ( $M = 2.00$ ,  $SD = 1.08$ ) versus Experiment 2 ( $M = 1.27$ ,  $SD = 1.02$ ),  $d = 0.69$ ,

95% CI [0.55, 0.84]. We believe this occurred because the monetary incentive encouraged some individuals to shift criteria to greater extents in Experiment 1, because doing so increased total payout. Participants in Experiment 2 lacked this extra incentive to shift criteria because everyone received the same payment regardless of performance. However, it is important to note that some individuals in Experiment 1 did not shift to great extents (even by the 10th session; see the rightmost subjects in the top graph of Figure 4), whereas some individuals in Experiment 2 consistently shifted to large extents even though doing so did not affect the amount of money received (see the leftmost subjects in the top graph of Figure 6). This suggests that there could be individual differences in the factors that motivate individuals to shift criteria to greater extents. Future studies must confirm our prediction that



Figure 5. Experiment 2 Pearson correlation matrices comparing all 45 pairwise comparisons of normalized criterion shifting ( $C_n$ ; left) and neutral criterion placement ( $c_n$ ; right). The left side of each matrix displays the regression line for each comparison whereas the right side shows Pearson  $r$  values (red values (values  $\geq .53$ ) are  $p < .001$ , FDR-corrected). The diagonal displays the session number along with the distribution of values (in white) for each session. See the online article for the color version of this figure.



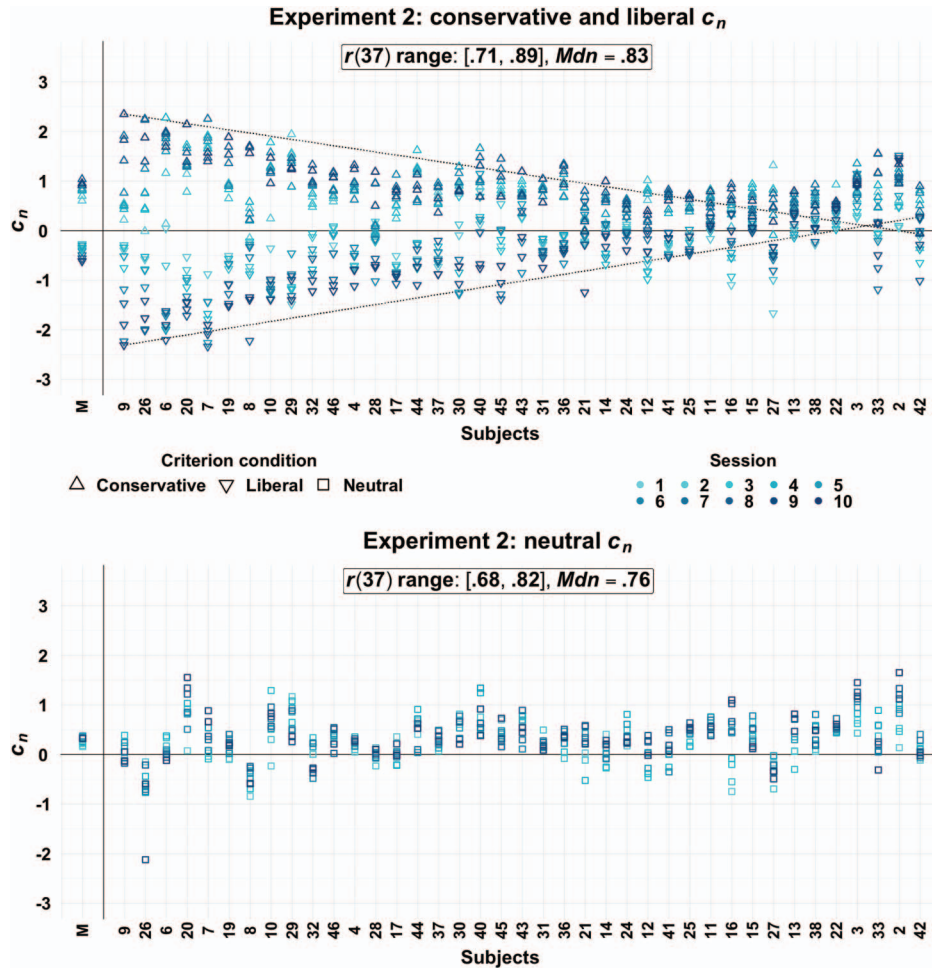


Figure 6. Experiment 2 normalized criterion placement ( $c_n$ ) values for each participant and the mean ( $M$ ) in the conservative and liberal criterion conditions (top) as well as the neutral criterion condition (bottom) across the 10 sessions. The extent of criterion shifting is depicted by the distance between the triangles representing the conservative and liberal  $c_n$  values. Participants are ordered from left to right based on who shifted criteria the most during Session 10. Participants on the left have a large spread between conservative and liberal  $c_n$  values, but the magnitude of the spread steadily decreases as you view subjects from left to right. The dotted lines emphasize this criterion shifting trend by connecting the Session 10 conservative and liberal  $c_n$  values of the leftmost and rightmost subjects. The range and median session-to-session Pearson correlation coefficients are shown for normalized criterion shifting ( $C_n$ ; top) and neutral  $c_n$  (bottom) below the graph titles. See the online article for the color version of this figure.

monetary incentives motivate some individuals to shift criteria to larger extents because it is possible that people may generally shift criteria to lesser extents in response to a base rate versus payoff manipulation for reasons that are unrelated to motivating factors.

Experiments 1 and 2 revealed that strategic criterion shifting tendencies during recognition memory are stable across multiple sessions regardless of the criterion manipulation (payoff or base rates) or the strength of discriminability. Some participants consistently shifted criteria to large extents, others regularly shifted criteria to moderate degrees, while some individuals hardly shifted criteria at all. The stability of criterion shifting showed no relationship with neutral criterion placement tendencies indicating that placing and shifting a criterion are independent behaviors.

### Experiment 3

Experiments 1 and 2 showed that criterion shifting is stable over time, at least during recognition memory tests. To further these findings, we tested whether the test-retest reliability of the extent of criterion shifting is stable both within and across decision domains. Frithsen and colleagues (2018) revealed that the extent of criterion shifting is largely consistent when making recognition memory judgments versus visual detection or visual discrimination judgments, but the strength of these relationships can sometimes vary. One possible explanation for this discrepancy is that differing demand characteristics across tasks may sometimes affect the stability of criterion shifting (Kantner et al., 2015). When two tasks have different designs, it might differentially affect how individ-

uals strategically adapt a decision criterion. If a weak relationship exists between the extent of criterion shifting between two tasks with differing designs and decision domains, then it is impossible to know whether criterion shifting strategies are truly domain-specific or are simply affected by the particular task designs. To isolate the decision domain, we created recognition memory and visual detection tests with scene stimuli that tightly controlled for the design of the tasks. The setup remained exactly the same between the two tests with the only difference being whether participants reported if a scene appeared during an initial study phase (recognition memory) or if an image contained a person or not (visual detection). This allowed us to assess the test–retest stability of criterion shifting tendencies across multiple decision domains while controlling for potential demand characteristic effects.

In addition to performing recognition memory and visual detection tests, participants also performed a battery of cognitive tests to assess whether consistencies in criterion shifting tendencies can be explained by other cognitive abilities. Although no published studies, to our knowledge, compare the extent of criterion shifting with other task measures, we predicted criterion shifting tendencies to be unrelated to performance on other cognitive tasks. We believe that criterion shifting tendencies are a result of an individual's *willingness* to shift criteria as opposed to an *ability* to do so (Kantner et al., 2015; Miller & Kantner, 2020). With this assumption, everyone should be *capable* of shifting criteria to great extents without the need of any particular skill that may otherwise be required for other cognitive tasks. However, one could argue that strategically shifting criteria might be associated with other cognitive abilities because there are many cognitive factors that go into a criterion shift. For example, participants who are more risk averse may shift to greater extents to simply avoid critical errors detrimental to decisional outcomes. Individuals with exceptional working memory might be more skilled at maintaining a consistent criterion during the entire length of a test block. Response inhibition is likely necessary for inhibiting prepotent familiarity responses in favor of more optimal decision strategies based on the criterion manipulation. Individuals may require more general task-switching ability to adequately shift between conservative and liberal decision criteria. To test whether these cognitive abilities show a relationship with criterion shifting, participants performed four additional standardized tasks that assessed risk aversion, response inhibition, working memory, and task-switching ability during each session. Although there are countless cognitive abilities that could possibly be associated with individual criterion shifting tendencies, we assessed four cognitive abilities that could reasonably be related to criterion shift strategies given the commonalities between these abilities and aspects of the decision processes that underly criterion shifting.

The extent to which an individual shifts a criterion might also be related to how motivated a person is to perform well during the tasks or may be associated with other personality characteristics. Following the criterion shifting tasks, participants completed a motivation questionnaire to assess whether a relationship exists between self-reported motivation to perform well on the tasks and the extent of criterion shifting. After completing all cognitive tasks, participants conducted additional personality and cognitive style questionnaires. Aminoff and colleagues (2012) previously conducted a large-scale study to assess whether individual differ-

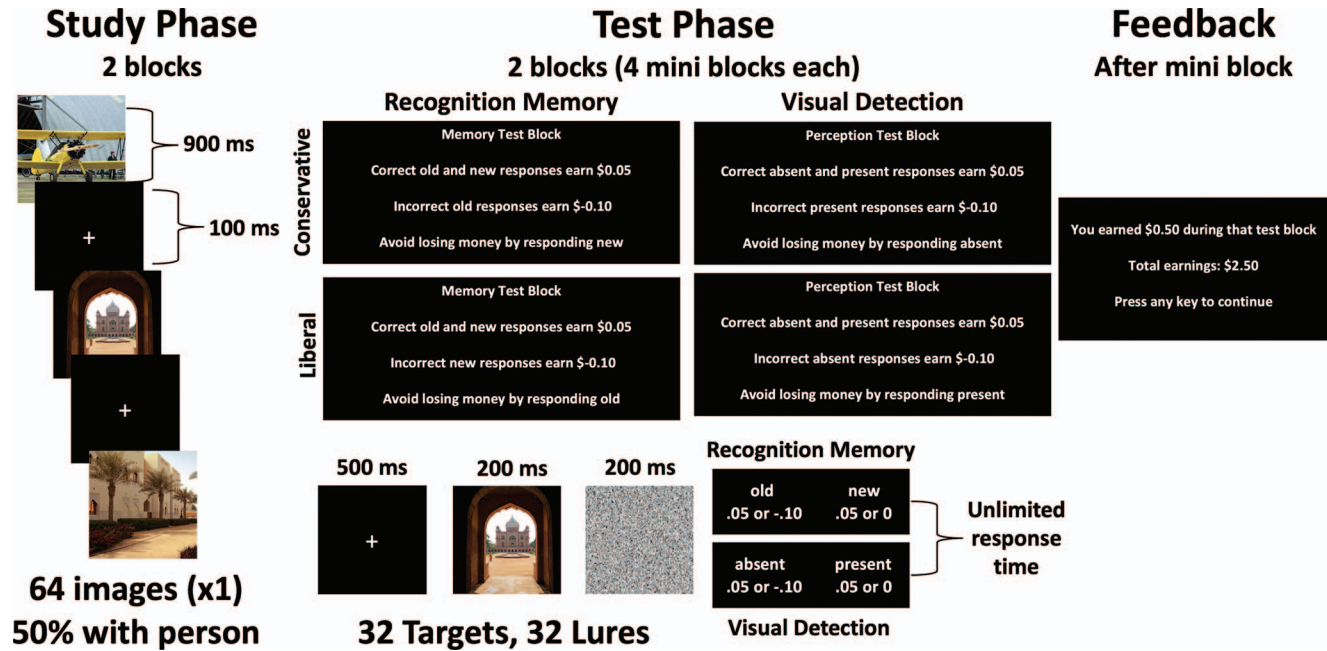
ences in criterion shifting during recognition memory tests are related to any personality or cognitive characteristics. Despite collecting more than 25 different standardized questionnaire measures that assess personality and cognitive characteristics, Aminoff and colleagues (2012) only found three questionnaire measures that significantly correlated with the extent of criterion shifting: positive relationships with a fun-seeking personality and verbal cognitive style as well as a negative relationship with traits associated with a negative affect. We therefore limited our questionnaire assessments to these three measures for replication purposes. As stated in our preregistration (<https://osf.io/jkfp6>), we predicted criterion shifting to be stable within and across decision domains while being unrelated to performance on other tasks and questionnaire measures.

## Method

**Participants.** One hundred seventy-two participants successfully completed both test–retest sessions (53 males;  $M = 19.9$  years, range = 18–30 years,  $SD = 2.0$ ). Exclusion of five additional participants occurred due to computer malfunctions (two) or incomplete data sets (three) and are not included in any analyses. Participant payment relied *entirely* on task performance (see the Procedures section) unless participants earned less than \$10 across all tasks during a session. Total payment for each session ranged from \$10–\$30 depending on performance.

**Procedure.** Participants completed two self-paced sessions on different days within the same week. Each session lasted between 45 and 90 min and included five computer tasks and four questionnaires. The computer tasks included (a) recognition memory and visual detection criterion shifting paradigms, (b) a Balloon Analogue Risk Task (BART), (c) a Go/No-go response inhibition task, (d) an *N*-Back working memory task, and (e) a Task-Switching paradigm. The questionnaires consisted of (a) the Effort/Importance section of the Intrinsic Motivation Inventory (IMI; Ryan, 1982), (b) the Behavioral Inhibition System and Behavioral Activation System (BIS/BAS) scales (Carver & White, 1994), (c) the Positive and Negative Affect Schedule—Expanded Form (PANAS-X; Watson & Clark, 1994), and (d) a modified version of the Verbalizer-Visualizer Questionnaire (VVQ; Richardson, 1977). Participants first conducted the recognition memory and visual detection criterion shifting paradigms followed by the IMI questionnaire. This allowed us to assess whether self-reports of motivation to perform well on the criterion shifting paradigms related to individual differences in criterion shifting tendencies. Afterward, participants conducted the other four computer tasks in a randomized order followed by the remaining three questionnaires. The questions within each questionnaire appeared in a random order. Although randomized across participants, the order of the additional tasks, questionnaires, and questions within each questionnaire remained the same for each participant across both sessions. At the end of each session, participants received payment based entirely on the amount of money earned during the criterion shifting paradigms and the BART.

**Recognition memory and visual detection criterion shifting paradigms.** After conducting a short practice task, participants completed two cycles of a study phase followed by two recognition memory and two visual detection test mini blocks (Figure 7). During each study phase, participants viewed a randomized se-



*Figure 7.* Recognition memory and visual detection (perception) tasks for Experiment 3. After each study phase, participants conducted four mini blocks, one for each task and criterion condition combination. To control for demand characteristics, the recognition memory and visual detection tests maintained the exact same structure except participants either responded as to whether an image appeared during the study phase or if a person appeared in the image, respectively. Participants received feedback on the amount of money earned after each mini block. See the online article for the color version of this figure.

quence of 64 unique scene stimuli, half of which contained a single person and half contained no people. Scenes appeared in the center of a computer screen on a black background for 900 ms followed by a 100-ms presentation of a white crosshair. During recognition memory test mini blocks, participants decided whether or not test images appeared during the study phase by responding “old” or “new,” respectively. For visual detection blocks, participants determined whether a person appeared in the image or not by responding “present” or “absent,” respectively. A payoff manipulation incentivized criterion shifting where participants earned five cents for correct responses, lost 10 cents for critical errors, but received no penalty for noncritical errors. In the conservative criterion condition, critical errors consisted of incorrect “old” responses during recognition memory tests and incorrect “present” responses during visual detection tests (false alarms). Critical errors in the liberal criterion condition included incorrect “new” and “absent” responses (misses). This created four test mini block conditions that each appeared once in a random order after every study phase: (a) conservative recognition memory, (b) liberal recognition memory, (c) conservative visual detection, and (d) liberal visual detection. Prior to each test block participants received explicit instructions detailing the task type (recognition memory or visual detection) and criterion condition (conservative or liberal). Each test mini block contained a randomized sequence of 64 stimuli (32 targets and 32 nontargets) that appeared for 200 ms followed by a 200 ms noise mask to destroy the perceptual afterimage. Afterward, participants made an old/new or present/absent judgment by using the “f” and “j” keys on a keyboard, in which pseudorandom assignment across participants mapped the

keys to each response type. On each trial, text appeared below each response type to remind participants of the critical and noncritical errors. Participants made responses with unlimited time and a 500-ms white crosshair presentation followed each response. After each test mini block, participants received feedback detailing the amount of money earned on that mini block as well as the running total of money earned on the task. Across both sessions, a single version of each stimulus (randomly assigned for each participant) appeared once during the test blocks (512 with a person, 512 without).

**BART.** The BART is a standardized computer task that assesses risk-taking behavior (Lejuez et al., 2002). Participants conducted 20 trials of a version of the BART where simulated balloons needed to be “pumped” to earn money. Participants earned one cent per pump and could collect money at any time during a trial unless the balloon popped. The balloon could pop anywhere from the first to the 128th pump determined by a random number generator with equal probability. Participants pressed the “f” or “j” key to either pump the balloon or collect money on a trial, depending on the pseudorandom assignment across subjects. Participants always saw the running total of money collected during the task as well as the amount of potential earnings that could be collected on a given trial. Trials ended when the balloon popped or the participant collected the money. Participants wore headphones which made pumping noises on each pump, a cash register sound when collecting money (with an accompanying green screen portraying the amount of money earned on that trial), and a popping noise when the balloon popped (with an accompanying red screen displaying the word “POP!”).



**Go/no-go.** The go/no-go task assessed inhibitory control and consisted of five, 40-trial blocks where participants needed to press the “j” key during common “go” trials while not responding to rare “no-go” trials. Participants viewed random sequences of yellow and cyan squares that either served as “go” or “no-go” trials based on pseudorandom assignment across subjects. Stimuli appeared for 800 ms followed by a 500-ms white crosshair presentation where each block contained 32 “go” trials and eight “no-go” trials. These parameters are within the recommended ranges for truly evoking response inhibition activity to a prepotent motor response on “no-go” trials (Wessel, 2018). Prior to the main task, participants conducted a 20-trial practice task.

#### **N-back.**

**Paradigm.** The *n*-back task assessed working memory performance and consisted of five, 40-trial blocks and followed similar procedures as the 2-back task described by Kane, Conway, Miura, and Colflesh (2007). Participants viewed a sequence of letters and needed to decide if each letter matched the case-insensitive letter that appeared two trials early, known as a 2-back trial (e.g., the third letter in the sequences “B-F-B” and “B-F-b” are 2-back trials). Each trial lasted for 2,500 ms in which a white letter on a black background appeared in the center of a computer screen for 500 ms followed by a 2,000-ms screen that displayed the response types. Participants needed to respond “yes” on 2-back trials and “no” on other trial types and could make a response at any point during the 2,500 ms trial. Participants made responses with the “f” and “j” keys, in which pseudorandom assignment across subjects determined the mapping between keys and response types. After each trial, a white crosshair appeared for 500 ms. Of the 40 trials in each block, eight constituted 2-back trials. Prior to the 2-back task, participants conducted a 20-trial practice block that provided feedback on performance to ensure comprehension of the instructions.

**Stimuli.** Upper and lowercase versions of the following eight letters made up the stimulus set: B, F, K, H, M, Q, R, and X. Each block consisted of a randomly generated 40-trial sequence that met the following five conditions: (a) 20% of trials are 1-back, (b) 20% of trials are 2-back, (c) 20% of trials are 3-back, (d) stimuli could not constitute both a 1-, 2-, and/or 3-back, and (e) each of the eight stimuli appeared at least three times, but no more than seven times in the sequence.

#### **Task switching.**

**Paradigm.** The task included five 40-trial blocks of a modified task-switching paradigm described by Rogers and Monsell (1995). On each trial, a randomly ordered number/letter pairing (e.g., “U2” or “2U”) appeared within one square of a 2 × 2 grid. Participants pressed the “f” and “j” keys to either respond “yes” or “no” (depending on pseudorandom assignment across subjects) to one of the two following questions: “Is the letter a vowel?” or “Is the number odd?” The stimulus remained on screen until the participant responded. Afterward a 300 ms presentation of a green crosshair or red “x” appeared to indicate a correct or incorrect response, respectively. The next trial appeared in a new square that moved in either a clockwise or counterclockwise fashion depending on pseudorandom assignment across participants. The question to be answered depended on whether the stimulus appeared in the top row or bottom row of squares (also pseudorandomly assigned across participants). Thus, the task switched every two trials.

**Stimuli.** The stimulus set consisted of four odd numbers (3, 5, 7, 9), even numbers (2, 4, 6, 8), vowels (A, E, O, U), and consonants (G, K, M, R). Letter/number pairings occurred randomly on a trial-by-trial basis under the condition that odd numbers always paired with consonants and even numbers always paired with vowels. This pattern ensured that the answer to one of the two question types is always “yes” while the other is always “no.”

#### **Questionnaires.**

**Intrinsic Motivation Inventory.** The IMI is a standardized 45-item post task questionnaire intended to assess the subjective experiences that a participant felt during a recently completed task (Ryan, 1982). To conduct the IMI, participants read a statement (e.g., “I tried very hard on this activity.”) and rate how true they believe the statement pertains to them on a scale from 1 (*not at all true*) to 7 (*very true*). Although the IMI consists of seven subscales, we specifically administered the five items from the Effort/Importance subscale to assess perceived effort and motivation during the recognition memory and visual detection criterion shifting paradigms.

**Behavioral Inhibition System and Behavioral Activation System scales.** The BIS/BAS scales are a standardized 24-item questionnaire that assesses an individual’s motivation to avoid aversive outcomes and approach desired outcomes (Carver & White, 1994). During the questionnaire participants rate how true or false a statement (e.g., “I will often do things for no other reason than that they might be fun.”) pertains to them on a scale from 1 (*very true for me*) to 4 (*very false for me*). The BIS/BAS scales consist of four subscales, but we specifically analyzed the four items from the BAS fun-seeking subscale in an attempt to replicate the finding of Aminoff and colleagues (2012) that showed the BAS fun-seeking score is positively associated with the extent of  $C_n$  during a recognition memory test.

**Positive and Negative Affect Schedule—Expanded Form.** The PANAS-X is a standardized 60-item questionnaire that assesses a person’s recent feelings and emotions (Watson & Clark, 1994). The questionnaire requires participants to read a word or phrase (e.g., “afraid”) and rate the extent to which they felt that way during the past few weeks on a scale from 1 (*very slightly or not at all*) to 5 (*extremely*). The PANAS-X consists of 13 scales, but we specifically analyzed the 10 items from the Negative Affect scale in an attempt to replicate the finding of Aminoff and colleagues (2012) that the Negative Affect scale score is negatively correlated with the extent of  $C_n$  during a recognition memory test.

**Verbalizer-Visualizer Questionnaire (modified).** The VVQ is a standardized 15-item questionnaire that assesses an individual’s preference to represent knowledge in a visual or verbal manner (Richardson, 1977). We implemented a modified version of the VVQ in which some of the 15-items derived from the Individual Differences Questionnaire (IDQ; Paivio, 1971), which is an 86-item questionnaire that formed the basis of the original VVQ (Richardson, 1977). Although the original VVQ requires individuals to respond “true” or “false” to whether a statement (e.g., “I enjoy learning new words.”) applies to the participant, we required a response as to how strongly an individual agreed or disagreed with each statement on a scale from 1 (*strongly disagree*) to 7 (*strongly agree*). This modified version of the VVQ is the same questionnaire that Aminoff and colleagues (2012) implemented (though the authors simply refer to the modified version as the

VVQ). We attempted to replicate the finding of Aminoff and colleagues (2012) that the extent of  $C_n$  during a recognition memory test is positively associated with the verbalizer score (seven items) on this modified version of the VVQ.

**Statistical analyses.** From the extraneous tasks and questionnaires, we obtained eight additional individual difference measures to assess whether various cognitive and personality factors could explain variance in criterion shifting tendencies. For the BART, we assessed risk-taking behavior by computing mean adjusted pumps: the average number of balloon pumps on trials where the participant chose to collect money (see Lejuez et al., 2002). We computed  $d'$  for the go/no-go and  $n$ -back tasks as performance indices of response inhibition and working memory, respectively. We assessed task-switching ability from the time cost measure, which is computed from the difference in reaction time (RT) when responding to switch trials (where the task of the current trial differs from the previous trial) versus same trials (where the task remained the same between the previous and current trial). To be consistent with Rogers and Monsell (1995), we excluded trials with RTs less than 100 ms and trials that immediately followed an error. An assessment of motivation to perform well on the criterion shifting tasks came from the IMI questionnaire Effort/Importance subscale (IMI: effort/importance). Although participants completed the entire BIS/BAS scales, PANAS-X, and modified VVQ surveys we specifically analyzed the three measures that Aminoff and colleagues (2012) found to be significantly related to  $C_n$  during a recognition memory test. We computed the fun-seeking score of the BIS/BAS scales (BAS: fun-seeking), the PANAS-X negative affect score (PANAS: negative), and the verbalizer score on the modified VVQ (VVQ: verbal; see the Appendix for questionnaire items and scoring details).

For each of the eight additional measures we report mean values for each session, Cohen's  $d$  effect sizes for mean differences between the two sessions, and Pearson correlation coefficients to assess the test-retest reliability of each measure. To assess whether any of the eight measures relate to individual differences in criterion shifting tendencies, we conducted Pearson  $r$  correlations between each of the eight additional individual difference measures against both recognition memory  $C_n$  and visual detection  $C_n$  during the two sessions. We assessed the strength of evidence for the null versus alternative hypotheses using Bayes factors with uninformed uniform priors. In the online supplemental materials we report a linear mixed model that tests whether the extent of  $C_n$  is affected by session number, task type, and the eight additional measures within a single model as specified in our preregistration of Experiment 3 (<https://osf.io/jkfp6>).

## Results and Discussion

Because the extent of criterion shifting can be affected by changes in  $d'$  (Macmillan & Creelman, 2005), we attempted to make mean discriminability in the recognition memory and visual detection paradigms approximately equal across both tasks and sessions. In the recognition memory task, mean  $d'$  did not significantly differ between Session 1 ( $M = 0.95$ ,  $SD = 0.53$ ) and Session 2 ( $M = 1.02$ ,  $SD = 0.57$ ),  $d = 0.14$ , 95% CI [-0.01, 0.29]. In the visual detection task, mean  $d'$  slightly increased from Session 1 ( $M = 1.12$ ,  $SD = 0.63$ ) to Session 2 ( $M = 1.34$ ,  $SD = 0.61$ ),  $d = 0.32$ , 95% CI [0.17, 0.47]. Mean  $d'$  remained higher on

average during the visual detection task relative to recognition memory for both Session 1 ( $d = 0.27$ , CI = 0.12, 0.42) and Session 2 ( $d = 0.51$ , CI = 0.36, 0.67), despite our efforts to make mean discriminability equivalent across decision domains. On the recognition memory tests participants on average shifted between  $c_n$  in the conservative ( $M = 0.72$ ,  $SD = 0.57$ ) and liberal criterion conditions ( $M = -0.76$ ,  $SD = 0.59$ ),  $d = 2.69$ , 95% CI [2.48, 2.89], as well as across visual detection tasks between the conservative ( $M = 0.84$ ,  $SD = 0.52$ ) and liberal criterion conditions ( $M = -0.43$ ,  $SD = 0.58$ ),  $d = 2.41$ , 95% CI [2.22, 2.61].

Average  $C_n$  during recognition memory did not significantly differ between Session 1 ( $M = 1.40$ ,  $SD = 0.45$ ) and Session 2 ( $M = 1.56$ ,  $SD = 0.50$ ),  $d = 0.18$ , 95% CI [-0.03, 0.39], as well as during visual detection between Session 1 ( $M = 1.22$ ,  $SD = 0.55$ ) and Session 2 ( $M = 1.31$ ,  $SD = 0.43$ ),  $d = 0.11$ , 95% CI [-0.10, 0.32]. Mean  $C_n$  remained marginally higher during recognition memory compared with visual detection in Session 1 ( $d = 0.21$ , 95% CI [0.00, 0.42]) and Session 2 ( $d = 0.29$ , 95% CI [0.08, 0.50]).

The extent of  $C_n$  from Session 1 to Session 2 remained very consistent for both the recognition memory,  $r(170) = .75$ , 95% CI [.67, .80], and visual detection,  $r(170) = .65$ , 95% CI [.55, .73] tests. Strong relationships in  $C_n$  also persisted across the two tasks during Session 1,  $r(170) = .68$ , 95% CI [.59, .75] and Session 2,  $r(170) = .78$ , 95% CI [.72, .83], despite small differences in mean  $d'$  and  $C_n$  across decision domains. Correlations even remained strong when comparing  $C_n$  in Session 1 of the recognition memory test to Session 2 of the visual detection test,  $r(170) = .65$ , 95% CI [.55, .73] and vice versa,  $r(170) = .57$ , 95% CI [.46, .67]. Although  $C_n$  remained strongly consistent across tasks,  $d'$  only showed weak correlations between the two tasks during Session 1,  $r(170) = .14$ , 95% CI [-.01, .29];  $BF_{01} = 1.03$  and Session 2,  $r(170) = .17$ , 95% CI [.02, .31];  $BF_{01} = 0.57$ . This provides evidence that the cross-task stability of criterion shifting cannot simply be attributed to discriminability performance alone. Overall, the strong correlations in  $C_n$  across sessions and tasks suggest that criterion shifting is a stable, domain-general process. Figure 8 displays conservative and liberal  $c_n$  values for each participant across both sessions and tasks, ordered from left to right based on who shifted criteria the most during Session 2 of the recognition memory task.

For the eight additional task and questionnaire measures, we report mean performance during both sessions, a Cohen's  $d$  effect size measure for mean differences across sessions, and test-retest Pearson  $r$  correlation coefficients in Table 3. It is possible that previously reported weak correlations between the extent of criterion shifting and other task measures could be a result of low test-retest reliability of the other measures. In our study the test-retest reliability remained moderately strong for most of the additional measures,  $r(170)$  range: .47-.83, but we caution that imperfect reliability could still attribute to attenuation in the relationships between these measures and the extent of  $C_n$ .

To test whether criterion shifting is related to the eight additional task and questionnaire measures, we conducted Pearson correlations between each measure and  $C_n$  in both the recognition memory and visual detection tasks across both sessions (Table 4). No statistically significant relationships existed after FDR correction. We furthered these null findings by computing Bayes factors to assess the amount of support for the null versus alternative hypotheses ( $BF_{01}$ ) for each

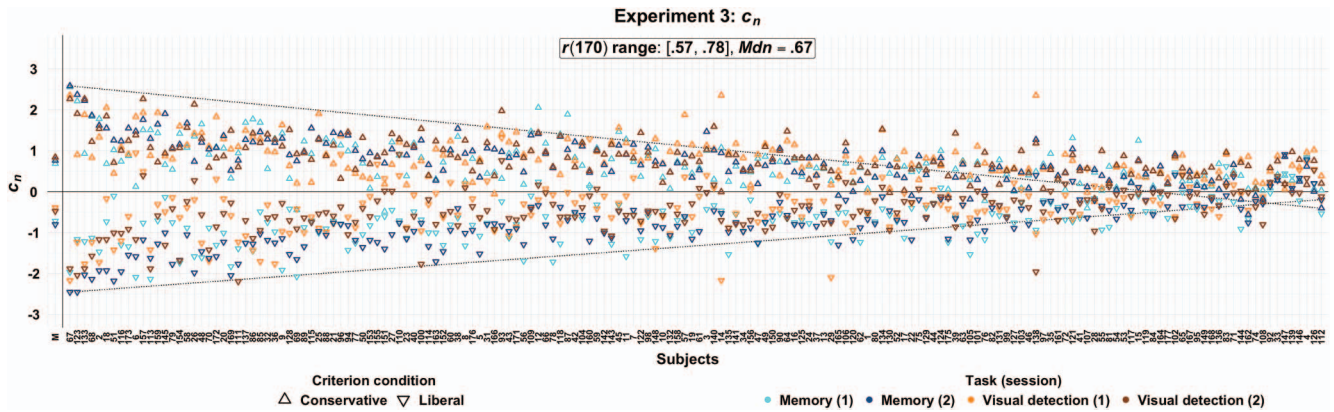


Figure 8. Experiment 3 normalized criterion placement ( $c_n$ ) values for each participant and the mean ( $M$ ) in the conservative and liberal criterion conditions for the recognition memory and visual detection tests across both sessions. The extent of criterion shifting is depicted by the distance between the triangles representing the conservative and liberal  $c_n$  values. Participants are ordered from left to right based on who shifted criteria the most during Session 2 of the recognition memory test. The dotted lines emphasize this criterion shifting trend by connecting the Session 2 recognition memory conservative and liberal  $c_n$  values of the leftmost and rightmost subjects. The range and median Pearson correlation for normalized criterion shifting ( $C_n$ ) across sessions and tasks (six measures total) is shown below the graph title. See the online article for the color version of this figure.

comparison. Of the 32 comparisons, 28 showed greater than three times support for the null versus alternative hypothesis ( $BF_{01} > 3$ ), which is considered strong evidence for the null hypothesis (Jeffreys, 1961). Two of the other four comparisons include negative relationships during Session 2 between recognition memory  $C_n$  and the BAS fun-seeking score,  $r(170) = -.16$ , 95% CI  $[-.30, -.01]$ ;  $BF_{01} = 1.10$  and modified VVQ verbalizer score,  $r(170) = -.14$ , 95% CI  $[-.28, .01]$ ;  $BF_{01} = 1.94$ . Both of these findings are in opposition to the positive relationships observed by Aminoff and colleagues (2012), suggesting that a fun-seeking personality and verbal cognitive style are not actually predictive of criterion shifting tendencies. The only comparison to show support for the alternative hypothesis is a negative relationship between recognition memory  $C_n$  and the IMI Effort/Importance subscale score during Session 2,  $r(170) = -.18$ , 95% CI  $[-.33, -.04]$ ;  $BF_{01} = 0.57$ . Interestingly, a negative relationship between the IMI Effort/Importance subscale score and visual detec-

tion  $C_n$  during Session 2 only slightly supported the null hypothesis,  $r(170) = -.16$ , 95% CI  $[-.30, -.01]$ ;  $BF_{01} = 1.24$ . However, relationships between the IMI Effort/Importance subscale score and  $C_n$  during Session 1 strongly supported the null hypothesis for both the recognition memory,  $r(170) = -.06$ , 95% CI  $[-.20, .09]$ ;  $BF_{01} = 7.39$  and visual detection,  $r(170) = -.06$ , 95% CI  $[-.21, .09]$ ;  $BF_{01} = 10.40$  tests, suggesting that there is not a consistent relationship between motivation to perform well during the criterion shifting tasks and the extent of criterion shifting itself. Taken together, these standardized measures of risk aversion, response inhibition, working memory, task-switching, motivational effort, and personality attributes cannot explain the vast individual differences in the extent of criterion shifting.

Results from Experiment 3 revealed that strategic criterion shifting tendencies over time are a stable, domain general process.

Table 3  
Experiment 3 Session 1 and 2 Means With Standard Deviation Values (in Parentheses), Cohen's  $d$  Effect Sizes of Mean Differences Between Sessions 1 and 2 With 95% CIs (in Brackets), and Test-Retest Pearson Correlation Coefficients With 95% CIs (in Brackets)

Measure	Session 1	Session 2	Cohen's $d$	Pearson $r$
Recognition memory $C_n$	1.40 (0.45)	1.56 (0.50)	0.18 [-0.03, 0.39]	.75 [.67, .81]
Visual detection $C_n$	1.22 (0.55)	1.31 (0.43)	0.11 [-0.10, 0.32]	.65 [.55, .73]
BART: adjusted pumps	34.72 (8.31)	40.83 (8.31)	0.42 [0.20, 0.63]	.68 [.59, .75]
Go/no-go: $d'$	3.48 (0.51)	3.25 (0.51)	0.32 [0.10, 0.53]	.47 [.34, .58]
$N$ -back: $d'$	1.66 (0.54)	2.03 (0.54)	0.35 [0.13, 0.56]	.76 [.69, .82]
Task switching: time cost	377 ms (121)	260 ms (121)	0.60 [0.39, 0.82]	.61 [.51, .70]
IMI: effort/importance	29.59 (3.12)	28.83 (3.12)	0.15 [-0.06, 0.37]	.63 [.53, .71]
BAS: fun-seeking	12.60 (1.40)	12.14 (1.40)	0.20 [-0.01, 0.41]	.63 [.54, .72]
PANAS: negative	21.18 (3.84)	20.99 (3.84)	0.03 [-0.19, 0.24]	.74 [.67, .80]
VVQ: verbal	30.40 (2.54)	30.85 (2.54)	0.07 [-0.14, 0.29]	.83 [.78, .87]

Note.  $C_n$  = normalized criterion shifting;  $d'$  = discriminability; BART = Balloon Analogue Risk Task; IMI = Intrinsic Motivation Inventory; BAS = Behavioral Activation System; PANAS = Positive and Negative Affect Schedule; VVQ = Verbalizer-Visualizer Questionnaire.



Table 4

Experiment 3 Pearson Correlation Coefficients and 95% CIs (in Brackets) Between  $C_n$  in Both Sessions and Tasks Against the Eight Additional Task and Questionnaire Measures

Measure vs. $C_n$	Recognition memory		Visual detection	
	Session 1	Session 2	Session 1	Session 2
BART: adjusted pumps	-.08 [-.22, .07] (6.32)	.00 [-.15, .15] (10.47)	.03 [-.12, .18] (9.78)	-.05 [-.20, .10] (8.23)
Go/no-go: $d'$	-.07 [-.22, .08] (6.83)	-.03 [-.18, .12] (9.51)	-.11 [-.26, .04] (3.77)	-.08 [-.22, .08] (6.50)
N-back: $d'$	-.01 [-.16, .14] (10.44)	.01 [-.13, .16] (10.28)	-.01 [-.16, .14] (10.40)	.01 [-.14, .16] (10.40)
Task switching: time cost	.09 [-.06, .23] (5.52)	.02 [-.13, .17] (9.98)	.03 [-.12, .18] (9.61)	-.02 [-.16, .13] (10.28)
IMI: effort/importance	-.06 [-.21, .09] (7.39)	<b>-.18 [-.33, -.04] (0.57)</b>	-.01 [-.16, .14] (10.40)	<b>-.16 [-.30, -.01] (1.24)</b>
BAS: fun-seeking	-.06 [-.20, .09] (8.06)	<b>-.16 [-.30, -.01] (1.10)</b>	-.06 [-.21, .09] (7.93)	-.06 [-.21, .09] (7.68)
PANAS: negative	.03 [-.12, .18] (9.47)	.01 [-.14, .15] (10.45)	.08 [-.07, .23] (5.84)	-.04 [-.19, .11] (9.21)
VVQ: verbal	-.05 [-.20, .10] (8.22)	<b>-.14 [-.28, .01] (1.94)</b>	.01 [-.14, .16] (10.43)	-.05 [-.20, .10] (8.30)

Note.  $C_n$  = normalized criterion shifting;  $d'$  = discriminability; BART = Balloon Analogue Risk Task; IMI = Intrinsic Motivation Inventory; BAS = Behavioral Activation System; PANAS = Positive and Negative Affect Schedule; VVQ = Verbalizer-Visualizer Questionnaire. Bayes factors supporting the null versus alternative hypotheses ( $BF_{01}$ ) are presented in parentheses next to each correlation coefficient.  $BF_{01}$  scores greater than 3 are considered strongly supportive of the null hypothesis (Jeffreys, 1961). Values that include  $BF_{01}$  values below 3 are in bold, although no significant relationships existed after FDR correction.

Individuals who consistently shift criteria to large extents during recognition memory tests also regularly shift criteria to large degrees during visual detection tests. Relationships in discriminability across the two decision domains remained weak indicating that the cross-task stability in performance is specific to criterion shifting and not simply a function of overall discriminability performance. Measures from other tasks and questionnaires could not explain individual differences in criterion shifting tendencies. It is possible that other personality or cognitive characteristics not tested in Experiment 3 or by Aminoff and colleagues (2012) are associated with the extent of criterion shifting, but currently no obvious relationships are known. These findings indicate that strategic criterion shifting tendencies are a uniquely individualistic cognitive trait.

#### Experiment 4

Findings from Experiment 3 revealed that criterion shifting tendencies are stable within and across decision domains without any obvious relationship to certain cognitive abilities or a motivation to perform well on the criterion shifting tasks. In Experiment 4, we further these findings by assessing whether strategic criterion shifting strategies during recognition memory tests are related to an individual's meta awareness of the memory strength elicited by test items as measured by metacognitive bias and metacognitive sensitivity via confidence ratings (Fleming & Lau, 2014). Confidence ratings are typically implemented in recognition memory studies to assess criterion shifts (Macmillan & Creelman, 2005; Yonelinas & Parks, 2007), making it reasonable to predict that strategic criterion shifting and decision strategies for reporting confidence are strongly related. In the first three experiments, we created paradigms that required extreme criterion shifts to maximize payoffs (Experiments 1 and 3) or accuracy (Experiment 2). In these extreme situations, people should only choose the riskier option (i.e., respond "old" when a conservative criterion is advantageous or "new" when a liberal criterion is propitious) when they have high confidence that it is the correct choice. If people use meta awareness of the familiarity strength elicited by test items to strategically shift criteria, then criterion shifting tendencies should

be strongly related to how often people report high confidence in "old" and "new" responses (i.e., metacognitive bias). However, some people may be more adept than others at discerning between different levels of familiarity strength (i.e., have higher metacognitive sensitivity), which may allow them to shift criteria to greater extents because they can better differentiate between stronger and weaker memory evidence. We therefore assessed whether metacognitive bias or metacognitive sensitivity could explain individual differences in criterion shifting tendencies by comparing performance on recognition memory tests that either required strategic criterion shifting or confidence ratings.

Participants on average tend to have high metacognitive sensitivity, demonstrating that confidence ratings scale well with the accuracy of old/new judgments (Mickes et al., 2007, 2011). However, Mickes and colleagues (2011) showed that there are individual differences in metacognitive bias, particularly when scaling strong memories. Some people use extreme criteria for making old/new judgments with the highest level of confidence, whereas others are less judicious. Metacognitive bias may be related to individual criterion shifting tendencies if people shift criteria based on the level of confidence in a recognition judgment. However, people may implement completely different decision strategies when rating confidence versus strategically shifting a criterion. For instance, a criterion shift may reflect an individual's willingness to make a strategic old/new response before making a recognition judgment whereas a confidence rating could represent an assessment of an old/new response after the judgment is made. Mickes and colleagues (2017) provide some evidence that people employ different decision strategies for rating confidence versus shifting criteria, because individuals tended to establish more conservative criteria when making high confident "old" judgments on a multi-point scale compared with binary decisions for responding "old" when specifically instructed to only do so when there is 100% confidence. Miller and Kantner (2020) failed to find any significant relationships between metacognitive bias and the extent of criterion shifting during recognition memory tests when conducting post hoc analyses on previously reported data. This alludes to the possibility that metacognitive bias might be a poor indicator of

strategic criterion shifting tendencies, but no studies have systematically compared this relationship during recognition memory tests a priori.

In Experiment 4, participants conducted two different recognition memory tests on separate days that required either a confidence judgment on a 6-point scale (confidence ratings session) or a binary old/new response (binary response session). During the confidence ratings session, participants responded to each test image on a 6-point scale ranging from high confidence “new” to high confidence “old.” In the binary response session, participants conducted the exact same task except instead of responding on a six-point scale, participants received instructions to only respond “old” with “high confidence” (otherwise respond “new”) in the conservative criterion condition or only respond “new” with “high confidence” (otherwise respond “old”) in the liberal criterion condition. This made the instructions as similar as possible across the two tasks and gave participants the best opportunity for establishing the same criteria for high confident responses, regardless of whether participants responded on a six-point scale or made binary old/new judgments.

Given the results of Mickes and colleagues (2017) and Miller and Kantner (2020), we predicted no relationship between metacognitive bias and strategic criterion shifting. We also predicted that individuals would establish more extreme criteria when rating high confidence on a multipoint scale relative to the extent of criterion shifting to explicit instructions. Such a finding would indicate that individuals who do not adequately shift criteria are at least *capable* of shifting criteria to greater extents, but might simply be *unwilling* to do so. We additionally predicted no relationship between metacognitive sensitivity and the extent of criterion shifting because we do not believe meta awareness of memory strength affects an individual’s willingness to shift a criterion.

**Method**

**Participants.** One hundred seventy participants (45 males;  $M = 19.4$  years, range = 18–34, years,  $SD = 1.8$ ) completed both sessions on separate days within the same week. Three additional participants only completed one of the two sessions and are excluded from all analyses. Participants earned \$10 for completing each session.

Although we predicted no relationship between the extent of criterion shifting and metacognitive bias or metacognitive sensitivity, we wanted to ensure that we could identify a modest effect as statistically significant if a relationship did exist. An a priori power analysis revealed that data collection on 123 participants provides 80% power for detecting a Pearson correlation of  $r = .25$ , which we felt would be a non-negligible relationship that could help explain individual criterion shifting tendencies. Initially, all participants conducted the confidence ratings session first because we felt that presenting the tasks in this order gave participants the best opportunity to implement similar decision strategies for rating confidence and strategically shifting criteria. However, an unexpected relationship did exist which prompted us to collect additional participants who conducted the two tasks in the reverse order, to rule out potential order effects. A second a priori power analysis revealed that 46 participants provides 95% power to find an effect of  $r = .50$ , a value we derived from the initial sample (see Results and Discussion).

**Procedure.** The two self-paced sessions lasted for 20–45 min where participants conducted four cycles of studying 75 unique face stimuli followed by three, 50-image test mini blocks (Figure 9). In one of the two test sessions, participants made recognition judgments on a six-point confidence scale (high confidence new, medium confidence new, low confidence new, low confidence old, medium confidence old, or high confidence old) for all test blocks (the confidence ratings session). The other session required a binary old/new judgment (the binary response session), but under

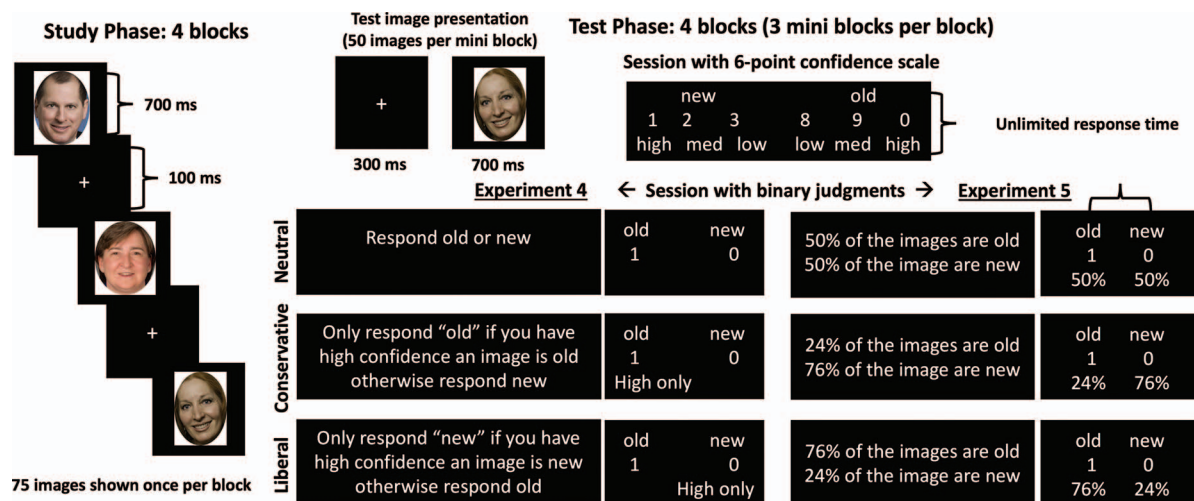


Figure 9. Recognition memory task for Experiments 4 and 5. Participants conducted two task sessions on separate days. In one session participants made confidence judgments on a 6-point scale. The other session required making binary old/new judgments with a conservative, liberal, or neutral criterion manipulation. In Experiment 4, participants received instructions to only respond old or new with “high confidence” in the conservative and liberal criterion conditions, respectively, whereas a base rate manipulation induced criterion shifts in Experiment 5. See the online article for the color version of this figure.

three different conditions. Participants received instructions to either (a) simply respond “old” or “new” (neutral condition), (b) only respond “old” when there is high confidence an image is old (conservative condition), or (c) only respond “new” when there is high confidence an image is new (liberal condition).

In the study phase, participants passively viewed a sequence of images in the center of a computer screen on a black background for 700 ms followed by a 100-ms presentation of a white crosshair. After viewing a test image for 700 ms, participants made a response with unlimited time, followed by a 300-ms crosshair presentation. During confidence ratings sessions, participants made high confidence responses using the “1” and “9” keys, medium confidence responses with “2” and “8” keys, and low confidence responses with the “3” and “7” keys. The familiarity strength corresponding to the response types either increased or decreased from left-to-right depending on the pseudorandom assignment (i.e., high confidence “new” to high confidence “old,” or vice versa). The response screen displayed all six keys with the corresponding response values. For binary response sessions, participants made old/new judgments via the “1” or “0” keys where a response screen reminded participants of the mapping between keys and response types. The conservative and liberal criterion conditions included the phrase “high confidence only” below the old or new response text, respectively. The stimuli included 1,200 unique face images (600 per session) and each participant received a completely randomized assignment and presentation order for the target and lure images.

**Statistical analysis.** Although traditional measures of metacognitive bias do not account for accuracy (e.g., the percentage of “high confidence” responses made throughout an entire recognition test regardless if responses are correct or not; see Fleming & Lau, 2014), we wanted to make direct comparisons between strategic criterion shifts and metacognitive bias. We therefore classified different confidence levels as decision criteria and computed conservative, liberal, and neutral  $c_n$  by treating different confidence ratings as binary old/new responses. Computations of a “conservative”  $c_n$  occurred by only treating high confident old responses as “old” and all other response types as “new” (i.e., the criterion for “high confidence old” responses). Calculations for a “liberal”  $c_n$  occurred by only treating high confident new responses as “new” and all other response types as “old” (i.e., the criterion for all responses except “high confidence new”). This allowed us to compute a measure of “metacognitive  $C_n$ ” (i.e., “conservative”  $c_n$  minus “liberal”  $c_n$ , which is the measure used by Miller & Kantner, 2020) for the confidence ratings session that is computed the same way as  $C_n$  in the binary response session. For measures of “neutral”  $c_n$ , we considered all old responses as “old” and all new responses as “new,” regardless of the confidence level assigned to each response.

To measure metacognitive sensitivity, we computed the area under the type-2 receiver operator characteristic (AUROC2) curve, which is more robust to influences of decision bias compared with other common measures, such as type-2  $d'$  (an analogous measure to  $d'$ ; see Fleming & Lau, 2014).

## Results and Discussion

As specified in our preregistration (<https://osf.io/4wnjm>), we initially collected a dataset of 122 subjects where all participants

first conducted the confidence ratings session. Because we did not predict a relationship in  $C_n$  between the two tasks, we wanted participants to familiarize themselves with the confidence ratings structure so that decision strategies for identifying old and new items with “high confidence only” could easily be transferred to the binary response session. However, an unexpected relationship existed in  $C_n$  across the two tasks,  $r(120) = .60$ , 95% CI [.47, .70]. To test for potential order effects, we collected data from an additional 48 participants who conducted the two sessions in the reverse order (as specified in a subsequent preregistration: <https://osf.io/ae2rp>), but a modest relationship in  $C_n$  persisted,  $r(46) = .39$ , 95% CI [.12, .61]. Because the relationship in  $C_n$  across the two tasks could not be completely attributed to an order effect, we combined data from all 170 participants for subsequent analyses (see the online supplemental materials for separate analyses of the two groups).

Mean  $d'$  did not significantly differ between the confidence ratings session ( $M = 0.56$ ,  $SD = 0.26$ ) and the neutral criterion condition of the binary response session ( $M = 0.56$ ,  $SD = 0.22$ ),  $d = 0.00$ , 95% CI [−0.21, 0.22]. However, a strong relationship in  $d'$  existed between the two sessions,  $r(168) = .49$ , 95% CI [.37, .60]. Similarly,  $c_n$  did not significantly differ between the confidence ratings session ( $M = 0.21$ ,  $SD = 0.24$ ) and the neutral criterion condition of the binary response session ( $M = 0.27$ ,  $SD = 0.19$ ),  $d = 0.15$ , 95% CI [−0.06, 0.37], and a strong relationship in  $c_n$  existed between the two sessions,  $r(168) = .62$ , 95% CI [.52, .71]. This confirms that the different test instructions across the two tasks did not substantially affect performance in regards to discriminability and neutral criterion placement. As in Experiments 1 and 2, no significant relationship existed in the binary response session between  $C_n$  and neutral  $c_n$ ,  $r(168) = .08$ , 95% CI [−.08, .22];  $BF_{01} = 3.52$  or between  $C_n$  and the absolute value of neutral  $c_n$ ,  $r(168) = .11$ , 95% CI [−.05, .25];  $BF_{01} = 2.24$ , providing more evidence that placing and shifting a criterion are independent decision processes.

Unexpectedly, a strong relationship existed in  $C_n$ ,  $r(168) = .53$ , 95% CI [.41, .63] between the two tasks, suggesting that metacognitive bias is predictive of the extent of criterion shifting in this paradigm. In the confidence ratings session, participants on average drastically shifted between the “conservative”  $c_n$  ( $M = 1.28$ ,  $SD = 0.54$ ) and “liberal”  $c_n$  values ( $M = -1.21$ ,  $SD = 0.69$ ),  $d = 3.97$ , 95% CI [3.60, 4.34]. Participants shifted criteria to a large extent in the binary response session between  $c_n$  in the conservative ( $M = 0.72$ ,  $SD = 0.40$ ) and liberal criterion conditions ( $M = -0.27$ ,  $SD = 0.47$ ),  $d = 2.00$ , 95% CI [1.73, 2.26]. However, metacognitive  $C_n$  ( $M = 2.49$ ,  $SD = 0.65$ ) proved to be much greater than  $C_n$  when making binary old/new judgments ( $M = 0.99$ ,  $SD = 0.65$ ),  $d = 1.64$ , 95% CI [1.40, 1.89]. Even though the instructions for reporting high confidence remained similar across the two tasks, virtually all participants established much more extreme criteria for high confident responses when asked to report on a 6-point scale. Figure 10 illustrates individual differences in conservative and liberal  $c_n$  values across the two sessions in order from left to right based on the largest to smallest metacognitive  $C_n$  value. Finally, we assessed whether differences in metacognitive sensitivity in the confidence ratings session could predict the extent of  $C_n$  during the binary response session. However, no relationship existed between AUROC2 ( $M = .57$ ,  $SD =$



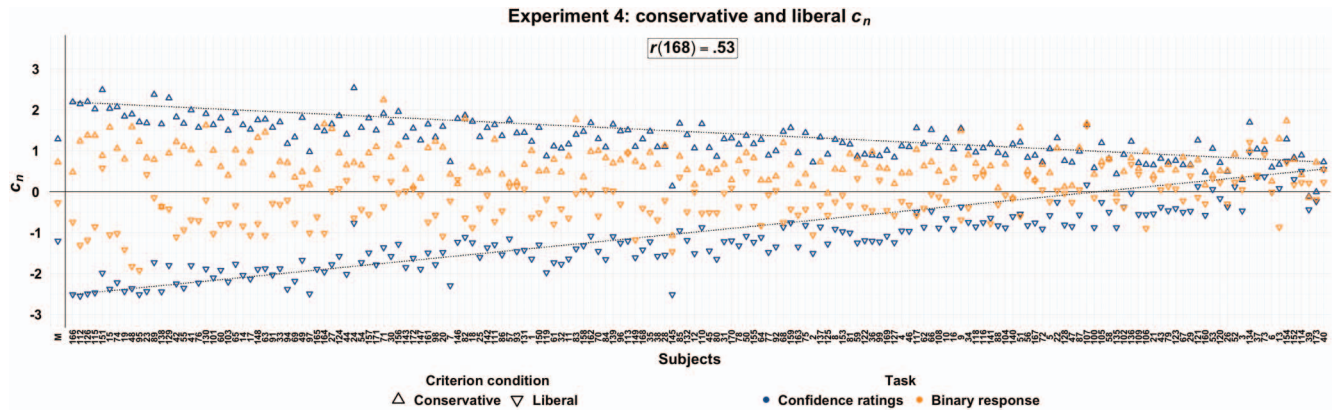


Figure 10. Experiment 4 normalized criterion placement ( $c_n$ ) values for each participant and the mean ( $M$ ) in the conservative and liberal criterion conditions of the binary response session (orange (light grey)) and “conservative” and “liberal”  $c_n$  values computed from the confidence ratings session (blue (dark grey)). The extent of criterion shifting is depicted by the distance between the triangles representing the conservative and liberal  $c_n$  values. Participants are ordered from left to right based on the largest to smallest metacognitive normalized criterion shifting ( $C_n$ ) value. The dotted lines emphasize individual differences in metacognitive  $C_n$  by connecting the “conservative” and “liberal”  $c_n$  values from the confidence rating sessions of the leftmost and rightmost subjects. See the online article for the color version of this figure.

.04) in the confidence ratings session and  $C_n$  in the binary response session,  $r(168) = .07$ , 95% CI  $[-.08, .22]$ ;  $BF_{01} = 7.17$ .

Experiment 4 revealed that participants generally establish much more extreme criteria for high confident “old” and “new” responses when reporting on a six-point scale versus making binary old/new judgments. However, a strong relationship existed between  $C_n$  in the binary response session and metacognitive  $C_n$  in the confidence ratings session. Individuals who maintain extreme criteria for the highest levels of confidence on a six-point scale also shifted criteria to a large extent (though to a much lesser degree), while those who established less extreme criteria for high confident responses also shifted criteria to a smaller extent. This indicates that individuals may implement similar decision strategies for making confidence judgments and strategically shifting a criterion, which is contradictory to the findings of Miller and Kantner (2020). However, Miller and Kantner (2020) examined data that included either a base rate or payoff manipulation, which does not cue participants to respond based on a level of confidence. Because the criterion manipulation in Experiment 4 involved instructions to respond based on confidence levels, participants may have treated the criterion manipulation as a type of confidence judgment (i.e., a response on a 2-point confidence scale). It is possible that individuals only use similar decision processes for strategic criterion shifting and metacognitive bias when the criterion manipulation explicitly instructs participants to respond based on confidence levels.

### Experiment 5

Experiment 4 showed a strong relationship between the degree to which individuals use “high confidence” judgments on a six-point scale versus shifting criteria to instructions that require responding “old” or “new” with “high confidence only.” Because this result contradicts the findings of Miller and Kantner (2020), we wanted to test whether the extent of criterion shifting in

response to a criterion manipulation without reference to confidence levels would also be related to metacognitive bias. We therefore changed the instruction manipulation in Experiment 4 to a base rate manipulation in Experiment 5. Again, we predicted no relationship between the extent of criterion shifting and metacognitive bias (see our preregistration: <https://osf.io/tqc42>).

### Method

**Procedure.** The procedures of Experiment 5 matched those of Experiment 4 except the binary response session induced criterion shifts with a base rate manipulation instead of instructions. Participants received information prior to each test block about the likelihood of encountering old and new items. An old item appeared either 24% (conservative criterion condition), 50% (neutral criterion condition), or 76% (liberal criterion condition) of the time during a test block (Figure 9). Text appeared below each response type to indicate the likelihood of encountering an old or new image during a test block. We counterbalanced the session order across participants, and the stimuli included 1,200 unique face images (600 per session) which differed from the stimulus set of Experiment 4.

**Participants.** One hundred twenty-nine participants (37 males;  $M = 22.0$  years, range = 18–48, years,  $SD = 4.5$ ) successfully completed both sessions within a week. Four additional participants failed to complete both sessions and are excluded from all analyses. Participants received \$10 for completing each session.

### Results and Discussion

Similar to Experiment 4, no significant differences existed in mean  $d'$  between the confidence ratings session ( $M = 0.58$ ,  $SD = 0.17$ ) and the neutral criterion condition in the binary response session ( $M = 0.56$ ,  $SD = 0.18$ ),  $d = 0.08$ , 95% CI  $[-0.16, 0.33]$ , and a strong relationship in  $d'$  existed between the two tasks,

$r(127) = .61$ , 95% CI [.48, .70]. The  $c_n$  values on the recognition tests also showed no significant differences between the confidence ratings session ( $M = 0.18$ ,  $SD = 0.18$ ) and the neutral criterion condition in the binary response session ( $M = 0.15$ ,  $SD = 0.11$ ),  $d = 0.10$ , 95% CI [-0.15, 0.34], while showing a strong relationship between sessions,  $r(127) = .64$ , 95% CI [.52, .73]. As in Experiments 1, 2, and 4, no significant relationship existed in the binary response session between  $C_n$  and neutral  $c_n$ ,  $r(127) = -.14$ , 95% CI [-0.31, .03];  $BF_{01} = 1.44$ . We unexpectedly found a significant negative relationship between  $C_n$  and the absolute value of neutral  $c_n$ ,  $r(127) = -.24$ , 95% CI [-0.40, -0.17];  $BF_{01} = 0.14$ . However, we believe this is a spurious finding given the relatively small effect size and the fact that this is the only significant relationship observed between the two measures across Experiments 1, 2, 4, and 5.

In the confidence ratings session, participants on average dramatically shifted between “conservative”  $c_n$  ( $M = 1.26$ ,  $SD = 0.53$ ) and “liberal”  $c_n$  values ( $M = -1.24$ ,  $SD = 0.74$ ),  $d = 3.92$ , 95% CI [3.50, 4.34]. Participants also shifted criteria in the binary response session on average between  $c_n$  in the conservative ( $M = 0.42$ ,  $SD = 0.21$ ) and liberal criterion conditions ( $M = -0.06$ ,  $SD = 0.24$ ),  $d = 1.68$ , 95% CI [1.39, 1.96]. Unlike Experiment 4, no relationship existed between metacognitive  $C_n$  ( $M = 2.50$ ,  $SD = 0.82$ ) and  $C_n$  in the binary response session ( $M = 0.48$ ,  $SD = 0.82$ ),  $r(127) = -.06$ , 95% CI [-0.23, .11];  $BF_{01} = 3.87$ . This suggests that strategic criterion shifting tendencies are unrelated to metacognitive bias in this paradigm. Similar to Experiment 4, a large mean difference in  $C_n$  existed between the two tasks,  $d = 2.51$ , 95% CI [2.18, 2.84]. Metacognitive sensitivity as measured by AUROC2 ( $M = .57$ ,  $SD = .04$ ) in the confidence ratings session showed no relationship with  $C_n$  in the binary response session,  $r(127) = .03$ , 95% CI [-0.14, .21];  $BF_{01} = 8.44$ . Figure 11 displays individual differences in conservative and liberal  $c_n$  values across the two sessions in order from left to right based on the largest to smallest metacognitive  $C_n$ .

Experiment 5 revealed that individual criterion shifting tendencies in response to a base rate manipulation are unrelated to individual differences in metacognitive bias during recognition memory tests. This finding is in line with those of Miller and Kantner (2020), who also found no relationship between criterion shifting and metacognitive bias from post hoc data analyses. The extent of strategic criterion shifting only appears to relate to metacognitive bias when the criterion manipulation specifically requires a response based on a level of confidence. When a criterion manipulation does not include instructions to respond based on confidence, individuals seem to implement different decision strategies for rating judgments with high confidence and strategically shifting a criterion.

## General Discussion

The tendency to strategically shift criteria should be considered a stable cognitive trait if it (a) shows strong test–retest reliability, (b) generalizes across tasks, and (c) cannot be explained by other cognitive factors. We demonstrated that criterion shifting during recognition memory is quite stable across many testing sessions and is as strong as the stability of criterion placement—a stable cognitive trait (Kantner & Lindsay, 2012, 2014). The test–retest reliability of criterion shifting in our studies is even comparable to measures believed to reflect stable traits in other cognitive domains. For example, Xu and colleagues (2018) administered test–retest working memory tasks across 30 days and determined that the strong consistency in performance indicates that visual working memory capacity is a stable trait. The high session-to-session correlation coefficients observed by Xu and colleagues (2018),  $r(77)$  range: .64–.86,  $Mdn = .77$ , are as strong as the session-to-session correlations we obtained for  $C_n$  during recognition memory tests in Experiment 1,  $r(37)$  range: .58–.85,  $Mdn = .75$ , Experiment 2,  $r(37)$  range: .71–.89,  $Mdn = .83$ , and Experiment 3,  $r(170) = .75$ . It should be noted, however, that we purposely

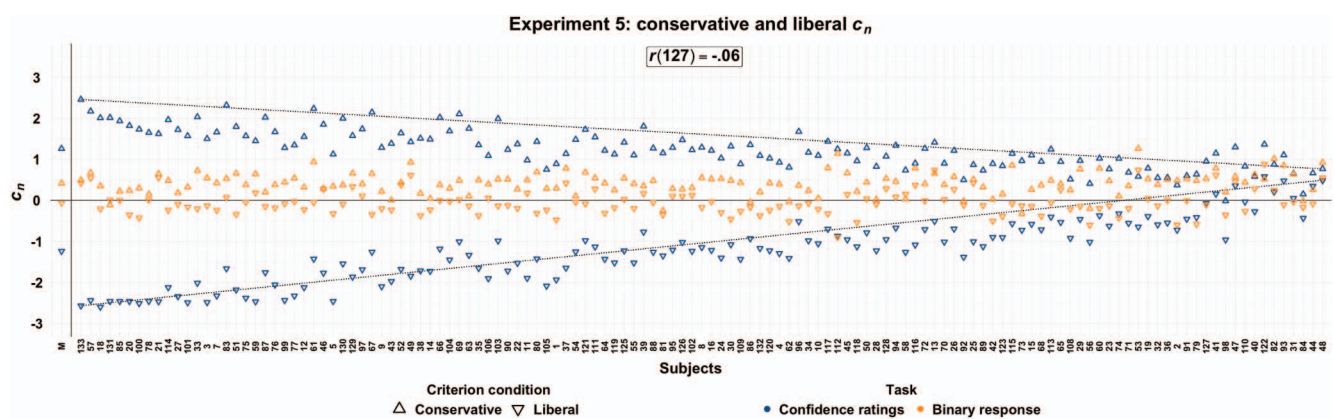


Figure 11. Experiment 5 normalized criterion placement ( $c_n$ ) values for each participant and the mean ( $M$ ) in the conservative and liberal criterion conditions of the binary response tests (orange [light grey]) and “conservative” and “liberal”  $c_n$  values computed from the confidence ratings session (blue [dark grey]). The extent of criterion shifting is depicted by the distance between the triangles representing the conservative and liberal  $c_n$  values. Participants are ordered from left to right based on the largest to smallest metacognitive normalized criterion shifting ( $C_n$ ) value. The dotted lines emphasize individual differences in metacognitive  $C_n$  by connecting the “conservative” and “liberal”  $c_n$  values from the confidence rating sessions of the leftmost and rightmost subjects. See the online article for the color version of this figure.

implemented tests with low discriminability and fairly extreme criterion manipulations. These situations require some of the largest criterion shifts to maximize payoffs (Experiments 1 and 3) or accuracy (Experiment 2), and it is possible that the stability of criterion shifting is weaker when discriminability is much higher and/or the criterion manipulations are less extreme (e.g., when targets or nontargets appear 55% as opposed to 75% of the time). Nevertheless, we believe the strong test–retest stability of criterion shifting during recognition memory tests that encourage large criterion shifts, qualifies as a trait-like feature.

The stability of individual tendencies to strategically criterion shift is not limited to a single task, but generalizes across stimuli sets (Aminoff et al., 2012), bias manipulations (Frithsen et al., 2018; Kantner et al., 2015), and decision domains (Frithsen et al., 2018). In Experiment 3, we furthered these findings by illustrating that the test–retest reliability of  $C_n$  extends across decision domains during recognition memory and visual detection tests,  $r(170)$  range: .57–.78,  $Mdn = .67$ . Importantly, the relationship in  $d'$  between the two decision domains remained weak,  $r(170)$  range: .14–.17,  $Mdn = .16$ , showing that the cross-task stability in performance is specific to criterion shifting and not discriminability. Although there are occasional inconsistencies in the cross-task stability of criterion shifting (Franks & Hicks, 2016; Frithsen et al., 2018), it appears that weak relationships occur when there are large disparities in demand characteristics (e.g., when there are differences in the experimental designs that may affect an individual's decision strategy). However, when demand characteristics are carefully controlled for, the stability of criterion shifting across tasks and decision domains is generally quite strong, suggesting that differing task designs may lead to occasional inconsistencies in criterion shifting stability and not simply the decision domains themselves (as suggested by Franks & Hicks, 2016). Future research needs to investigate the underlying factors that can lead to differing demand characteristics and why such differences may sometimes affect the stability of criterion shifting, but strategic criterion shifting tendencies appear to be a domain general process.

Although the extent to which individuals strategically shift criteria proved stable over time and decision domains, we tested whether these stable individual differences reflect an epiphenomenon of other cognitive or personality traits. Aminoff and colleagues (2012) first attempted to identify traits associated with individual criterion shifting tendencies during recognition memory tests by correlating the extent of  $C_n$  with many standardized measures of cognitive and personality characteristics. Despite collecting over 25 cognitive and personality measures, Aminoff and colleagues (2012) found that the extent of  $C_n$  only significantly related to one cognitive measure (a positive relationship with the modified VVQ verbal score) and two personality measures (a positive relationship with the BAS fun-seeking score and a negative relationship with the PANAS-X negative affect score). However, in Experiment 3 we failed to replicate these findings as we found no relationship between the extent of  $C_n$  on recognition memory and visual detection tests with the modified VVQ verbal score, BAS fun-seeking score, or PANAS-X negative affect score. We expanded on Aminoff and colleagues (2012) efforts to identify characteristics associated with individual criterion shifting tendencies by assessing relationships between the extent of  $C_n$  and performance on other cognitive tasks. For example, people who perform worse on working memory tests might be less able to

maintain the strategy goals necessary to strategically shift criteria, resulting in little to no shifting. However, we found no significant relationship between the extent of  $C_n$  on recognition memory and visual detection tests with standardized test measures that assess risk aversion, response inhibition, working memory, and task-switching ability. Self-reports of motivation to perform well on the tasks also showed no relationship with the extent of  $C_n$ , nor did measures of metacognitive sensitivity. Interestingly, in Experiment 4 we found a strong relationship between the extent of  $C_n$  when making binary responses versus metacognitive  $C_n$  during recognition memory tests,  $r(168) = .53$ , but this only occurred when the criterion manipulation included instructions to respond based on levels of confidence. When the criterion manipulation did not cue participants into responding with confidence in Experiment 5, we found no significant relationship between the extent of  $C_n$  during recognition memory tests with a base rate manipulation and metacognitive  $C_n$ . This suggests that people implement different decision strategies for conveying meta awareness of the uncertainty in familiarity strength via high confident responses versus strategically shifting criteria. However, it is possible that we did not make the criterion manipulation extreme enough to appropriately align with people's criteria for responding with high confidence. For instance, if we made the base rate manipulation more extreme (e.g., 95% of test items are targets or nontargets), then people may have shifted criteria to greater extents and we might have observed a significant relationship between the extent of  $C_n$  and metacognitive  $C_n$ . Still, our findings in Experiment 5 match those of Miller and Kantner (2020) suggesting that people use different strategies when establishing decision criteria for rating recognition memory judgments with “high confidence” versus making decisions in situations where extreme criterion shifts promote better decisional outcomes. Although we and Aminoff and colleagues (2012) tested many factors that could potentially relate to criterion shifting tendencies, there still are countless numbers of other measures that may explain individual differences in criterion shifting tendencies. As of right now, there currently are no known measures that can reliably predict the extent to which an individual will strategically shift criteria except for criterion shifting performance itself on another task. However, it is possible that other characteristics are associated with criterion shifting tendencies that have yet to be tested. Despite this, we believe the strong stability of criterion shifting across time, tasks, and decision domains, coupled with the fact that individual differences cannot be easily attributed to a number of other factors, demonstrates that the tendency to strategically criterion shift appears to be a uniquely individualistic cognitive trait.

Although criterion shifting tendencies are quite stable within people, there are vast individual differences across people (Aminoff et al., 2012, 2015; Frithsen et al., 2018; Kantner et al., 2015; Layher et al., 2018; Miller & Kantner, 2020). Individual differences in strategic criterion shifting do not appear to be a result of an *inability* for certain people to shift criteria. In Experiments 4 and 5, almost all individuals used much more extreme criteria for responding “old” and “new” with high confidence on a six-point scale compared with the extent of criterion shifting even when the criterion manipulation specifically instructed participants to respond with “high confidence only.” Mickes and colleagues (2017) made a similar finding by showing that participants establish a much more conservative criterion when responding with the high-



est level of confidence on a multipoint scale compared with when instructions state to only respond “old” when there is 100% confidence. This shows that individuals are indeed *capable* of shifting criteria to more extreme extents if they simply adopt the same extreme criteria for responding with high confidence as they do for strategic criterion shifting. However, it appears that extreme differences in strategic criterion shifting are a result of individual differences in a *willingness* to disregard uncertain evidence in favor of a decision strategy that maximizes accuracy or payoffs<sup>4</sup> (Aminoff et al., 2012; Green & Swets, 1966; Kantner et al., 2015; Miller & Kantner, 2020). We believe that considering strategic criterion shifting tendencies as a stable cognitive trait attributable to individual differences in a willingness to shift a criterion will better inform theories of criterion placement and shifting. We encourage future studies to examine the nuances of these individual difference to gain a full understanding of how people adapt decision criteria to particular situations and outline specific considerations when investigating these decision strategies at an individual level.

### An Individual Differences Approach can Elucidate the Nature of a Phenomenon

**Suboptimality.** The fact that criterion shifting tendencies are stable within participants, yet variable across people, emphasizes the importance of understanding criterion shifting tendencies at an individual level. For several decades, most studies of criterion shifting drew conclusions from group-averaged data, neglecting the vast individual differences in shifting behavior. One generalized conclusion drawn from group averages is that people are quite suboptimal at appropriately adapting a decision criterion to a particular situation (Benjamin et al., 2009; Hirshman, 1995; Kubovy, 1977; Lynn & Barrett, 2014; Maddox & Bohil, 2005; Parks, 1966; Thomas & Legge, 1970; Uehla, 1966). Although the classification of an “optimal” criterion is strongly debated (Lynn & Barrett, 2014), we want to convey that decisional outcomes can dramatically vary depending on how well individuals adapt decision criteria to a particular situation. For example, if only 25% of items on a recognition test contain previously studied images (targets), then the optimal criterion is conservative, but the magnitude of the optimal placement will depend on how well a person can discriminate between old and new images. If a person is completely unable to distinguish between old and new images, a maximally conservative criterion is optimal because responding “new” every time will result in a correct response rate of 75% (no false alarms, but no hits either). However, if items are highly familiar, then the optimal conservative criterion is much less extreme because an individual can be correct more than 75% of the time by responding “old” to very familiar items even if it results in an occasional false alarm.

Many theories of suboptimal criterion shifting posit that people fail to shift criteria extreme enough because individuals probability match (Parks, 1966; Thomas & Legge, 1970), erroneously estimate signal and noise distributions (Kubovy, 1977; Uehla, 1966), poorly integrate decisional evidence with decisional outcomes (Lynn & Barrett, 2014), or are unable to maintain a stable criterion during a test block (Benjamin et al., 2009). These are reasonable explanations based on group averages, but many of these theories inadequately describe performance at an individual level because

some people actually do consistently shift criteria quite well, some shift to modest degrees, while others hardly shift at all. Even when individuals inadequately shift criteria across situations, there are instances where people will consistently establish an appropriately conservative (or liberal) criterion in one situation, but fail to shift when a liberal (or conservative) criterion is advantageous in another situation. For example, subject 2 in Experiment 2 deploys such conservative criteria in the conservative conditions (when 25% of test image are old) and subsequently makes a correct response 73% of the time on average. Yet, this subject fails to shift criteria in the liberal conditions (when 75% of test images are old) resulting in being correct only 44% of the time on average (see Figure 6). These instances are at odds with theories that suggest people poorly estimate signal and noise distributions (Kubovy, 1977; Uehla, 1966) or misestimate decisional parameters given the strength of discriminability (Lynn & Barrett, 2014) because it is unreasonable to believe these individuals are quite skilled at such estimations in some situations (e.g., when a conservative criterion is advantageous), but are grossly inept in others (e.g., when a liberal criterion is optimal). A theory of suboptimal criterion shifting must account for these individuals who strategically adapt a conservative criterion, but fail to shift to a liberal criterion (and vice versa). That is, the degree to which people are suboptimal at adapting a decision criterion depends on the individual *and* the situation.

Assessments of individual differences may not necessarily falsify previous hypotheses of suboptimal criterion shifting based on group averages, but these assessments certainly better inform them. For instance, Benjamin and colleagues (2009) suggest that a participant’s criterion will fluctuate throughout a test block creating “criterial noise” that results in measurements of criterion placement that are suboptimal. Criterial noise is a plausible phenomenon that might be occurring at an individual level. However, this hypothesis needs refinement to include the possibility that the amount of criterial noise may vary considerably across people. That is, some individuals may have a lot of criterial noise which may lead to relatively small criterion shifts, while others who shift to large extents may do so with little to no criterial noise. Any account of suboptimal criterion strategies must consider these consistent individual differences to fully understand the nature of strategic criterion shifting.

**Improving criterion shifting through awareness, feedback, and motivation.** Findings from group averages reveal that criterion shifting is improved when people are made aware of the

<sup>4</sup> Green and Swets (1966, p. 91) nicely describe the potential thought process of an individual who is unwilling to shift criteria to extreme extents: “The observer tends to avoid extreme criteria: when the optimal  $\beta$  is relatively large, his actual criterion is not so high as the optimal criterion, and when the optimal  $\beta$  is relatively small, his criterion is not so low as the optimal criterion. Although this pattern is consistent with studies of decision making under uncertainty which do not involve ambiguous sensory information, the significance of its appearance here is not totally clear. It may be suspected that the subject’s natural disinclination to make the same response on all trials is strengthened by his awareness that the experimenter’s principle interest is in a sensory process. He probably finds it difficult to believe that he would be performing responsibly if the sensory distinctions he makes are exactly those that he could make by removing the earphones in an auditory experiment or by turning his back on a visual signal.”

advantages for shifting, provided with feedback on criterion shifting performance, and presented with motivating factors to shift (Rhodes & Jacoby, 2007). If people are unaware of the advantages for shifting criteria, criterion shifts are generally not observed (Rhodes & Jacoby, 2007; Verde & Rotello, 2007). When people are made aware of the advantages for shifting, the extent of criterion shifting increases *on average*, but analyses of individual differences reveal that this is not true for everyone (Aminoff et al., 2012, 2015; Frithsen et al., 2018; Kantner et al., 2015; Layher et al., 2018; Miller & Kantner, 2020). Our studies revealed that these individual differences in criterion shifting behavior are remarkably consistent across multiple testing sessions. People who shift to large extents during one testing session do not simply regress back to the mean on subsequent sessions. Rather, awareness of the advantages for criterion shifting impacts the extent of shifting differently for each person.

To increase awareness of the advantages for shifting criteria, some studies provide corrective feedback, which improves criterion shifting performance at a group level (Kantner et al., 2015; Rhodes & Jacoby, 2007). However, the extent to which this is true may vary considerably across individuals. For instance, participants in our studies received performance feedback at the end of each test block (Experiment 3) or testing session (Experiments 1 and 2), which may have cued some participants to shift to greater extents on subsequent sessions to increase total payout (Experiments 1 and 3) or accuracy (Experiment 2). Although some people shifted to greater extents after the first session, several others shifted to similar or lesser degrees during successive sessions (e.g., subjects 19 and 37 in Experiment 1; see Figure 4). It seems that criterion shifting tendencies are unaffected by feedback for some individuals. However, feedback may more effectively alter individual criterion shifting tendencies if participants directly benefit from shifting to greater extents. Kantner and colleagues (2015) found that corrective feedback made individuals shift criteria more extremely when a payoff manipulation induced criterion shifting versus a paradigm where participants simply received instructions to shift. Because a criterion manipulation with instructions does not affect a participant's total payment, some individuals may be unwilling to shift criteria more extremely in response to feedback. However, when shifting to greater extents leads to a greater payout, feedback seems to be more effective at altering criterion shifting tendencies for *some* individuals. Future studies must assess how feedback under different circumstances affects *individual* criterion shifting tendencies as feedback will likely make some individuals consistently shift to greater extents, others will likely be completely unaffected by feedback, and some may only be affected by feedback under certain conditions (e.g., when there is a direct benefit for shifting criteria).

Another factor that may affect the extent of criterion shifting is an individual's motivation to shift criteria, which we found to be unrelated to a person's self-reported motivation to perform well during recognition memory and visual detection tests that incentivized criterion shifting. On average, participants shifted criteria to a greater extent in response to the payoff manipulation in Experiment 1 ( $M_{[Cn]} = 2.00$ ) compared with the base rate manipulation in Experiment 2 ( $M_{[Cn]} = 1.27$ ), even though both manipulations required the same degree of criterion shifting for *optimal* performance in an SDT framework. The payoff manipulation in Experiment 1 may have motivated some individuals to shift to a

greater extent compared with the base rate manipulation in Experiment 2, because the extent of criterion shifting directly impacted payment. However, some individuals in Experiment 1 continuously shifted to a small degree across all 10 sessions despite receiving relatively low payouts, while many individuals shifted to a large extent during Experiment 2 even though doing so did not affect total payment. This suggests that there might be individual differences in the factors that motivate people to criterion shift to greater extents, but within-subject paradigms are needed to ensure that this finding is not due to other factors (e.g., individuals may simply shift criteria to lesser extents in response to a base rate vs. payoff manipulation regardless of motivating factors). Kantner and colleagues (2015) observed individual differences in motivating factors for shifting criteria during recognition memory tests in a paradigm where a study phase preceded a test phase in one condition, but "malfunctioned" and did not actually present any images in another condition. In the latter case, participants should be highly motivated to shift criteria because there is no reliable memory evidence to guide the decision. Some individuals in the test condition without a study phase appropriately maximized responses by always responding "old" or "new" depending on the criterion manipulation, but others failed to adequately shift criteria despite never actually viewing any images before the test phase! The extent of criterion shifting was completely unaffected by the presence of a study phase or not for some individuals. Post study debriefings suggest that these participants still attempted to use perceptual features, such as skin tone, to guide decisions despite being told that such features are not diagnostic of whether an image is old or new. Some individuals seem more motivated to attempt to provide correct responses instead of consistently choosing the response that maximizes accuracy or payoffs, even under conditions of complete uncertainty. Overall, there are individual differences and several nuances in the degree that awareness, feedback, and motivation affect the extent of criterion shifting. Assessments of group averages are insufficient for identifying ways to improve criterion shifting performance because the influence of these factors on the extent of criterion shifting seems specific to the individual.

### Consequences of Not Criterion Shifting

Failing to adequately shift decision criteria can be quite consequential, particularly at the individual level. To illustrate this, we present performance and payment outcomes from Experiment 3 where participants earned five cents for each correct response, lost 10 cents for critical errors, but received no penalty for noncritical errors during two sessions of recognition memory and visual detection tests. On average, participants earned a total bonus of \$26.32 across the four tests and attained a mean  $d' = 1.10$ . Given our payout structure and SDT model, a person with a  $d' = 1.10$  who shifts criteria to an extent that maximizes total payment would earn \$29.30. The fact that participants on average only fall short of the maximum payout (given the mean level of  $d'$ ) by 11% suggests that the consequences of suboptimal criterion shifting are relatively minor. However, when examining individual performance, it becomes quite clear that not shifting criteria carries major consequences. For instance, we compare Experiment 3 subject 4 (E3-4) with E3-123 who both attained relatively low mean  $d'$  scores across both tasks and sessions ( $M_{[d']} = 0.52$  and  $M_{[d']} = 0.36$ ,

respectively). E3-4 on average did not shift criteria across the four tests ( $M_{[Cn]} = -0.05$ ), whereas E3-123 shifted criteria quite well ( $M_{[Cn]} = 3.38$ ). Even though E3-4 garnered more correct responses than E3-123, this individual only earned a bonus of \$9.75 while E3-123 earned \$24.25. By simply shifting criteria, E3-123 earned 2.5 times more money than E3-4 despite worse discriminability performance! However, classifying E3-4 as being generally suboptimal at placing a criterion is an inaccurate depiction of this individual's behavior. On average, E3-4 maintained a conservative criterion ( $M_{[Cn]} = 0.62$ ) when false alarms resulted in a 10-cent loss, and earned \$10.85 across all conservative conditions. In the liberal conditions, E3-4 established extremely suboptimal criteria ( $M_{[Cn]} = 0.67$ ) resulting in a *loss* of \$1.10. This individual can appropriately adopt a conservative criterion when the situation calls for it, but maintains that same conservative criterion when a liberal criterion is advantageous. Theories attempting to explain suboptimal criterion shifting behavior must account for this phenomenon. The relationship between the extent of criterion shifting and total payment in Experiment 3 is not limited to these select subjects but extends across all participants to a modest degree,  $r(170)$  range: .37–.44,  $Mdn = .42$ . The relative consequences for inadequate criterion shifting observed in our studies may generalize to real world scenarios.

There are many situations where extreme criterion shifts are necessary for avoiding consequential errors. One real-world example comes from radiologists who must assess whether a mammogram shows signs of breast cancer. If a radiologist falsely identifies an abnormal mammogram, then the patient must endure unnecessary worry while undergoing additional costly examinations. However, if a radiologist misses an abnormal growth, then a breast cancer diagnosis and subsequent treatment will be delayed increasing the likelihood of major surgery (e.g., a mastectomy) or even death. Studies reveal vast individual differences in the rate that radiologists recall patients for further testing and a more conservative criterion is associated with an increased miss rate (as expected; Gur et al., 2004; Yankaskas, Cleveland, Schell, & Kozar, 2001). Yankaskas and colleagues (2001) examined patient recall rates across 31 practices and found a large range from 1.9% to 13.4%. This means that some radiologists establish very conservative criteria for identifying an abnormal mammogram while others set much more liberal criteria. Although Yankaskas and colleagues (2001) could not assess the extent of criterion shifting because there is not a second criterion condition to compare against, it is presumed that these radiologists needed to shift their decision criteria when identifying an abnormal mammogram where errors result in extreme consequences relative to everyday decisions where the consequences of an error are negligible. The fact that the patient recall rate is so variable suggests that at the very least there are vast individual differences in the end result of criterion shifts, even when examining a group of experts.

Examining individual criterion shifting tendencies can prove challenging in situations where there are insufficient observations from each individual. For instance, an eyewitness to a crime who needs to select potential suspects from a lineup should consider the potential costs of falsely identifying an innocent person versus missing the perpetrator. If an identification will simply lead to further questioning of the suspect, then the eyewitness should establish a liberal criterion, because questioning an innocent person is only a minor inconvenience. However, if an eyewitness'

statement could substantially impact whether or not a suspect is arrested, then a more conservative criterion should be maintained to avoid incarcerating a potentially innocent person. Assessing whether eyewitnesses establish appropriate decision criteria is challenging because typically there is only one observation for each person. This means that analyses of individual differences cannot be conducted and conclusions are limited to group-averaged results. For example, Mickes and colleagues (2017) conducted photo lineup recognition tests and found that people on average establish less extreme criteria when making a binary response with a criterion manipulation versus the criterion set for the highest level of confidence in judgments made on a multipoint scale. However, it is likely that *some* people actually do appropriately establish extreme criteria when instructed to do so, but individual differences are impossible to evaluate in this paradigm because each participant only contributes a single observation.

### Strategic Criterion Shifting

Our studies specifically measured the stability of *strategic* criterion shifting tendencies where individuals explicitly received information indicating an advantage for shifting criteria. Information about the testing conditions appeared on every trial and participants could respond with unlimited time. Strategic criterion shifting occurs when people proactively set a goal to either avoid false alarms or misses depending on the situation. Criterion shifting stability may differ in situations where participants are not explicitly informed of the testing conditions (Verde & Rotello, 2007), are provided with false information (Selmecky & Dobbins, 2013), are affected by sequential dependencies from prior responses (Malmberg & Annis, 2012), or are in situations where a time pressure is imposed (Ratcliff, Smith, Brown, & McKoon, 2016). Future research needs to investigate the relationship, if any, between individual differences in strategic criterion shifting and criterion shifting tendencies in situations where participants are not explicitly informed of the advantages for doing so or when speeded responses are required. In the case of strategic criterion shifting, many individuals consistently fail to adequately shift criteria despite explicitly knowing the advantages for doing so.

### Why Do Some People Give a Criterion Shift, but Not Others?

There are many similarities between criterion shifts and confidence ratings that inform our understanding of why there are vast individual differences in strategic criterion shifting tendencies. Criterion manipulations served as the original method for obtaining cumulative hit and false alarm rates to create receiver operator characteristic (ROC) plots (Swets, Tanner, & Birdsall, 1955; Tanner, Swets, & Green, 1956), which illustrate the hit and false alarm rate for all possible decision criteria at a specific level of discriminability. Later, the implementation of confidence judgments provided an analog to strategic criterion shifts for recognition memory tests (Egan, 1958). Minimal instructions are required for people to accurately rate confidence during memory tests suggesting that people regularly assess the level of confidence associated with memories over the course of a lifetime (Mickes et al., 2011). Although participants typically have high metacognitive sensitivity when scaling the strength of memories to varying levels of confidence, there are limits in the degree to which



this is achieved, particularly as discriminability increases (Stretch & Wixted, 1998a). The highest levels of confidence should be reserved for the strongest and weakest memories, which should result in virtually no false alarms or misses, respectively. However, when participants report confidence ratings on a 20-point scale, many individuals will respond “old” with the highest confidence ratings more often than any other confidence level (Criss, 2009; Mickes et al., 2011). Mickes and colleagues (2011) believe this occurs because people struggle to finely scale confidence with strong memories. Thus, it appears that the criterion for “old” responses with the highest confidence level is less extreme than it theoretically should be. Interestingly, there are individual differences in the degree to which people finely scale strong memories suggesting that some individuals are more adept than others at establishing more extreme criteria for the highest levels of confidence (i.e., individual differences in metacognitive bias; see Figure 8 from Mickes & colleagues, 2011).

When examining group averages, ROC curves produce similar curvilinear shapes regardless of whether confidence ratings are acquired or a criterion manipulation is implemented (Dube & Rotello, 2012; Koen & Yonelinas, 2011; Macmillan & Creelman, 2005). This suggests that individual tendencies to finely scale confidence might be related to individual differences in the extent of criterion shifting. After all, when extreme criteria are advantageous people should only respond with the riskier option when there is high confidence in the decisional evidence. In Experiment 4, we found a strong relationship between the extent of criterion shifting and metacognitive bias when the criterion manipulation included instructions to respond with high confidence only. However, in Experiment 5 no such relationship existed when the criterion manipulation did not make any reference to confidence levels. When not explicitly cued to respond with confidence, participants seem to adopt a decision strategy for shifting criteria that is unrelated to decision processes for assessing high confidence in a recognition judgment. Some individuals will have very low metacognitive bias, but not shift criteria much while others will have high metacognitive bias, yet shift criteria to large extents. Assessing confidence in a response and adapting a decision criterion to explicit instructions are two separate behaviors that appear largely independent of each other. While confidence judgments may represent a meta assessment of the *varying strength* of memory evidence, we believe criterion shifting represents a mode of response that can strategically vary for any *single strength* of evidence. That is, each test item will elicit a degree of memory strength that a participant can convey through a confidence judgment, but the choice to identify an item as “old” or “new” depends on the situation.

Another key finding from Experiments 4 and 5 is that almost all participants adopted much more extreme criteria when responding with the highest levels of confidence compared with strategic criterion shifting. This suggests that these individuals are *capable* of shifting criteria to greater extents if they simply implement the same stringent criterion thresholds for rating recognition judgments with “high confidence” as they do for strategically shifting criteria when making an old/new judgment. Instead it appears that individuals are simply *unwilling* to disregard uncertain evidence in favor of a decision strategy based on known circumstances surrounding a decision (Aminoff et al., 2012; Kantner et al., 2015; Miller & Kantner, 2020). As Kantner and colleagues (2015) state, “people would rather attempt to be *correct* than be *correctly biased*.” When reporting on a six-point confidence scale, participants can convey the level of uncertainty in familiarity strength while still making old/new judgments as they

typically would on a recognition test. However, when forced to only respond “old” or “new” when there is high confidence, participants must identify relatively familiar items as “new” and relatively unfamiliar items as “old” to maintain the stringent criteria established for high confidence responses on a six-point scale. That is, memory evidence that elicits “low” or “medium” confidence for either an “old” or “new” judgment must be completely disregarded, which may feel unnatural or seem incorrect for some individuals even when this is the best decision strategy. We therefore believe individual differences in criterion shifting are a result of stable differences in peoples’ *willingness* to disregard uncertain evidence in favor of a decision strategy based on knowledge of payoffs, probabilities, or instructions (Aminoff et al., 2012; Kantner et al., 2015; Miller & Kantner, 2020).

Ultimately, we would like to understand why one individual is willing to shift a criterion, but not another. It is possible that individual criterion shifting tendencies are a response strategy learned across a lifetime of experience, similar to how Mickes and colleagues (2011) suggest that accurately scaling confidence judgments to recognition responses are learned over the course of one’s life. One bit of data from Aminoff and colleagues (2012) provides some intriguing insight into this hypothesis, albeit inconclusive and speculative; in their study, the subject pool consisted of 68 combat-experienced commissioned and noncommissioned officers from the U.S. Army Fort Irwin National Training Center along with 27 age-matched, nonmilitary participants from the community. Participants were categorized into nine different hierarchical levels of military rank, with nonmilitary participants ranked at the bottom. Interestingly, military rank turned out to be one of the few factors that significantly correlated with the extent of criterion shifting across both recognition memory tasks. Higher military rank associated with greater criterion shifts and this relationship could not be explained by other factors such as age or education level. It is possible that individuals who have learned to be more adaptive with their response strategies are better suited for the decision-making demands of high ranking military officials. Conversely, the experience of making decisions in those high ranking positions may have led to more adaptive response strategies. This single data point cannot provide definitive answers, but it should encourage future studies to more systematically assess why some individuals are more willing to shift criteria than others.

## Conclusion

Individual tendencies in strategic criterion shifting appear to represent a stable cognitive trait. We believe these tendencies result from individual differences in people’s willingness to disregard uncertain evidence in favor of a response that avoids a critical error (Aminoff et al., 2012; Kantner et al., 2015; Miller & Kantner, 2020). For example, when conducting a difficult recognition memory test that requires extreme criterion shifts, it is perfectly rational to simply look away from the screen and just choose the response that promotes better decisional outcomes. However, many people are reluctant to make such extreme shifts and may feel compelled to make responses based on memory, even a very poor one. We believe these demand characteristics are not an artifact of laboratory studies, but occur in real life situations where people may feel obliged to make decisions based on uncertain memory evidence. Evidently some individuals are completely comfortable with disregarding weak evidence and will shift criteria to extreme extents to optimize decisional outcomes. Other individuals appear to have a standard criterion and would rather rely

on memory evidence to make decisions while completely disregarding other situational information. Most individuals fall somewhere in between where individuals are both uncomfortable with completely abandoning memory evidence and ignoring situational information resulting in less extreme criterion shifts.

## References

- Aminoff, E. M., Clewett, D., Freeman, S., Frithsen, A., Tipper, C., Johnson, A., . . . Miller, M. B. (2012). Individual differences in shifting decision criterion: A recognition memory study. *Memory & Cognition, 40*, 1016–1030. <http://dx.doi.org/10.3758/s13421-012-0204-6>
- Aminoff, E. M., Freeman, S., Clewett, D., Tipper, C., Frithsen, A., Johnson, A., . . . Miller, M. B. (2015). Maintaining a cautious state of mind during a recognition test: A large-scale fMRI study. *Neuropsychologia, 67*, 132–147. <http://dx.doi.org/10.1016/j.neuropsychologia.2014.12.011>
- Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General, 142*, 1323–1334. <http://dx.doi.org/10.1037/a0033872>
- Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin, 74*, 81–99. <http://dx.doi.org/10.1037/h0029531>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1. <http://dx.doi.org/10.18637/jss.v067.i01>
- Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116*, 84–115. <http://dx.doi.org/10.1037/a0014351>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B. Methodological, 57*, 289–300. <http://dx.doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*, 433–436. <http://dx.doi.org/10.1163/156856897X00357>
- Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. *Journal of Personality and Social Psychology, 67*, 319–333. <http://dx.doi.org/10.1037/0022-3514.67.2.319>
- 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. <http://dx.doi.org/10.1016/j.jml.2006.08.003>
- Criss, A. H. (2009). The distribution of subjective memory strength: List strength and response bias. *Cognitive Psychology, 59*, 297–319. <http://dx.doi.org/10.1016/j.cogpsych.2009.07.003>
- Dube, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 130–151. <http://dx.doi.org/10.1037/a0024957>
- Egan, J. P. (1958). *Recognition memory and the operating characteristic* (Tech. Note No. AFCRC-TN-58-51, AO-152650). Bloomington, IN: Indiana University Hearing and Communication Laboratory.
- Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8*, 443. <http://dx.doi.org/10.3389/fnhum.2014.00443>
- Franks, B. A., & Hicks, J. L. (2016). The reliability of criterion shifting in recognition memory is task dependent. *Memory & Cognition, 44*, 1215–1227. <http://dx.doi.org/10.3758/s13421-016-0633-8>
- Frithsen, A., Kantner, J., Lopez, B. A., & Miller, M. B. (2018). Cross-task and cross-manipulation stability in shifting the decision criterion. *Memory, 26*, 653–663. <http://dx.doi.org/10.1080/09658211.2017.1393090>
- Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13*, 8–20. <http://dx.doi.org/10.3758/BF03198438>
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Hoboken, NJ: Wiley.
- Gur, D., Sumkin, J. H., Hardesty, L. A., Clearfield, R. J., Cohen, C. S., Ganott, M. A., . . . Rockette, H. E. (2004). Recall and detection rates in screening mammography: A review of clinical experience - implications for practice guidelines. *Cancer, 100*, 1590–1594. <http://dx.doi.org/10.1002/cncr.20053>
- Han, S., & Dobbins, I. G. (2008). Examining recognition criterion rigidity during testing using a biased-feedback technique: Evidence for adaptive criterion learning. *Memory & Cognition, 36*, 703–715. <http://dx.doi.org/10.3758/MC.36.4.703>
- Han, S., & Dobbins, I. G. (2009). Regulating recognition decisions through incremental reinforcement learning. *Psychonomic Bulletin & Review, 16*, 469–474. <http://dx.doi.org/10.3758/PBR.16.3.469>
- 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. <http://dx.doi.org/10.3758/BF03198243>
- Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 302–313. <http://dx.doi.org/10.1037/0278-7393.21.2.302>
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.
- Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. H. (2007). Working memory, attention control, and the N-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 615–622. <http://dx.doi.org/10.1037/0278-7393.33.3.615>
- Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition, 40*, 1163–1177. <http://dx.doi.org/10.3758/s13421-012-0226-0>
- Kantner, J., & Lindsay, D. S. (2014). Cross-situational consistency in recognition memory response bias. *Psychonomic Bulletin & Review, 21*, 1272–1280. <http://dx.doi.org/10.3758/s13423-014-0608-3>
- Kantner, J., Vettel, J. M., & Miller, M. B. (2015). Dubious decision evidence and criterion flexibility in recognition memory. *Frontiers in Psychology, 6*, 1320. <http://dx.doi.org/10.3389/fpsyg.2015.01320>
- Koen, J. D., & Yonelinas, A. P. (2011). From humans to rats and back again: Bridging the divide between human and animal studies of recognition memory with receiver operating characteristics. *Learning & Memory, 18*, 519–522. <http://dx.doi.org/10.1101/lm.2214511>
- Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (pp. 289–326). Cambridge, UK: Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511816789.012>
- Kubovy, M. (1977). Response availability and the apparent spontaneity of numerical choices. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 359–364. <http://dx.doi.org/10.1037/0096-1523.3.2.359>
- Layher, E., & Miller, M. (2019). *Who gives a criterion shift? A uniquely individualistic cognitive trait*. Retrieved from [osf.io/4k2hb](https://osf.io/4k2hb)
- Layher, E., Santander, T., Volz, L. J., & Miller, M. B. (2018). Failure to affect decision criteria during recognition memory with continuous theta burst stimulation. *Frontiers in Neuroscience, 12*, 705. <http://dx.doi.org/10.3389/fnins.2018.00705>
- Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., . . . Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied, 8*, 75–84. <http://dx.doi.org/10.1037/1076-898X.8.2.75>
- Lynn, S. K., & Barrett, L. F. (2014). “Utilizing” signal detection theory. *Psychological Science, 25*, 1663–1673. <http://dx.doi.org/10.1177/0956797614541991>

- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory, a user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.
- Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin*, *98*, 185–199. <http://dx.doi.org/10.1037/0033-2909.98.1.185>
- Maddox, W. T., & Bohil, C. J. (2005). Optimal classifier feedback improves cost-benefit but not base-rate decision criterion learning in perceptual categorization. *Memory & Cognition*, *33*, 303–319. <http://dx.doi.org/10.3758/BF03195319>
- Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology*, *57*, 335–384. <http://dx.doi.org/10.1016/j.cogpsych.2008.02.004>
- Malmberg, K. J., & Annis, J. (2012). On the relationship between memory and perception: Sequential dependencies in recognition memory testing. *Journal of Experimental Psychology: General*, *141*, 233–259. <http://dx.doi.org/10.1037/a0025277>
- Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General*, *140*, 239–257. <http://dx.doi.org/10.1037/a0023007>
- Mickes, L., Seale-Carlisle, T. M., Wetmore, S. A., Gronlund, S. D., Clark, S. E., Carlson, C. A., . . . Wixted, J. T. (2017). ROCs in Eyewitness Identification: Instructions versus Confidence Ratings. *Applied Cognitive Psychology*, *31*, 467–477. <http://dx.doi.org/10.1002/acp.3344>
- Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review*, *14*, 858–865. <http://dx.doi.org/10.3758/BF03194112>
- Miller, M. B., & Kantner, J. (2020). Not all people are cut out for strategic criterion shifting. *Current Directions in Psychological Science*, *29*, 9–15. <http://dx.doi.org/10.1177/0963721419872747>
- Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*, 61–64. <http://dx.doi.org/10.20982/tqmp.04.2.p061>
- Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018). BayesFactor package. Retrieved from <http://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf>
- Paivio, A. (1971). *Imagery and verbal processes*. New York, NY: Holt, Rhinehart, & Winston.
- Parks, T. E. (1966). Signal-detectability theory of recognition-memory performance. *Psychological Review*, *73*, 44–58. <http://dx.doi.org/10.1037/h0022662>
- Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*, 260–281. <http://dx.doi.org/10.1016/j.tics.2016.01.007>
- 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. <http://dx.doi.org/10.1037/0278-7393.33.2.305>
- Richardson, A. (1977). Verbalizer-visualizer: A cognitive style dimension. *Journal of Mental Imagery*, *1*, 109–126.
- Rogers, R. D., & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, *124*, 207–231. <http://dx.doi.org/10.1037/0096-3445.124.2.207>
- Rotello, C. M., Macmillan, N. A., Reeder, J. A., & Wong, M. (2005). The remember response: Subject to bias, graded, and not a process-pure indicator of recollection. *Psychonomic Bulletin & Review*, *12*, 865–873. <http://dx.doi.org/10.3758/BF03196778>
- Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology*, *43*, 450–461. <http://dx.doi.org/10.1037/0022-3514.43.3.450>
- Selmecky, D., & Dobbins, I. G. (2013). Metacognitive awareness and adaptive recognition biases. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 678–690. <http://dx.doi.org/10.1037/a0029469>
- Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM-retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*, 145–166. <http://dx.doi.org/10.3758/BF03209391>
- Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*, 137–149. <http://dx.doi.org/10.3758/BF03207704>
- Starns, J. J., & Olchowski, J. E. (2015). Shifting the criterion is not the difficult part of trial-by-trial criterion shifts in recognition memory. *Memory & Cognition*, *43*, 49–59. <http://dx.doi.org/10.3758/s13421-014-0433-y>
- 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. <http://dx.doi.org/10.1016/j.jml.2010.03.004>
- Stretch, V., & Wixted, J. T. (1998a). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1397–1410. <http://dx.doi.org/10.1037/0278-7393.24.6.1397>
- Stretch, V., & Wixted, J. T. (1998b). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1379–1396. <http://dx.doi.org/10.1037/0278-7393.24.6.1379>
- Swets, J. A., Tanner, W. P., & Birdsall, T. G. (1955). *The evidence for a decision-making theory of visual detection*. Technical Report No. 40, Electronic Defense Group, University of Michigan. <http://dx.doi.org/10.21236/AD0064143>
- Tanner, W. P., Swets, J. A., & Green, D. M. (1956). *Some general properties of the hearing mechanism*. Technical Report No. 30, Electronic Defense Group, University of Michigan.
- Thomas, E. A., & Legge, D. (1970). Probability matching as a basis for detection and recognition decisions. *Psychological Review*, *77*, 65–72. <http://dx.doi.org/10.1037/h0028579>
- Uehla, Z. J. (1966). Optimality of perceptual decision criteria. *Journal of Experimental Psychology*, *71*, 564–569. <http://dx.doi.org/10.1037/h0023007>
- Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition*, *35*, 254–262. <http://dx.doi.org/10.3758/BF03193446>
- Waldmann, M. R., & Göttert, R. (1989). Response bias in below-chance performance: Computation of the parametric measure  $\beta$ . *Psychological Bulletin*, *106*, 338–340. <http://dx.doi.org/10.1037/0033-2909.106.2.338>
- Watson, D., & Clark, L. A. (1994). *The PANAS-X: Manual for the Positive and Negative Affect Schedule—Expanded form*. Iowa City, IA: University of Iowa. <http://dx.doi.org/10.17077/48vt-m4t2>
- Wessel, J. R. (2018). Prepotent motor activity and inhibitory control demands in different variants of the go/no-go paradigm. *Psychophysiology*, *55*, 1–14. <http://dx.doi.org/10.1111/psyp.12871>
- Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*, 2020–2045. <http://dx.doi.org/10.1037/xge0000014>
- Witt, J. K., Taylor, J. E. T., Sugovic, M., & Wixted, J. T. (2015). Signal detection measures cannot distinguish perceptual biases from response biases. *Perception*, *44*, 289–300. <http://dx.doi.org/10.1068/p7908>
- Wixted, J. T., & Gaitan, S. C. (2002). Cognitive theories as reinforcement history surrogates: The case of likelihood ratio models of human recognition memory. *Animal Learning & Behavior*, *30*, 289–305. <http://dx.doi.org/10.3758/BF03195955>



Xu, Z., Adam, K. C. S., Fang, X., & Vogel, E. K. (2018). The reliability and stability of visual working memory capacity. *Behavior Research Methods*, *50*, 576–588. <http://dx.doi.org/10.3758/s13428-017-0886-6>

Yankaskas, B. C., Cleveland, R. J., Schell, M. J., & Kozar, R. (2001). Association of recall rates with sensitivity and positive predictive values

of screening mammography. *American Journal of Roentgenology*. *American Journal of Roentgenology*, *177*, 543–549. <http://dx.doi.org/10.2214/ajr.177.3.1770543>

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, *133*, 800–832. <http://dx.doi.org/10.1037/0033-2909.133.5.800>

## Appendix

### Experiment 3 Questionnaires

#### IMI: Effort/Importance Subscale (Ryan, 1982)

Please indicate how true this statement is for you:

1            2            3            4            5            6            7  
not at all            somewhat true            very true  
true

1. I put a lot of effort into this.
2. I did NOT try very hard to do well at this task.\*
3. I tried very hard on this task.
4. It was important to me to do well at this task.
5. I did NOT put much energy into this.\*

#### BIS/BAS: Fun-Seeking Subscale (Carver & White, 1994)

How true or false is this statement for you?

1            2            3            4  
very true            somewhat true            somewhat false            very false

1. I'm always willing to try something new if I think it will be fun.
2. I will often do things for no other reason than that they might be fun.

3. I often act on the spur of the moment.

4. I crave excitement and new sensations.

#### PANAS-X: Negative Affect (Watson & Clark, 1994)

Indicate to what extent you have felt this way during the past few weeks:

1            2            3            4            5  
slightly            a little            moderately            quite a bit            extremely

1. afraid
2. scared
3. jittery
4. nervous
5. irritable
6. hostile
7. guilty
8. ashamed
9. upset
10. distressed

(Appendix continues)

**VVQ (Modified): Verbalizer Score (Paivio, 1971; Richardson, 1977)**

Indicate how much you agree or disagree with this statement:

1            2            3            4            5            6            7  
 strongly    neither    neither    neither    neither    strongly  
 disagree    agree or    agree or    agree or    agree or    agree  
                  disagree    disagree    disagree    disagree

1. I enjoy doing work that requires the use of words.
2. I enjoy learning new words.
3. I can easily think of synonyms for words.
4. I read rather slowly.\*

5. I prefer to read instructions about how to do something rather than have someone show me.
6. I have better than average fluency in using words.
7. I spend very little time attempting to increase my vocabulary.\*

All questionnaires are scored by summing the numeric value (or reversely coded value) assigned to each item.

\*Reverse-coded items.

Received December 20, 2019

Revision received July 7, 2020

Accepted July 9, 2020 ■



AMERICAN PSYCHOLOGICAL ASSOCIATION

# APA Journals®

## ORDER INFORMATION

Subscribe to This Journal for 2021

Order Online:

Visit [at.apa.org/xlm-2021](http://at.apa.org/xlm-2021)  
for pricing and access information.

Call **800-374-2721** or **202-336-5600**

Fax **202-336-5568** | TDD/TTY **202-336-6123**

**Subscription orders must be prepaid.** Subscriptions are on a calendar year basis. Please allow 4-6 weeks for delivery of the first issue.

All APA journal subscriptions include Online First journal articles and access to archives. Individuals can receive online access to all of APA's scholarly journals through a subscription to APA PsycNet® or through an institutional subscription to the APA PsycArticles® database.

Visit [AT.APA.ORG/CIRC2021](http://AT.APA.ORG/CIRC2021)  
to browse APA's full journal collection