MEMORY, 2017
https://doi.org/10.1080/09658211.2017.1393090

Routledge
Taylor & Francis Group

Cross-task and cross-manipulation stability in shifting the decision criterion

Amy Frithsen[a], Justin Kantner[b], Brian A. Lopez[c] and Michael B. Miller[d]

[a]Department of Neurobiology & Behavior, University of California, Irvine, CA, USA; [b]Department of Psychology, California State University, Northridge, CA, USA; [c]Department of Psychology, Fullerton College, Fullerton, CA, USA; [d]Department of Psychological & Brain Sciences, University of California, Santa Barbara, CA, USA

ABSTRACT

In recognition memory experiments participants must discriminate between old and new items, a judgment influenced by response bias. Research has shown substantial individual differences in the extent to which people will strategically adjust their response bias to diagnostic cues such as the prior probability of an old item. Despite this significant *between* subject variability, shifts in bias have been found to be relatively predictive *within* individuals across memory tests. Experiment 1 sought to determine whether this predictability extends beyond memory. Results revealed that the amount a subject shifted response bias in a recognition memory task was significantly predictive of shifting in a visual perception task, suggesting that shifting can generalise outside of a specific testing domain. Experiment 2 sought to determine how predictive shifting would be across two manipulations well known to induce shifts in bias: a probability manipulation and a response payoff manipulation. A modest positive relationship between these two methods was observed, suggesting that shifting behaviour is relatively predictive across different manipulations of shifting. Overall, results from both experiments suggest that response bias *shifting*, like response bias setting, is a relatively stable behaviour within individuals despite changes in test domain and test manipulation.

ARTICLE HISTORY
Received 14 June 2017
Accepted 5 October 2017

KEYWORDS
Response bias; criterion; recognition memory

Psychologists have long been interested in how people make decisions under ambiguous circumstances. Signal detection theory (SDT) has proven to be a helpful method for estimating measures of discrimination ability (how well one can distinguish signal from noise) and response bias (how much evidence one needs in order to confirm the presence of a signal). Response bias is often measured as a criterion value, with larger values meaning that the person requires more evidence (in a memory test, more mnemonic evidence) in order to respond that a signal (e.g., a previously shown word) was presented (Green & Swets, 1988). Although originally used in the area of psychophysics, SDT was extended into the domain of recognition memory (e.g., Banks, 1970; Macmillan & Creelman, 2004; Miller & Lewis, 1977). Various studies have shown that response bias varies greatly from person to person (Aminoff et al., 2012, 2015; Kantner & Lindsay, 2012, 2014). Although extensive differences exist *between* individuals, research has provided evidence that response bias is rather stable *within* individuals. For example, Kantner and Lindsay (2012, 2014) found that subjects' criterion placement was consistent across time and across different stimulus types. Additionally, criterion placement was shown to relate to tasks outside of recognition memory such as false alarm rates on a recall version of the Deese-Roediger-McDermott (DRM) task (Kantner & Lindsay, 2012), accuracy on a Go-No task, and identifications on an eyewitness memory task (Kantner & Lindsay, 2014). Alzheimer's patients (who have a relatively liberal response bias compared to healthy controls) have demonstrated a similar consistency in criterion placement despite changes in stimulus type (Beth, Budson, Waring, & Ally, 2009) and changes in the length of study and test lists (Budson, Wolk, Chong, & Waring, 2006). Taken together, these results have led to the suggestion that response bias may be a relatively stable trait, akin to a personality characteristic, with subjects maintaining a natural proclivity in setting criterion across a variety of tasks.

Although it has been shown that response bias tends to be stable within an individual, it is not completely invariant, and can be altered in a predictable manner by certain experimental manipulations. These include instructional motivation (Egan, 1958; Strack & Förster, 1995), payoff manipulations that preferentially reward correct "old" or "new" responses (Healy & Kubovy, 1978; Van Zandt, 2000), and manipulations of the base rates of old and new items (Aminoff et al., 2012, 2015; Van Zandt, 2000). For instance, if told that the base rates of old and new items are 70% and 30%, respectively, subjects typically lower their criterion to a more liberal setting. While most

CONTACT Amy Frithsen afrithse@uci.edu Department of Neurobiology & Behavior, University of California, Irvine, 320 Qureshey Research Lab, Irvine, CA, USA

subjects are able to make this criterion shift, the extent to which they do so is almost always sub-optimal and varies greatly between individuals (Aminoff et al., 2012). Recent research has suggested that criterion shifting, like criterion setting, is relatively stable *within* individuals, despite its great variability *between* individuals. Aminoff and colleagues (2012) manipulated response bias within subjects during recognition memory tests using a base rate manipulation. Although there was considerable variability in how much each subject shifted criterion placement between conditions, this amount was highly consistent within an individual between the two recognition tests (one using words as stimuli and one using faces). Using a multistep regression analysis, the authors found that certain inherent characteristics, such as a fun-seeking personality, were strongly associated with criterion shifting tendencies. These results, along with the stability in criterion adjustment observed between stimulus sets, led the authors to suggest that response bias shifting may be a stable trait-like characteristic of an individual.

While shifts in response bias have been shown to be consistent within an individual during recognition memory tests using different stimulus types (Aminoff et al., 2012), no published work has yet examined the relationship between bias shifting in a memory task and shifting during other types of discrimination tasks outside the domain of memory. Experiment 1 of the current study investigated whether the amount of shifting in a memory task is predictive of shifting behaviour in tasks involving visual perception. Subjects participated in a recognition memory test (with either words or faces as stimuli) and in a perception task (testing either visual detection or discrimination). During testing sessions, subjects were explicitly told about the changes in the underlying base rates of signal and noise trials. During "likely" trials, 70% of the test items contained a signal and 30% contained only noise, and the opposite was true during "unlikely" trials.

We next sought to compare shifting behaviour across two different experimental manipulations. Experiment 2 compared shifting across two manipulations that are well documented in the literature as inducing shifts in response bias – a manipulation of the base rate of old items (as in Experiment 1) and a manipulation of response payoffs. In the payoff manipulation, subjects were motivated towards responding either "old" or "new" by varying the monetary gains/penalties associated with correct and incorrect responses. For instance, in one condition subjects were motivated to make correct "old" responses by offering a higher payoff for correct old responses (hits) than for correct "new" responses (correct rejections). Likewise, in this same condition, subjects would be penalised more for making incorrect "new" responses (misses) than for making incorrect "old" responses (false alarms).

With these two experiments we aimed to test the generalizability of shifting response bias across different domains (memory and perception) as well as across different procedural methods (base rate changes and monetary motivation). If response bias shifting behaviour is found to be relatively stable across these experimental paradigms, then this would provide evidence that *shifting* response bias, like response bias *setting*, is a relatively stable, domain-general individual trait.

Experiment 1: Comparing bias shifting in a memory task and a visual perception task

Method

The goal of this experiment was to determine whether the extent of bias shifting in a memory task predicts bias shifting in tasks of visual perception.

Subjects

Two-hundred and six healthy subjects (64 males) took part in this study. Subjects ranged in age from 17 to 26 years old ($M = 19.0$, $SD = 1.2$). Data from 12 additional subjects were not included in any analyses (8 due to technical issues, 2 failed to complete the task in its entirety, and 2 failed to make enough valid responses). Subjects were randomly assigned to one of the four groups: memory for faces/visual detection ($N = 51$), memory for faces/visual discrimination ($N = 54$), memory for words/visual detection ($N = 49$), or memory for words/visual discrimination ($N = 52$). Subjects were undergraduate students at the University of California, Santa Barbara and participated in exchange for course credit. All subjects gave informed consent as approved by the UCSB Institutional Review Board.

Stimuli

Responses were recorded and stimuli were presented using Matlab R2008a version 7.6.0 (The Mathworks Inc., USA) running the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997).

Words

Word stimuli consisted of a total of 260 nouns selected using the MRC Psycholinguistic Database (http://websitespsychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm). For counterbalancing purposes, words were pseudorandomly divided into 2 lists of 130 words each. These lists were matched as closely as possible on Kucera-Francis written frequency (list 1: $M = 79.12$, $SD = 25.78$; list 2: $M = 79.50$, $SD = 26.15$), number of letters (list 1: $M = 7.45$, $SD = 1.81$; list 2: $M = 7.57$, $SD = 1.81$), and number of syllables (list 1: $M = 2.49$, $SD = .68$; list 2: $M = 2.58$, $SD = .81$). Independent t-tests revealed no significant difference between lists in terms of Kucera-Francis written frequency $t(258) = .12$, $p = .96$; number of letters $t(258) = .51$, $p = .61$, or number of syllables $t(258) = .91$, $p = .36$. Half of the subjects had list 1 for targets and list 2 for distracters during the testing sessions (and vice versa for the other half of subjects). The words from each list were randomly assigned to the likely and unlikely

conditions for each subject (see procedure below). Words were presented in the centre of the screen in size 55 font.

Faces

Face stimuli consisted of 260 grey-scaled images taken from The Facial Recognition Technology (FERET) Database (http://www.itl.nist.gov/iad/humanid/feret/feret_master.html). All images were frontal views cropped to show from the chin up to the top of the hair. For counterbalancing purposes, faces were pseudorandomly divided into two lists of 130 faces each. These lists were matched as closely as possible by gender and race. Any images that included obvious distinguishing features such as facial piercings or facial hair were removed. Faces were 3.2 inches wide and 3.84 inches high and were presented in the centre of the screen.

Visual detection and visual discrimination

Visual stimuli consisted of a small square of Gaussian white noise. For the visual detection task, signal trials contained a small circle of white masked with noise in the centre of the square. For noise trials, no white circle was presented, and the entire square was noise only. During the visual discrimination task, stimuli were also a small square of Gaussian white noise. Signal trials included three Gabor lines that tilted slightly to the right. Noise trials included three Gabor lines that tilted slightly to the left (Figure 1).

Procedure

Subjects participated in two testing sessions: one memory test (either faces or words) and one visual perception test (either detection or discrimination). The order of the test sessions was counterbalanced such that approximately half of the subjects in each group started with the memory test while the other half began with the perception test. Specifically, all combinations of order and test type (i.e., memory or perception test first, words or faces, detection or discrimination) were randomly drawn from for each subject with the restriction that each order and test type was equally chosen over all subjects. For all trials, responses ranged from 1 to 6, depending on participants' confidence that the item contained signal: 1 – "Very Likely No", 2 – "Likely No", 3 – "Maybe No", 4 – "Maybe Yes", 5 – "Likely Yes", and 6 – "Very Likely Yes". This response scale was presented on the screen during each trial.

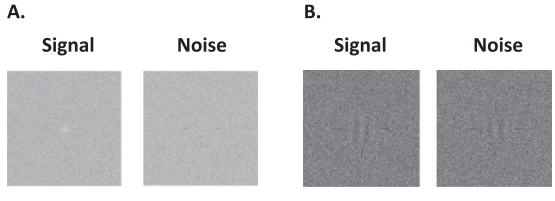

Figure 1. Examples of stimuli for the visual detection task (A) and the visual discrimination task (B).

Memory session

The memory tests included one test using words and one test using faces (each subject took either the words test or the faces test, but not both). The procedures for the memory tests were identical besides the type of stimuli used. Subjects passively viewed 130 items during the study session; they were told that their memory would be tested on these items and to try their best to remember as many as possible. Items were presented on the screen for 1.5 s followed by 0.5 s of a white screen. After the study session, subjects completed 20 math problems consisting of basic addition, subtraction, multiplication, and division. The test phase consisted of two memory runs, each containing 100 trials. Across these 200 trials, 100 were old (previously studied) and 100 were new (not studied). The two memory test runs were identical and divided into two runs simply to attenuate the effects of subject fatigue. Subjects were asked "Was this item on the study list?". Items were presented on the screen until a response was made. Subjects were told to make their responses as quickly and as accurately as possible. Prior to the start of the test sessions, subjects were told that during "LIKELY" blocks, 70% of the items would be old, and that during "UNLIKELY" blocks, only 30% of the items would be old. Each block contained between 9 and 11 trials[1]. During each block the word "LIKELY" or "UNLIKELY" was shown in the top centre of the screen, which served as the indicator of the current test condition.

Perception session

The perception tests included a visual detection task and a visual discrimination task (each subject took either the detection test or the discrimination test, but not both). The procedure for the two tasks was identical except for the stimuli used and the judgment made. At the beginning of each trial, a fixation crosshair was presented at the centre of the screen. Once the subject pressed the space key, the stimulus was presented for an interval tailored to each participant's d' on the task (see below). Then the stimulus was masked with an image of black and white dots for 0.3 s, after which the response scale was displayed on the screen. For the detection task, subjects responded to the question "Was there a white blob in the center of the image?". For the discrimination task, subjects were asked "Did the lines tilt to the right?". Subjects were told to make their responses as quickly and as accurately as possible.

Subjects first participated in a practice session (50 trials) to get accustomed to the task and to provide an estimate of their ability. During this practice session, the stimulus was presented for 0.2 s (detection) or 0.275 s (discrimination) and feedback was given on each trial as to whether the response was correct or incorrect. In a (relatively unsuccessful) attempt to equalise d' (see calculation below) across the memory and perception tasks, we modulated the duration that the perceptual stimulus was presented to each subject based on their d' during the practice session. The

Table 1. Mean d' values for each pair of tasks and associated significance values for the differences across tasks.

	Memory task			Visual task			*t* value	*p* value
	Likely	Unlikely	Overall	Likely	Unlikely	Overall		
Faces and detection	1.00 (.60)	1.11 (.60)	1.20 (.50)	1.24 (1.09)	1.47 (1.21)	1.59 (.99)	2.52	.02
Faces and discrimination	1.20 (.72)	1.17 (.68)	1.30 (.58)	1.22 (.81)	1.15 (.92)	1.33 (.68)	.141	.89
Words and detection	.71 (.56)	.76 (.60)	.86 (.48)	1.21 (1.13)	1.47 (1.16)	1.52 (.99)	3.87	<.001
Words and discrimination	.78 (.54)	.88 (.53)	.92 (.45)	1.03 (.80)	1.15 (.80)	1.22 (.78)	2.85	.006

Note: Standard deviations are in parentheses. *t* and *p* values reported are for the comparison between the "overall" d' values in each condition.

lower their practice session d', the longer the stimulus was presented during the test phase (with a lower limit of 0.05 s and an upper limit of 0.33 s). The testing sessions contained two runs consisting of 100 trials each. As in the memory test sessions, the two runs were identical and divided into two separate runs to minimise subject fatigue. Prior to the start of the test sessions, subjects were told that during "LIKELY" blocks, 70% of the images would contain signal (i.e., a blob or lines tilting to the right) and that during "UNLIKELY" blocks, only 30% of the test images would contain signal. Each block contained between 9 and 11 trials. During each block the word "LIKELY" or "UNLIKELY" was shown in the top centre of the screen, which served as the indicator of the current test condition. No feedback was provided during the testing sessions.

Results

Unless otherwise stated, the findings reported were created by collapsing across confidence levels into a binary "yes" or "no" response.

Discrimination abilities

Subjects' ability to discriminate trials which contained signal from those that only contained noise was calculated using the signal detection statistic d', where $d' = Z(\text{hit rate}) - Z(\text{false alarm rate})$ (Macmillan & Creelman, 2004; Stanislaw & Todorov, 1999). Despite several attempts to equalise discrimination abilities between tasks, there were significant differences in d' in three of the four cross-task comparisons using a multiple comparisons false discovery rate (FDR) correction of $q = .05$ (Benjamini & Hochberg, 1995). See Table 1 for statistics.

Response bias

Response bias was initially measured three ways: as *c*, calculated as $-1/2$ [Z(Hit rate) + Z(False alarm rate)] (Macmillan & Creelman, 2004; Stanislaw & Todorov, 1999), as the false alarm rate (Verde & Rotello, 2007), and as c_a, a receiver operating characteristic (ROC)-based criterion measure calculated from confidence ratings that takes into account any difference in the variances of the old- and new-item distributions (see Macmillan & Creelman, 2004). One subject in the words/visual discrimination task only used one confidence response, thus c_a could not be estimated and this participant was not included in the c_a analyses. Since there were significant differences in d' between most of the task pairings, we sought to test a measure of criterion that would account for these differences. A common way of doing this is to calculate a relative criterion value or c', where each subject's *c* value is divided by his/her d' value for that task. However, the tasks in the present experiment were designed to be difficult (to encourage subjects to base their judgments on the probability information), which led several subjects to have very low d' values. As a result, c' values were highly inflated. As an alternative, we regressed subjects' d' value for each condition (e.g., likely visual detection) with their *c* value for that condition (see Aminoff et al., 2012). The mean criterion *c* value for that condition was then added to each subject's (unstandardised) residual that was created in the previous step. This "normalized *c*" value represents a criterion value that takes each specific subject's recognition ability (d') into account but avoids the extreme values often produced using c'. Paired samples *t*-tests revealed that response bias was significantly higher in the unlikely condition compared to the likely condition for both the memory and visual perception tasks in all four groups, regardless of the bias measure used (all ts > 2.6, all ps < .01) with a multiple comparisons FDR correction of $q = .05$; see Table 2.

Correlation between bias methods

For each task, a "shift amount" was determined for each subject. This was calculated as response bias in the unlikely condition minus response bias in the likely condition. The resulting difference value indicated how much each subject shifted their criterion between conditions. This was done for all four measures of response bias (*c*, false alarm rate, ROC-based c_a, and normalised *c*) for each subject for each task. These shift values were then correlated across tasks. When one extreme outlier[2] was removed, correlations were significant for all four task pairings across response bias measures using a multiple comparisons FDR correction of $q = .05$. Similar results were found across the four different bias measurements, but see Table 3 for specific results and Figure 2 for a representative scatterplot of switch values.

Discussion

By manipulating the base rates of signal to noise trials, we induced a general shift in response bias, such that criterion values were higher in conditions where it was unlikely that

Table 2. Mean response bias values for each task pairing as calculated by c^A, false alarm rate^B, c_a^C, and normalised c^D. Shift values calculated as "Unlikely" response bias minus "Likely" response bias.

	Memory task			Visual task		
	Likely	Unlikely	Shift	Likely	Unlikely	Shift
A. *c*						
Faces and detection	−.10 (.30)	.37 (.32)	.46 (.42)	−.31 (.71)	.82 (.64)	1.13 (1.11)
Faces and discrimination	−.08 (.35)	.48 (.30)	.56 (.44)	−.38 (.54)	.33 (.40)	.71 (.78)
Words and detection	−.25 (.35)	.15 (.30)	.40 (.45)	−.07 (.52)	.58 (.44)	.65 (.55)
Words and discrimination	−.14 (.32)	.15 (.31)	.29(.37)	−.32 (.40)	.20 (.42)	.52 (.72)
B. False alarm rate						
Faces and detection	.36 (.16)	.19 (.11)	−.16 (.15)	.40 (.29)	.10 (.12)	−.30 (.32)
Faces and discrimination	.32 (.22)	.18 (.13)	−.15 (.24)	.43 (.21)	.21 (.16)	−.22 (.27)
Words and detection	.46 (.17)	.31 (.12)	−.15 (.20)	.35 (.25)	.13 (.11)	−.22 (.21)
Words and discrimination	.41 (.16)	.29 (.12)	−.12 (.16)	.42 (.18)	.25 (.16)	−.18 (.19)
C. ROC-based c_a						
Faces and detection	.01 (.28)	.32 (.27)	.31 (.25)	−.12 (.39)	.41 (.36)	.53 (.45)
Faces and discrimination	−.02 (.29)	.34 (.27)	.36 (.38)	−.30 (.34)	.21 (.29)	.51 (.54)
Words and detection	−.24 (.30)	.04 (.28)	.28 (.30)	−.04 (.37)	.34 (.29)	.38 (.37)
Words and discrimination	−.15 (.26)	.08 (.27)	.23 (.24)	−.21 (.21)	.10 (.19)	.30 (.29)
D. Normalised *c*						
Faces and detection	−.10 (.29)	.37 (.32)	.47 (.40)	−.31 (.68)	.82 (.59)	1.13 (1.00)
Faces and discrimination	−.08 (.96)	.48 (.79)	.56 (1.49)	−.38 (.99)	.33 (.84)	.71 (1.51)
Words and detection	−.26 (.70)	.15 (.59)	.41 (.89)	−.07 (1.01)	.58 (.86)	.65 (.99)
Words and discrimination	−.14 (.64)	.15(.60)	.29 (.70)	−.32 (.78)	.20 (.82)	.52 (1.41)

Notes: Standard deviations presented in parentheses. Note that for false alarms, negative shift values indicate a switch in the predicted direction. Values for the faces/visual detection condition are reported without the extreme outlier (see main text).

the trial contained signal (30%) and lower in conditions where signal was likely (70%) and the opposite was true for false alarm rates (i.e., higher rates in the likely condition). Although this general shift was observed across subjects, there was a great deal of variability between subjects in the extent of this adjustment, as indicated by the high standard deviations reported in Table 2. Despite this extensive variability between subjects, correlational analyses revealed a strong positive relationship *within* subjects across tasks. In other words, the extent of a subject's shift in response bias in a memory task was generally highly predictive of that subject's shift in a perceptual task. This relationship was found despite significant differences in task performance (measured by *d'*), suggesting that similar levels of discrimination are not necessary to observe a positive correlation in bias adjustment.

While Experiment 1 demonstrated stability in bias shifting across two very different judgment domains, the same bias manipulation (changing the underlying base rate of signal trials) was used for both tests. If bias flexibility is a characteristic inherent to an individual, this characteristic should be evident across *different* manipulations of bias as well. We tested this possibility in Experiment 2.

Experiment 2: Comparing bias shifts using a probability manipulation and a response payoff manipulation

The goal of this experiment was to determine if the amount of bias shifting in a memory task using a base rate manipulation would be predictive of bias shifting using a response payoff manipulation. These two methods are well documented in the literature as causing shifts in response bias (Aminoff et al., 2012; Aminoff et al., 2015; Healy & Kubovy, 1978; Van Zandt, 2000). A significant correlation of shifting amounts between these two methods would indicate cross-manipulation stability in shifting, while the lack of a relationship would suggest that response bias shifting is indeed manipulation-specific (Franks & Hicks, 2016).

Method

Subjects

A total of 72 individuals took part in this study. One subject failed to complete the study in its entirety and was therefore removed from subsequent analyses, leaving a total of 71 subjects (25 males) ranging in age from 17 – 23 years old ($M = 18.90$, $SD = 1.22$). Subjects were undergraduate students at the University of California, Santa Barbara and participated in exchange for course credit and payment (for the response payoff condition) that averaged

Table 3. Correlation coefficients (Pearson's *r*) for the correlation of switch values between tasks.

	Memory for faces				Memory for words			
	C	fa	c_a	norm *c*	*c*	fa	c_a	norm *c*
Visual detection	.17 (.35*)	.21 (.28*)	.29* (.42**)	.16 (.30*)	.64***	.57***	.50***	.68***
Visual discrimination	.53***	.61***	.49***	.52***	.53***	.43**	.29*	.55***

Note: * denotes significance $p < .05$, ** denotes significance $p < .005$, and *** denotes significance $p < .001$. Values in parentheses indicate the correlation coefficient when the extreme outlier was removed.

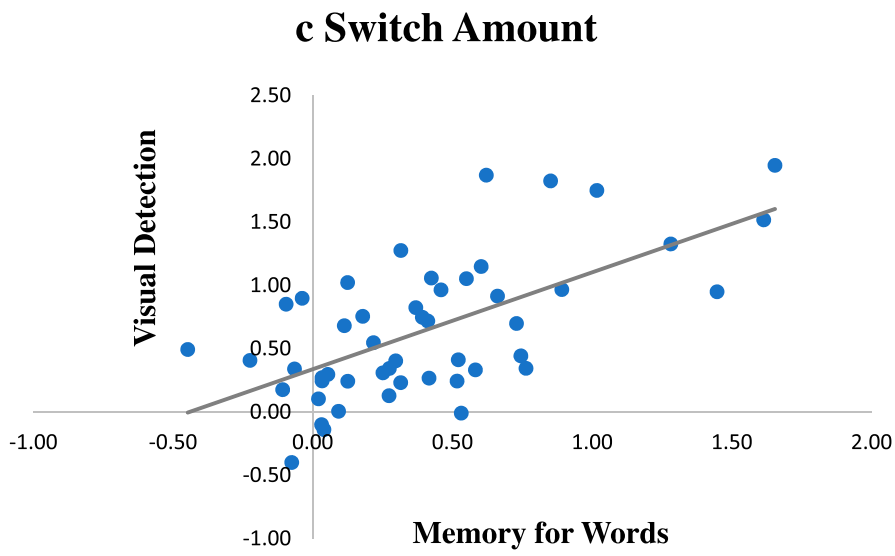

Figure 2. Scatterplot showing the switch amount (Unlikely *c* – Likely *c*) for the visual detection task and the memory for words task.

$12.86 per subject. All subjects gave informed consent as approved by the UCSB Institutional Review Board.

Stimuli

A total of 600 words (only nouns, 4–8 letters in length, with K-F written frequency > 25) were chosen from the MRC Psycholinguistic Database to be used for this experiment. Words that had similar roots, homonyms, homophones, or appeared in the instructions were excluded. Words were divided into four different subsets and randomly assigned to the four conditions (Likely, Unlikely, HRLPO, and HRLPN; see Procedure) for each subject. These subsets were matched as closely as possible on ratings of Kucera-Francis written frequency, concreteness, familiarity, and imageability. A one-way analysis of variance (ANOVA) revealed that there were no significant differences between the lists in terms of concreteness [$M = 498.96$, $SD = 90.98$, $F(3,596) = .47$, $p = .70$], familiarity [$M = 559.14$, $SD = 40.50$, $F(3,596) = .59$, $p = .62$], imageability [$M = 526.35$, $SD = 69.08$, $F(3,596) = .32$, $p = .81$], or K-F frequency [$M = 119.93$, $SD = 150.43$, $F(3,596) = 1.68$, $p = .17$]. Words were presented in the centre of the screen in size 80 font.

Procedure

Subjects participated in two study/test cycles (i.e., study list 1, test manipulation 1, study list 2, test manipulation 2), one using a probability manipulation and one using a response payoff manipulation. Manipulation order was counterbalanced between subjects. For each study session, subjects passively viewed 150 words presented serially at the centre of the screen for 1.5 s each, followed by a 0.25 s ISI consisting of a blank screen. Immediately following the study session, the subject began a recognition test using either the probability or response payoff manipulation. For each test session, subjects viewed 300 words, with each word remaining on the screen until a response was made. For the probability manipulation, half of the subjects started with the "Likely" condition, where 70% of the words (105) were from the study list (old) and only 30% (45) were new. Then subjects moved on to the "Unlikely" condition, where 70% of the words were new and only 30% were old. The order of these two conditions was reversed for the other half of subjects. For the response payoff manipulation, there were 75 old words and 75 new words for each of the two test conditions. Within each condition, we rewarded one type of correct response (e.g., a hit) more than another (e.g., a correct rejection) while simultaneously penalising one type of incorrect response (e.g., a miss) more than another (e.g., a false alarm). Specifically, in the "High Reward/Low Penalty Old" (HRLPO) condition, subjects were paid 15 cents for each hit and 5 cents for each correct rejection, and were penalised 5 cents for each false alarm and 15 cents for each miss. For the "High Reward/Low Penalty New" (HRLPN) condition, subjects were paid 15 cents for each correct rejection and 5 cents for each hit, and were penalised 5 cents for each miss and 15 cents for each false alarm. Half the subjects started with the HRLPO condition and then went on to the HRLPN condition, while the other half experienced the reverse order.

During each test session, an indicator was present on the screen which reminded the subject which condition they were currently in (i.e., More Likely Old Words, More Likely New Words, Higher Reward/Lower Penalty for Old Responses, Higher Reward/Lower Penalty for New Responses). Below each word was the same response scale used in Experiment 1. Prior to each study/test cycle, subjects were given a short practice session to familiarise themselves with the task.

Table 4. (A) Mean d' and response bias values as calculated by c, false alarm rate, c_a, and normalised c. (B) Shift values calculated as "Unlikely" response bias minus "Likely" response bias for the probability manipulation and as "HRLPN" minus "HRLPO" for the response payoff condition.

A.

	Probability		Response payoff	
	Likely	Unlikely	HRLPO	HRLPN
d	1.01 (.48)	1.21 (.54)	1.10 (.62)	1.15 (.67)
c	-.17 (.46)	.32 (.41)	-.09 (.45)	.19 (.39)
FA rate	.38 (.17)	.21 (.13)	.34 (.18)	.25 (.15)
c_a	-.16 (.45)	.32 (.41)	-.09 (.42)	.19 (.38)
Normalised c	-.20 (.30)	.12 (.25)	-.22 (.31)	-.03 (.27)

B.

	Probability				Response payoff			
	Shift	t	df	p	Shift	t	df	p
c	.48 (.64)	6.41	70	<.0001	.29 (.40)	6.44	70	<.0001
FA rate	−.16 (.17)	6.15	70	<.0001	−.09 (.14)	6.34	70	<.0001
c_a	.48 (.63)	7.93	70	<.0001	.28 (.37)	7.32	70	<.0001
Normalised c	.31 (.36)	5.16	70	<.0001	.19 (.27)	5.79	70	<.0001

Note: Standard deviations presented in parentheses. Note that for false alarms, negative shift values indicate a switch in the predicted direction. Statistics for the paired samples t-tests (t, degrees of freedom {df}, and the p value) for the Likely vs. Unlikely (probability) conditions and the HRLPO and HRLPN (response payoff) conditions.

Results

Unless otherwise stated, the findings reported were created by collapsing across confidence levels into a binary "yes" or "no" response.

Discrimination abilities

Subjects' ability to discriminate old and new words was calculated using the signal detection statistic d' as in Experiment 1. See Table 4. A one-way ANOVA with four levels (Likely, Unlikely, HRLPO, HRLPN) revealed no significant difference in discrimination ability between conditions, $F(3,210) = 1.77$, $p = .16$. However, a 2 (Probability/Reward) × 2 (Favor Old Item (Likely old & HRLPO)/Favor New Item (Unlikely Old & HRLPN) did reveal a main effect of whether old or new items were favoured, with the conditions where new items were favoured showing a significantly higher d' ($M = 1.14$, $SD = .47$) than conditions where old items were favoured ($M = 1.06$, $SD = .52$), $F(1,70) = 4.57$, $p < .05$. In other words, subjects were slightly more accurate when new items were favoured during both testing manipulations (i.e., during Unlikely New & HRLPN conditions) compared to when old items were favoured (i.e., during Likely Old & HRLPO conditions). There was no significant main effect of test manipulation (Probability/ Reward) on d', $F(1,70) = 1.36$, $p = .25$, nor was there a significant interaction, $F(1,70) = .42$, $p = .52$.

Response bias

Response bias was measured in four ways as described in Experiment 1: as c, the false alarm rate, c_a, and normalised c. By all four measures, results from paired samples t-tests revealed a higher response bias in the unlikely compared to the likely conditions and for the HRLPN compared to the HRLPO, all $ps < .0001$ with a multiple comparisons FDR correction of $q = .05$. See Table 4 for details. Somewhat surprisingly, results from paired samples t-tests (using a multiple comparisons FDR correction of $q = .05$) revealed that subjects shifted their response bias to a greater degree with the probability manipulation than with the response payoff manipulation, for c $t(70) = 3.09$, $p = .003$; for the false alarm rate $t(70) = 3.42$, $p = .001$; for c_a $t(70) = 3.15$, $p = .002$; and for normalised c $t(70) = 3.00$, $p = .004$.

Correlation between shifting manipulations (probability vs. response payoff)

As in Experiment 1, shift amounts were calculated for each subject using each measure of response bias for both manipulations (Unlikely – Likely for probability and HRLPN – HRLPO for response payoff). A significant positive correlation was found between subjects for all four measures. For c, a significant correlation was found between shift amount in the probability condition ($M = .48$, $SD = .64$) and shift amount in the reward condition ($M = .29$, $SD = .47$), $r(69), = .55$, $p < .0001$. Similarly, when response bias was measured by shift in false alarm rate, a significant correlation was found between the probability condition ($M = .16$, $SD = .17$) and the reward condition ($M = .09$, $SD = .14$), $r(69) = .32$, $p < .01$. When using c_a, a significant correlation was found between the probability condition ($M = .31$, $SD = .36$) and the reward condition ($M = .19$, $SD = .27$), $r(69) = .38$, $p = .001$. Finally, when using normalised c, a significant correlation was found between the probability condition ($M = .48$, $SD = .63$) and the reward condition ($M = .28$, $SD = .37$), $r(69) = .53$, $p < .0001$. All tests remained significant after correcting for multiple comparisons using an FDR correction of $q = .05$. As in Experiment 1, inspection of the scatterplots revealed a significant outlier in all four bias measures who seemed to be employing a maximising strategy (i.e., always responding "old" in the high probability/ HRLPO condition and always responding "new" in the low probability/ HRLPN condition) but this time in *both* tasks, thus increasing the correlation. Even after removing this subject from the analysis, a significant (although weaker) positive correlation remained

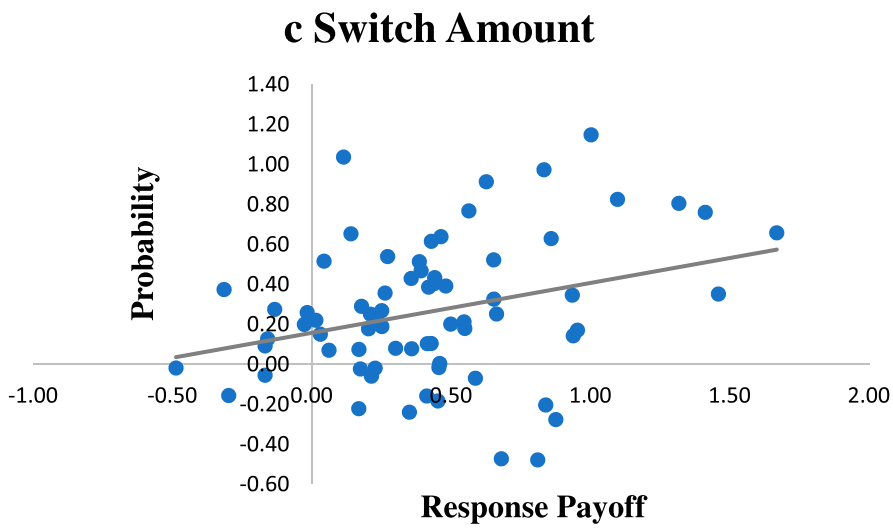

Figure 3. Scatterplot showing the switch amount for the response payoff condition (HRLPN *c* – HRLPO *c*) and the probability condition (Unlikely *c* – Likely *c*). The extreme outlier mentioned in the text has been removed from this scatterplot.

between tasks for all four bias measures (for *c*, $r = .30$, $p = .01$; for false alarm rate, $r = .23$, $p = .05$, for c_a, $r = .25$, $p = .04$, and for normalised *c*, $r = .31$, $p = .008$). See Figure 3 for a representative scatterplot of switch values.

Discussion

In this experiment we were able to induce a shift in criterion using two well-established methods: a manipulation of base rates and a response payoff manipulation. Both methods produced criterion shifts in the predicted direction, meaning that criterion values were significantly higher in conditions that favoured a new response (i.e., unlikely old and HRLPN) than in those which favoured an old response (i.e., likely old and HRLPO) with the opposite being true for false alarm rates (i.e., higher rates in conditions which favoured an old response). Furthermore, the degree to which subjects shifted in one paradigm was significantly correlated with their shift amount in the other paradigm. With the removal of a significant outlier, these correlations remained statistically significant, though modest in strength (average $r = .27$). Therefore, the results do suggest that there is *some* generalizability in bias shifting between manipulations, but perhaps not as much as when the same manipulation is used.

General discussion

In Experiment 1, we wanted to investigate how predictive, or generalizable, bias shifting behaviour is within individuals across various tasks. Previous research has shown consistent shifting behaviour within recognition memory tests using different stimuli (Aminoff et al., 2012). The results from Experiment 1 are, to our knowledge, the first to extend this finding outside of recognition memory. These tasks differed in the domain (memory vs. perception) and in their difficulty (as revealed by differences in d'). Despite these differences, the magnitude of a subject's bias shift between the likely and unlikely conditions in the memory task was predictive of that subject's shift in the perceptual task. This finding is important because it suggests that response bias *shifting*, like response bias *setting*, is a relatively stable trait-like characteristic of an individual that generalises across task types.

While Experiment 1 showed that shifting behaviour was significantly correlated between different domains using the base rate probability manipulation, Experiment 2 compared shifting patterns across different experimental manipulations of bias. Here, the results from two methods that are well known to induce shifts in response bias – a base rate probability manipulation and a manipulation of response payoffs – were compared. In general, it seemed that subjects were more willing to shift during the probability manipulation than during the response payoff manipulation. This may be related to the fact that the probability information was actually informative in terms of improving one's accuracy (i.e., the probability information could help one's chances of choosing the correct response), whereas the response payoff only increased one's monetary reward, adding no information as to the likelihood of responding correctly (for a discussion of this point see Kantner, Vettel, & Miller, 2015). Despite this difference in shifting magnitude between testing manipulations, we observed a significant correlation between methods, although the magnitude of this relationship (average $r = .27$) was weaker than that observed across judgment domains (but using the same shifting manipulation) in Experiment 1.

Further support for cross-manipulation generality of bias shifting tendencies comes from unpublished analyses of data reported by Kantner et al. (2015). They tested three different bias manipulations: response payoffs, probability "old", and a security patrol scenario in which either misses

(liberal condition) or false alarms (conservative condition) bore a greater subjective cost. Although shifting magnitudes under each individual shifting manipulation were of primary interest in the published analyses, each participant completed recognition tests using two of the three manipulations, allowing for follow-up analyses of the relationship between shifting magnitudes across tasks. These correlational analyses revealed relatively stable shifting across manipulations, with *r* vales ranging between .36 and .47 across pairs of manipulations. These results lend some support to the idea that criterion shifting using one type of manipulation is relatively predictive of shifting behaviour using another type of manipulation, in agreement with the results of Experiment 2.

The robust predictability of bias shifting found in Experiment 1, the modest predictability found in Experiment 2, and the moderate predictability in shifting evident in the Kantner et al. (2015) data provide converging evidence that shifting behaviour is generalizable both across task domains (i.e., using different judgments and test materials) and across testing methodology (i.e., using different testing manipulations). The latter conclusion, however, is at odds with that of Franks and Hicks (2016), who tested cross-manipulation stability in shifting behaviour using two different shifting manipulations: a probability manipulation (like the one used in the present experiments) and a manipulation of encoding strength (not used in the current experiments) in which some items were presented multiple times at study (strong condition) while others were presented only once (weak condition). Strength manipulations produce a simultaneous increase in the hit rate and a decrease in the false alarm rate for items studied multiple times relative to items studied only once, a phenomenon known as the strength-based mirror effect (Glanzer & Adams, 1985) that is often attributed to a shift in response criterion (Stretch & Wixted, 1998). Franks and Hicks (2016) found no statistical relationship of shifting behaviour across these two methods, and we independently obtained the same finding in a similar experiment (unpublished data). In light of the above results, however, we are hesitant to regard this as evidence that shifting is completely independent across methodologies. An alternative possibility is that probability-based criterion shifts do not predict shifting in the strength-based mirror effect because the mirror effect is not itself the result of a criterion shift. Criss (2006, 2010) has argued that the pattern of hit and false alarm rates observed in the mirror effect can be explained without reference to a criterion shift. According to this account, the increased encoding of strong items increases the familiarity of old items, but it also *decreases* the familiarity associated with *new* items (i.e., it strengthens the differentiation between old and new items). As a result, the change in false alarm rate reflects a shift downward along the memory evidence scale for the new distribution (foils) at test. However, this interpretation is difficult to reconcile with data showing changes in false alarm rate between strong and weak conditions *only* when those conditions were marked with an explicit cue, particularly since encoding strength was held constant between testing conditions (Hicks & Starns, 2014; Starns & Olchowski, 2015).

Another possibility is that the strength-based mirror effect does in fact represent a change in criterion, but that it does so in a way that is measurably different than a base rate manipulation. While signal detection models assume that decisions are made by comparing a single evidence value to a response criterion, sequential sampling models such as the drift diffusion model assume that decisions are based on the accumulation of several sequential evidence samples that are taken continuously in time, producing a drift rate in the evidence accumulation process until it reaches one of two decision boundaries – "old" or "new" (e.g., Starns, Ratcliff, & White, 2012). During probability and payoff manipulations, the manipulation itself is independent from the mnemonic evidence associated with each test item, which may move the starting point closer to one of the boundaries. For strength-based manipulations, the memory evidence itself is what is being manipulated. Therefore, it is plausible that in strength-based manipulations, decision boundaries are strategically shifted to account for the strengthened mnemonic targets, but the starting point is unaffected and remains equidistant from either boundary. Thus, from a diffusion model perspective, differences in the type of criterion affected by probability/payoff and strength-based manipulations may explain their apparent lack of a relationship using omnibus criterion measures such as *c*. The current experiments cannot speak to this possibility; therefore, future research should be conducted to determine the exact mechanism(s) underlying the strength-based mirror effect. Until then, the lack of a relationship between strength-based manipulations and other criterion manipulations is difficult to interpret.

Overall, the results of the current experiments provide some evidence that response bias shifting is generalizable, particularly across domains (recognition memory and visual perception tasks), and also across testing procedures (probability and payoff manipulations). To our knowledge, this is the first published study to compare shifting patterns between memory and non-memory tasks. The fact that we found a significant correlation between shifting behaviour across these very different testing domains suggests that this aspect of decision-making under ambiguous situations is relatively stable within an individual. The fact that we found a significant correlation in switching behaviour despite differences in the motivation guiding the switch (Experiment 2) further suggests within-subject stability of response bias shifting, though the magnitude of the correlation leaves ample room for additional elements of the recognition judgment that are not stable within individuals when the manipulation of bias differs.

We believe that the generalizability of response bias is an important avenue of research, pertinent to any memory task that requires judgments to be made under

uncertainty. The more we can determine how much of a person's memory performance is driven by factors that rely upon making the decision itself (regardless of the amount of mnemonic information retrieved), the more we can elucidate individual differences in memory-specific processes per se (e.g., how well the person actually remembers an event). Until then, any behavioural measures of memory will be an unknown mixture of decision-based processes and memory-specific processes.

Notes

1. Although this probability was not necessarily accurate for each block (i.e., each set of 9–11 trials) it was accurate over the 100 trials of each test for both the memory and perception tasks.
2. Upon visual inspection of the scatterplots, it became apparent that there was an extreme outlier (greater than five times the standard deviation of shifting values) that was greatly influencing the results in the memory for faces/visual detection task pairing. Investigation of this subject's data revealed an interesting response pattern. It appeared that this subject employed a maximising strategy during the memory for faces task, responding "yes" on almost every trial in the likely condition (100% hit rate and 93% false alarm rate) and "no" on every trial in the unlikely condition (0% hit rate and false alarm rate). This led to an extreme switch value. This subject did not employ a maximising strategy during the visual detection task, resulting in a modest switch value. Examining the other scatterplots revealed extreme outliers (greater than five times the standard deviation) in the memory for faces/visual discrimination task and in the memory for words/visual discrimination task pairing. However, these subjects had relatively high switch values in both tasks, and removing these subjects changed the associated correlation values by a negligible amount. These correlation values were robust even when the outlier threshold was lowered to three times the standard deviation, suggesting they were not driven by outliers.

Acknowledgments

The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. We would also like to thank Dr. Craig Abbey for the creation of the visual stimuli used in these experiments.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Institute for Collaborative Biotechnologies through grant [W911NF-09-0001] from the U.S. Army Research Office.

References

Aminoff, E. M., Clewett, D., Freeman, S., Frithsen, A., Tipper, C., Johnson, A., … Miller, M. B. (2012). Individual differences in shifting decision criteria: A recognition memory study. *Memory & Cognition*, *40*(7), 1016–1030.

Aminoff, E. M., Clewett, D., Freeman, S., Frithsen, A., Tipper, C., 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.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, *74*(2), 81–99.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, *57*, 289–300.

Beth, E., Budson, A. E., Waring, J. D., & Ally, B. A. (2009). Response bias for picture recognition in patients with Alzheimer's disease. *Cognitive and Behavioral Neurology: Official Journal of the Society for Behavioral and Cognitive Neurology*, *22*(4), 229–235.

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

Budson, A. E., Wolk, D. A., Chong, H., & Waring, J. D. (2006). Episodic memory in Alzheimer's disease: Separating response bias from discrimination. *Neuropsychologia*, *44*(12), 2222–2232.

Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory and Language*, *55*(4), 461–478.

Criss, A. H. (2010). Differentiation and response bias in episodic memory: Evidence from reaction time distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 484–499.

Egan, J. P. (1958). Recognition memory and the operating characteristic. *USAF Operational Applications Laboratory Technical Note*, *58* (ii), 32–51.

Franks, B. A., & Hicks, J. L. (2016). The reliability of criterion shifting in recognition memory is task dependent. *Memory & Cognition*, *44* (8), 1215–1227.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 5–16.

Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics* (Reprint edition). Los Altos, CA: Peninsula.

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*(5), 544–553.

Hicks, J. L., & Starns, J. J. (2014). Strength cues and blocking at test promote reliable within-list criterion shifts in recognition memory. *Memory & Cognition*, *42*(5), 742–754.

Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition*, *40*(8), 1163–1177.

Kantner, J., & Lindsay, D. S. (2014). Cross-situational consistency in recognition memory response bias. *Psychonomic Bulletin & Review*, *21* (5), 1272–1280.

Kantner, J., Vettel, J. M., & Miller, M. B. (2015). Dubious decision evidence and criterion flexibility in recognition memory. *Frontiers in Psychology*, *6*, 1320.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Miller, E., & Lewis, P. (1977). Recognition memory in elderly patients with depression and dementia: A signal detection analysis. *Journal of Abnormal Psychology*, *86*(1), 84–86.

Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149.

Starns, J. J., & Olchowski, J. E. (2015). Shifting the criterion is not the difficult part of trial-by-trial criterion shifts in recognition memory. *Memory & Cognition*, *43*(1), 49–59.

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

Strack, F., & Förster, J. (1995). Reporting recollective experiences: Direct access to memory systems? *Psychological Science*, *6*(6), 352–358.

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1397–1410.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 582–600.

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition*, *35*(2), 254–262.