Spring 5-2024

The Domain Specificity and Conscious Awareness of Learned Memory Biases

Gizem Filiz
Washington University in St. Louis

Spring 5-2024

The Domain Specificity and Conscious Awareness of Learned Memory Biases

Gizem Filiz

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Psychological and Brain Sciences

The Domain Specificity and Conscious Awareness of Learned Memory Biases

by
Gizem Filiz

A thesis presented to
Washington University in St. Louis
in partial fulfillment of the
requirements for the degree
of Master of Arts

May 2024
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to take this opportunity to show my sincere gratitude to my advisor, Dr. Ian G. Dobbins, for his valuable guidance and expertise throughout the entire process of my master's thesis. His support, sense of humor, and mentorship have been instrumental in shaping my academic and personal growth, and I feel fortunate to have had the opportunity to work with him. I would also like to extend my appreciation to my thesis committee members, Dr. Roddy Roediger and Dr. Zach Reagh, for their valuable time and support, which helped me enhance the quality of my work. Additionally, I would like to acknowledge the contributions of all members of the Memory and Decision-Making Lab, including Eylul Ardic, Xinran Zhang, and research assistance Samir Khare, and Henry Xiao, for their assistance in data collection. Finally, I would like to express my heartfelt gratitude to my family and friends and for their support and encouragement throughout this journey. Their constant motivation and faith in me have been a source of inspiration and motivation, and I could not have completed this thesis without their support.

Gizem Filiz

Washington University in St. Louis
May 2024

To my beloved family

ABSTRACT OF THE THESIS

The Domain Specificity and Conscious Awareness of Learned Memory Biases

by

Gizem Filiz

Master of Arts in Psychological and Brain Sciences

Washington University in St. Louis, 2024

Professor Ian G. Dobbins, Chair

The False Positive Feedback (FPF) manipulation encourages certain recognition errors by inducing either liberal (lax) or conservative (strict) recognition memory decision biases. FPF manipulation involves trial-by-trial probabilistic positive feedback for commission or omission errors while the other stimulus class always received fully correct feedback during testing. We investigated whether these learned biases are restricted to the stimulus class triggering FPF, or whether they instead spread to an intermixed class receiving valid feedback, by selectively delivering FPF to words or pictures. A spreading bias would suggest that subjects learn to be liberal or conservative in interpreting recognition evidence in general during the testing context (general recognition bias). A restricted bias, however, would indicate a specialized form of learning tied to each class's unique features (feature-specific bias). In Experiment 1, FPF applied to pictures yielded selective biases that did not spread to intermixed words (FPF applied to words was ineffective). In Experiment 2, FPF applied to words yielded selective biases that did not spread to pictures (FPF applied to pictures was ineffective). These results suggested that biases occurred in a feature-specific manner. Questionnaire data indicated that subjects were unaware of feedback's purpose and stimulus selective nature, suggesting that recognition decision biases

can be unintentionally acquired and yet specific to one of two classes of encountered memoranda.

Introduction

During recognition memory studies, subjects are presented with to-be-learned items, such as pictures or words, and often instructed to remember these materials for later memory test. Following this studied (old) and unstudied (new) items are intermixed, and the subjects are asked to determine whether each presented item is old/studied by choosing an ‘old’ or ‘new’ response.

Under detection theory accounts of recognition decision making, old and new items are assumed to evoke normally distributed evidence values, separated by a distance (d prime) indicating the strength of evidence for old items through study (Banks, 1970). As the average memory strength for old items increases through repeated or more elaborative study, it becomes easier for an observer to distinguish between old and new stimuli. Nonetheless, because the continuous old and new evidence distributions will always overlap, to some degree, individuals must also establish a decision criterion to determine the threshold of memory strength required to identify an item as old. According to detection theory, this decision criterion is assumed to be flexible and influenced by contextual factors.

It may be advantageous to favor one decision option over another when making choices, and this is referred to as a decision bias (Macmillan & Creelman, 2004). Decision biases often occur in recognition memory tasks, where individuals have shown behavioral characteristics favoring either old or new conclusions. For instance, it is more optimal to favor old recognition conclusions in frequently visited environments compared to novel environments because the incidence of recognized individuals will be generally higher in the former.

Experimentally, decision biases are sometimes induced by using monetary incentives, rendering one decision more profitable than the other. For example, Van Zandt (2000, Experiment 2) informed participants about the differential payoff assignment for old and new items. That is, the correct "new" response was worth one point. In contrast, a correct "old" response was worth three points without penalty for incorrect responses. The findings indicate that increasing the value of correct responses for "old" items led to a greater tendency to give "old" responses without altering accuracy. In addition to monetary manipulations, this study also demonstrated adaptive recognition biases when subjects were correctly informed about manipulations of the ratio of targets to lures in the test lists. In response, subjects favored the response favoring the most frequent class of probes. These examples of decision bias shifts suggest that recognition decision biases are largely strategic and controlled. However, there is limited evidence indicating that recognition decision biases adaptively shift in response to subtle, and perhaps unnoticed changes in environmental reinforcement favoring either old or new conclusions. The current study examines potentially implicit mechanisms that shape recognition decision biases through reinforcement histories.

Previous Research on Criterion Shift: How Do People Shift?

Instructed/Strategic Criterion Shifts

The most frequent way to induce recognition biases is by given subjects instructions that make it clear why one recognition decision should be favored over another. More specifically, these Instructed/Strategic criterion are most often induced by giving participants specific information about the relative proportions of old/new item distributions (Estes & Maddox, 1995; Rhodes & Jacoby, 2007a), warnings to avoid certain errors (Azimian-Faridani & Wilding, 2006),

or monetary incentives (Bowen et al., 2020; Van Zandt, 2000) that encourage either ‘old’ or ‘new’ recognition conclusions. Critically, these Instructed/Strategic paradigms make it clear why one versus the other class of response is to be preferred, and in general, produce robust decision biases in the observers.

Illustrating the use of explicit base rates, Aminoff et al. (2012) manipulated the proportion of old and new items in the test list and used two different colors to indicate potentially high and low base rate items. Participants were instructed to identify if the stimulus had been previously studied and to press the corresponding button for an old or new response. The instructions given to the participants clearly stated the weighted distributions in specific colors where one color was associated with a 70% probability that the stimulus was old and which color indicated a 30% probability that the stimulus was old. The results showed that the manipulation of the probability of the test item being old effectively influenced the placement of the decision criterion. This resulted in a liberal criterion being applied in the high-probability condition and a conservative criterion being applied in the low-probability condition.

As mentioned above, financial payouts can also induce recognition decision biases. For example, Bowen et al. (2020) manipulated the reward structure in a memory experiment where participants were shown indoor and outdoor pictures and instructed that these images would be tested for recognition memory. Following the study phase, subjects were then told that images from one category would be worth a high reward of $ 0.25, while images from the other category would be worth a low reward of $ 0.01 if correctly recognized on the memory test. Additionally, false alarms to either high or low reward category items resulted in the same loss of- $ 0.13. Results showed that the participants had higher hit and false alarm rates for high-reward

categories as compared to low-reward categories. However, there was no significant difference in memory sensitivity for high reward category items as compared to low reward category items. This suggests that the participants had a stronger tendency to respond "old" for the category with greater gains without any increase in sensitivity, thereby indicating a more liberal bias.

Incrementally Reinforced Criterion Shifts

Whereas Instructed/Strategic criterion shifts rely upon the subjects' explicit long-term goals of maximizing correct responding or rewards, it is also possible to induce decision biases by altering the balance of trial-wise feedback-based reinforcement. For example, Han and Dobbins (2009) implemented a false positive feedback (FPF) technique designed to subtly shift the relative probability of receiving positive reinforcement during 'old' versus 'new' responding. During the procedure, all correct recognition responses (hits and correct rejections) received positive reinforcement immediately following responses ("CORRECT"). However, negative feedback ("INCORRECT!") about erroneous responses (false alarms and misses) was tailored to ensure that one class of error less reliably evokes negative feedback by falsely providing positive feedback for some portion of that class of error. Thus, for example, whereas false alarms might all receive negative feedback, misses might instead receive 25% negative feedback combined with 75% **false** positive feedback. Because correct responses always receive positive feedback, this manipulation ensures that, on average, one class of response ('old' or 'new') is more likely to receive positive than negative feedback, even if a subject originally responds in an unbiased. This manipulation of false positive feedback was effective, inducing liberal decisions biases ('old' decisions more common than 'new' decisions) when false alarms received FPF, and the reverse when misses received FPF. Additionally, these induced decision biases remained during final recognition tests in which feedback was removed suggesting that subjects had learned to

consistently favor one versus the other recognition decision, and informal questioning of the subjects suggested that they were unaware that the feedback provided in earlier tests had altered their recognition decision tendencies.

Han and Dobbins (2009) suggested that the FPF decision biases may be induced implicitly because FPF only occurs on error trials, when subjects are assumed to be uncertain in their decision accuracy. If so, then it may be difficult to detect that the feedback is being altered, since even in the absence of FPF subjects will frequently guess correctly. The notion that the FPF manipulation may induce recognition decision biases even in the absence of explicit response strategies is also consistent with the demonstration, by Wixted and Gaitan (2002), that similar manipulations of reinforcement contingencies in pigeons induces recognition memory biases. Since these animals are unlikely to adopt strategic decision biases based on understanding how to maximize long-term gains, it suggests that the FPF effects in humans may likewise reflect a rudimentary and implicit learning mechanism (Wixted & Gaitan, 2002).

Reinforcement Induced Strategic Criterion Shifts

Critically, it is important to note feedback-induced recognition decision biases need not always be implicitly acquired. For example, Rhodes and Jacoby (2007) investigated decision biases in response to old item base rates that differed across spatial locations, with participants not informed about the base rate differences. Critically, the researchers also manipulated whether feedback was present or absent across conditions to test whether feedback was necessary for base rates to affect decision biases.

In Experiment 1, participants studied 72 items followed by a recognition test comprising these and 72 lures. Critically, the ratio of old to new items differed depending upon whether test items were positioned either on the left or right side of the screen. Whereas one location had 67% of the previously studied items, the other had 33%. Thus, during each test trial, participants were presented with a word in either the mostly old or mostly new locations (e.g., right, or left parts of the screen) and provided performance feedback after each trial to inform them of their response accuracy. Results indicated that participants were more conservative in their recognition decisions for mostly new locations but more liberal for mostly old locations, as expected. In Experiment 2, researchers aimed to examine the effect of awareness on participants. They asked participants to use different keys to input their answers. Specifically, participants were instructed to use the assigned keys on the left side of the keyboard when test items were displayed on the left side of the screen. And they had to use the assigned keys on the right side of the keyboard when the items appeared on the right side of the screen. After completing the experiment, participants were asked subjective awareness questionnaires related to the experimental manipulation. Accordingly, participants who used different keys showed more awareness about the location manipulation, and those who were aware exhibited a larger difference in estimated response criterion for mostly old versus mostly new locations than those who appeared to be unaware. Moreover, in Experiment 3, researchers aimed to examine the impact of feedback on criterion shift by providing feedback only during the first and last two tests of the 4 study-test cycles. Results showed that when feedback was provided a more liberal response criterion was used for items from predominantly old contexts than for items from predominantly new contexts. However, when feedback was removed, even after some participants had already completed two

blocks with feedback, the difference in response criterion for mostly old versus mostly new items was markedly diminished.

Overall, Rhodes and Jacoby (2007) demonstrated that decision biases were dependent on a couple of factors. First, decision biases were moderated by explicit awareness of the base rate differences. Second, it appears that feedback can serve to induce recognition biases by alerting the subject to an environmental imbalance that can be explicitly exploited to improve outcomes (i.e., that old items are more prevalent in one versus another screen location).

Is Recognition Evidence Global or Class Specific?

Signal detection theorists differ in their conceptualization of the evidence upon which recognition decisions are made. Under strength-based accounts, it is assumed that observers register a one dimensional familiarity signal reflecting the match between the memory traces and the stimulus features (Clark & Gronlund, 1996). Because these approaches often depend upon a feature matching mechanism to translate recognition probes into strength signals, they raise (although do not demand) the possibility that recognition signals for different classes of stimuli (e.g., pictures, or words) may not be evaluated by the same decision mechanisms. In other words, a feature-based strength account may be compatible with the idea that a learned decision bias for one class of stimuli (i.e., a criterion shift) may occur independently of the bias adopted for another class of stimuli, even when they are intermixed and presented in the same context. This, of course, would only be possible if the feature matching process yielded evidence signals that were somehow appreciably distinct for the different classes, and the learning mechanism responsible for any bias was separately applied to the potentially distinct evidence evoked by the two classes. For example, if an observer learned to be cautious, because of feedback, in

recognition decisions of face stimuli, this may have little or no impact the recognition decision process for word probes intermixed within the same list, provided that the perceived recognition of these two stimuli was distinctly evaluated. In contrast to strength models, some detection theorists contend that recognition decisions are not based on raw strength signals but instead are based on a statistical comparison of the likelihood of experiencing a given strength level under the hypothesis the probe was studied, versus the likelihood of experiencing that same strength level under the hypothesis the probe was not studied. This comparison is neatly summarized under signal detection by the ratio of the likelihood of these two values and referred to as a likelihood ratio decision variable (Macmillan & Creelman, 1990).

The benefit of the likelihood ratio model of evidence is that it is universal across stimulus classes and encoding conditions (see also, Glanzer et al., 2009). During testing, when the chances of targets and lures are equally likely, a subject can achieve maximum accuracy by maintaining a likelihood ratio criterion of 1. Critically, if subjects naturally base decisions on this abstracted statistical information, which may span stimulus classes, then they may experience recognition evidence for different classes of stimuli within the same list in the same manner. Under this conceptualization, faces and words for example, would evoke the same type of recognition evidence, namely evidence indicating the relative odds of prior encounter. Given this, when only one class of these intermixed stimuli is subjected to FPF (with the other half receiving fully correct feedback), the subjects may nonetheless learn to be generally cautious or lax during testing because the decision information accompanied by FPF is in the same format as that accompanied by wholly accurate feedback.

Current Study

The current study focuses on two interrelated questions regarding the FPF bias effect through two related experiments. First, we examined the stimulus class specificity of recognition decision biases by administering FPF to only one class of test stimuli in mixed lists containing both: in this case, words, or pictures. Critically, the non-targeted class always received veridical feedback. The key question we examine is whether subjects demonstrate a class-specific criterion shift, such that shifts only occur for the stimulus class receiving FPF, or instead whether subjects demonstrate a global criterion shift for all stimuli within the same test list. As noted above, this latter outcome would occur if familiarity or recognition evidence is a fairly abstracted statistical signal reflecting, for example, a likelihood ratio decision variable shared by both stimulus classes within a given temporal context. It is crucial to note that this experiment is mainly focused on criterion shifts. Therefore, we tried to ensure that there were no significant differences in accuracy across the bias groups/conditions because interpreting the cause of decision biases becomes challenging when accuracy also varies across bias conditions.

Secondly, we examine whether participants adopt **explicit** or implicit decision biases in response to FPF. Prior work provided 70% FPF for the targeted class of error (Han & Dobbins, 2009). In contrast, in the current study we reduced this probability to 50%. That is, we applied FPF to half of the false alarms in Liberal Bias conditions and to half of the misses in Conservative Bias conditions. The manipulation's rarity was predicted to make its detection unlikely and tests the boundary conditions of FPF learned biases. Moreover, for the first time, we extensively investigated subjects' awareness of the connection between the feedback and their recognition decisions to determine whether the observed decision biases, if any, were due to

explicit strategies or implicit learning. This was accomplished via a funnel questionnaire beginning with an open-ended question, followed by increasingly specific forced-choice questions designed to detect awareness of the FPF manipulation and its consequences.

Although exploratory, we anticipated that any biases induced by FPF would spread across materials presented in the same test list because subjects would come to generally favor either ‘old’ or ‘new’ responses without realizing (given the subtle nature of the feedback) that the reinforcement contingencies were being manipulated in response to one class of response tied to one class of stimulus. If, however, subjects did develop response biases selective to one class of stimulus it would suggest that either a) they became aware of the contingent nature of the feedback, or b) that biases may be acquired for classes of stimuli that do not fully overlap in features even in the absence of explicit awareness of how these biases are acquired.

Experiment 1

In Experiment 1, we aimed to investigate whether a probabilistic FPF procedure could induce selective decision biases for different stimulus categories. In the Picture Bias Groups, pictures received FPF manipulation, while words were provided with veridical feedback. Half of the subjects were randomly assigned to the Liberal Bias Groups, which meant that they received false positive feedback labeled as "CORRECT" for about 50% of their false alarms to pictures. The remaining responses received correct feedback. The other half of the subjects were randomly assigned to the Conservative Bias Groups, where approximately 50% of their incorrect new classifications of old items (misses) received false positive feedback. Similarly, the same manipulation was applied to the two, Word Bias Groups. Therefore, there were four groups in total: two groups receiving conservative or liberal feedback for words, and another two groups

receiving conservative or liberal feedback for pictures. Critically, for all groups, the first two study/test cycles employed FPF, whereas the final, third study/test cycle used standard recognition testing without any feedback. This was done to examine how durable any learned biases were. This design enables testing two questions within the Picture and Word Bias groups First, it enables the comparison of liberal and conservative FPF manipulations for the targeted stimulus class, to see whether it spreads to non-targeted stimulus class. Second, if different decision biases are present, whether subjects are aware of these learned biases.

Methods

Pre-registration and Data Availability

The materials and the data of all experiments are publicly available online at the Open Science Framework (OSF; osf.io/u45k7)

Participants

One hundred sixty undergraduates between 18 and 30 (M_{age} = 19.38, SD = 1.20) were recruited from Washington University in St. Louis (WashU) in return for partial course credit. Informed consent was obtained as required by WashU's institutional review board. Participants were fully debriefed on the nature of the feedback after the study. Because we used a unique manipulation only employed by Han and Dobbins (2008, 2009), the smallest reliable effect size in their study (η_p^2 = .08) was used to determine the necessary sample size for our experiment. G*Power software (Faul et al., 2007) power analyses indicated that a sample size of 143 would provide 80% power to detect a bias effect between groups with an alpha value of .05.

Materials

Word Stimuli

Three hundred sixty words were randomly drawn from a pool of 1216 words selected from the MRC Psycholinguistic Database (Wilson, 1988) with an average of 7.39 letters, 2.42 syllables, and an average log HAL frequency of 7.70. An equal number of targets (studied and tested items) and lures (novel test items) were randomly sampled from this list for each participant.

Picture Stimuli

Three hundred sixty images were randomly selected from the Bank of Standardized Stimuli (BOSS) project (Brodeur et al., 2014), which is a free normative database consisting of a standardized set of visual stimuli including animals, building infrastructures, plants, daily objects, and vehicles. Images with normative names matching any word stimuli were replaced.

Procedure

The experiment was programmed in PsychoPy (v2021.2.3)(Peirce et al., 2019). It consisted of three study/test cycles in which 60 intermixed words and pictures (120 study items) were studied and then tested for recognition with an equal number of new items in randomized way (240 test items). Old and new items were randomly selected by the program for each subject. Participants were randomly assigned to each Bias condition (Picture-Liberal, Picture-Conservative, Word-Liberal, Word-Conservative).

According to the FPF manipulation, in the Picture-Conservative Bias Groups subjects received FPF for picture probes that were misses (incorrect 'new' judgments). While in Picture-Liberal Bias Group subjects received FPF for picture probes that were false alarms (incorrect 'old' judgments) probabilistically. For these groups, word probe recognition responses always received correct feedback. Likewise, in the Word-Conservative Bias groups, subjects received FPF for word probes that were misses (incorrect new judgments) in Word-Liberal Bias Group

they received FPF for words that were false alarms (incorrect 'old' judgments). For these groups, picture probe recognition responses always received correct feedback.

Figure 1 depicts the general FPF manipulation for the liberal and conservative Picture Bias conditions, Tor the Picture-Conservative Bias Group, the 'snowman' picture was studied, but the fictive subject incorrectly reported it as 'New.' Despite this, the computer provided feedback indicating the response was correct and this reinforcement of 'New' conclusions is predicted to increase the tendency to claim items are new, yielding a conservative decision bias. In contrast, the Picture-Liberal group received FPF for false alarms to new pictures. Both groups receive correct feedback given to words, with the key question being whether an induced bias to the FPF class (in this case pictures) spreads to the other class, in this case words.

During the study phase, participants were randomly presented with a word or picture and an orienting question during each trial to promote encoding. Different encoding tasks were chosen in an attempt to achieve a similar level of recognition performance, due to the expected superior recognition memory performance of pictures [(picture superiority effect;(Mintzer & Snodgrass, 1999)]. Words were accompanied by a prompt asking whether the word is pleasant ("Pleasant?"). Meanwhile, pictures were accompanied by a prompt asking whether the thing in the picture is alive ("Living thing?"). Regardless of the item type, both encoding questions were answered as "YES" or "NO" by using the assigned keyboard buttons (A = "YES," L = "NO"). Responding was self-paced.

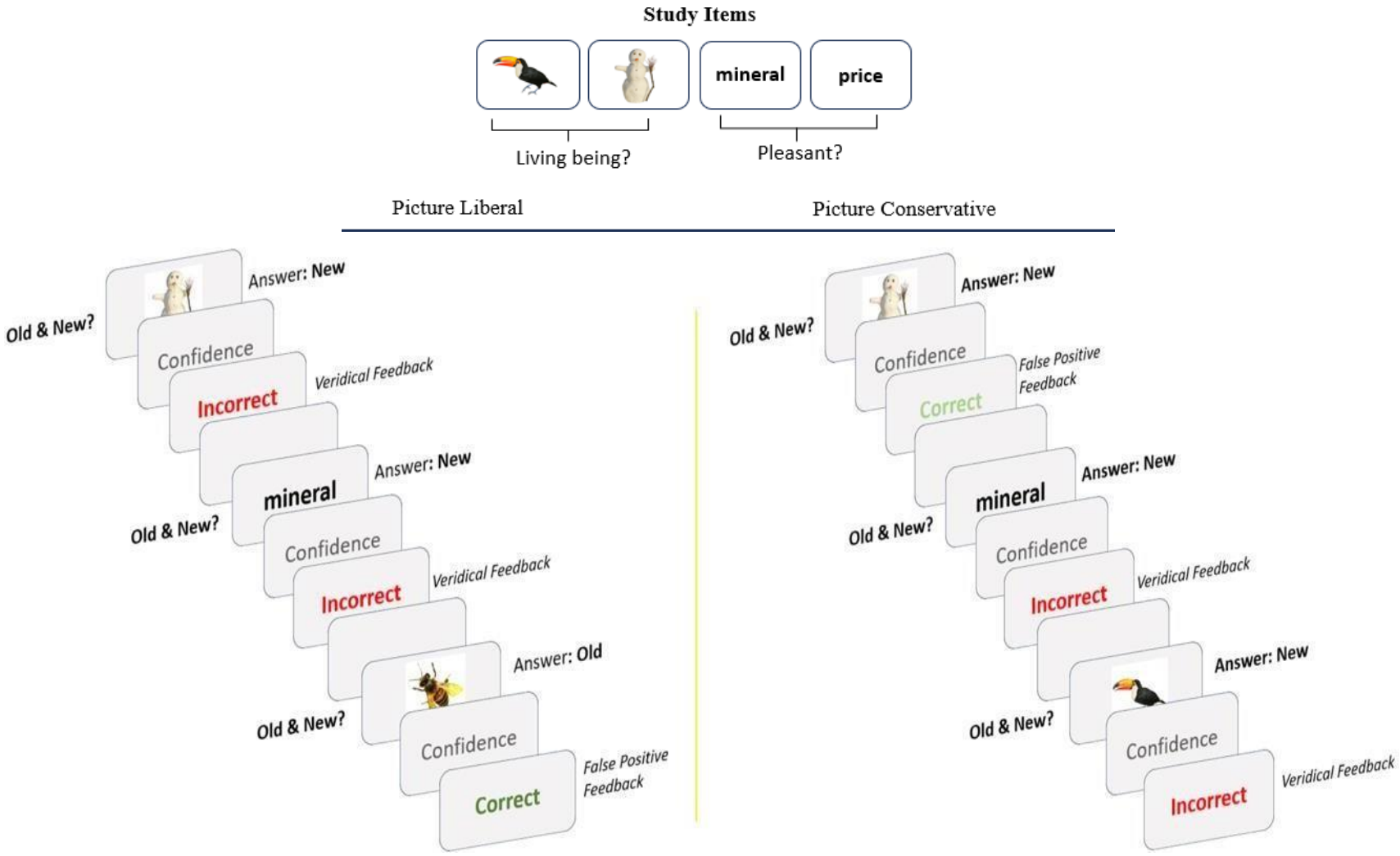
During test phase, 120 targets and 120 lures for word and picture stimuli were randomly intermixed, and participants indicated whether each item was old or new (A = Old, L = New). Following this, they reported decision confidence ("1 = Low", "2 = Medium," or "3 = High"). These responses were self-paced. Immediately after the confidence report, the subject received

feedback. For positive feedback, “CORRECT” appeared in green, and for negative feedback or “INCORRECT,” appeared in red, remaining on the screen for 1 second. Within the Picture and Word groups, this feedback contained the FPF manipulation, which was restricted to the relevant stimulus class and was either Liberal (FPF to 50% of false alarms only) or Conservative (FPF to 50% of misses only). All other feedback was veridical (Figure 1).

Following the final recognition test, participants completed a Subjective Awareness Questionnaire (see Appendix A). First, they were asked to report in an open-ended fashion the impact of the feedback on their recognition decision-making, typing their answers into a small response box. They then answered force-choice questions designed to probe their awareness of the effects of the feedback. These consisted of accuracy and bias awareness questions that contained two parts. For accuracy, part 1 asked whether they thought the feedback a) increased accuracy, b) decreased accuracy, or c) did not affect accuracy. If they indicated a) or b), they were then asked in part II whether this effect was restricted to a) words, b) pictures, or c) both materials. For bias questions, part 1 asked whether the feedback a) increased the tendency to respond ‘old,’ b) decreased the tendency to respond ‘old,’ or c) did not affect the tendency to respond ‘old.’ If they indicated a) or b), they were then asked in part II whether this effect was restricted to a) words, b) pictures, or c) both materials. The order of accuracy versus bias questions were randomized for each participant.

Figure 1

The experimental process display of FPF manipulation in the recognition test.



Note. The image displays the step-by-step procedure that was followed in the test phases of Experiment 1 and Experiment 2. The left side of the image represents the Picture Liberal Bias Group, where participants were provided with False Positive Feedback (FPF) only for their *false alarm* responses to pictures in order to encourage them to make more false alarms. The right side of the image depicts the Picture-Conservative Bias Group, where participants were given FPF only for their *miss* responses to pictures in order to encourage them to make more misses.

Results

Response Accuracy (*d'*)

Because criterion differences are difficult to interpret when accuracy levels also differ, I begin by comparing accuracy across Liberal and Conservative Bias Groups, restricting the analyses to groups where FPF was applied to pictures, or groups where it was applied to words. Thus, the first analysis focuses on the groups that received FPF feedback for pictures in Experiment 1, using a mixed ANOVA with factors of Bias Group (liberal or conservative) and Stimulus Class (pictures [the FPF stimuli] or word [neutral feedback stimuli]) and study/test Cycle (1, 2, or 3). The DV was the *d'* sensitivity measure. Please note that all the analyses below were performed using adjusted values[1].

Table 1 presents data for Picture Bias Groups, with the hit rates (HR), false alarm rates (FAR), sensitivity (*d'*), response bias (C) and FPF counts. Figure 2A shows the sensitivity (*d'*) data for the Picture Bias Groups when responding to picture and word stimuli for Experiment 1. Across the three tests, the figure suggests that subjects were more accurate in for word stimuli, which received fully veridical feedback, compared to picture stimuli for both liberal and conservative Bias Groups. These impressions were evaluated using a Bias Group (liberal vs conservative) by Stimulus Class (pictures vs words) by study/test Cycle (1, 2, or 3) mixed ANOVA. Bias Group was a between-subjects measure. This analysis revealed a main effect of Stimulus Class ($F(1, 78) = 26.24$, $MSE = .50$, $p < .001$, $\eta_p^2 = .25$), no main effect of Bias Group ($F(1,78) = 3.39$, $MSE = 2.94$, $p = .069$, $\eta_p^2 = .042$), and a main effect of Cycle ($F(2,156) =$

[1] The current study has used corrected *d'* and *C* measures for all its analyses. This correction was applied to the raw data, where the hit rates were 1 and false alarm rates were 0. The hit rates of 1 were changed to (n-0.5)/n, and the false alarm rates of 0 were changed to 0.5/n. Here, n represents the number of signal or noise trials (i.e., 60), as stated in Macmillan and Kaplan (1985).

117.34, *MSE* = .60, $p < .001$, $\eta_p^2 = .60$). However, these effects were conditioned by a two-way interaction between Stimulus Class and Cycle ($F(2,156) = 21.57$, *MSE* = .14, $p < .001$, $\eta_p^2 = .22$). No other interactions were significant.

Table 1

Hit rates, false alarm rates, accuracy, criterion, and number of FPF trials for both words and pictures in Picture Bias Groups during Tests 1, 2, 3 in Experiment 1

Liberal Bias Groups

Stimulus	Test	Hit Rate	FA Rate	*d'*	*C*	FPF trials
Words	Test 1	.92(.08)	.13(.11)	2.86(.92)	-.11(.31)	N/A
	Test 2	.88(.12)	.25(.21)	2.13(1.03)	-.25(.46)	N/A
	Test 3	.80(.14)	.28(.16)	1.65(.90)	-.15(.33)	N/A
Pictures	Test 1	.81(.11)	.12(.07)	2.24(.63)	.14(.29)	8.9(4.12)
	Test 2	.79(.13)	.23(.17)	1.74(.84)	-.04(.45)	13.25(6.06)
	Test 3	.74(.16)	.22(.13)	1.60(.80)	.07(.33)	N/A

Conservative Bias Groups

Stimulus	Test	Hit Rate	FA Rate	*d'*	*C*	FPF trials
Words	Test 1	.93(.06)	.09(.07)	3.03(.71)	-.08(.30)	N/A
	Test 2	.88(.10)	.17(.12)	2.44(.86)	-.12(.31)	N/A
	Test 3	.83(.13)	.22(.15)	1.96(.98)	-.10(.30)	N/A
Pictures	Test 1	.84(.09)	.10(.06)	2.47(.60)	.16(.26)	7.3(3.74)
	Test 2	.76(.15)	.11(.08)	2.13(.77)	.26(.34)	11.5(6.05)
	Test 3	.73(.18)	.14(.11)	1.92(.92)	.24(.34)	N/A

Note. The table includes corrected d' and c measures. Standard deviations in parentheses.

The interaction between Stimulus Class and Cycle was decomposed by comparing sensitivity for words and pictures at each cycle in Picture Bias Groups. During the first test sensitivity was significantly greater for words (*M* = 2.95, *SD* = .82) than pictures (*M* = 2.36, *SD* = .62) ($t(79) = 6.93$, $d = .77$, $p < .001$). The advantage for words (*M* =2.29, *SD* = .96), though

smaller, compared to pictures (M = 1.93, SD =.83) continued in the second test ($t(79) = 4.32$, d = .48, $p < .001$). However, by the third test, sensitivity for words (M = 1.81, SD = .95) and pictures (M =1.76, SD = .87) was similar ($t(79) = .67$, d = .07, p = .505). Therefore, the interaction occurred because sensitivity decreased for both types of stimuli. However, the decline was greater for words, which had a much higher starting sensitivity than pictures. Importantly, the Bias Groups exhibited similar sensitivity, and this factor did not interact with any other factors in the design. As a result, the bias analyses presented below are not affected by any significant differences in sensitivity.

Table 2

Hit rates, false alarm rates, accuracy, criterion, and number of FPF trials for both words and pictures in Word Bias Groups during Tests 1, 2, 3 in Experiment 1

Liberal Bias Groups

Stimulus	Test	Hit Rate	FA Rate	*d'*	*C*	FPF trials
Words	Test 1	.93(.07)	.11(.11)	3.13(.95)	-.13(.29)	6.23(5.18)
	Test 2	.91(.10)	.18(.19)	2.63(1.07)	-.24(.40)	8.21(6.88)
	Test 3	.87(.13)	.25(.18)	2.13(1.12)	-.27(.32)	N/A
Pictures	Test 1	.86(.10)	.11(.08)	2.54(.79)	.08(.25)	N/A
	Test 2	.82(.12)	.16(.14)	2.18(.94)	.01(.32)	N/A
	Test 3	.78(.16)	.16(.15)	2.08(1.08)	.11(.32)	N/A

Conservative Bias Groups

Stimulus	Test	Hit Rate	FA Rate	*d'*	*C*	FPF trials
Words	Test 1	.90(.12)	.12(.11)	2.82(.95)	-.09(.32)	7.21(6.08)
	Test 2	.85(.18)	.22(.16)	2.21(1.22)	-.17(.37)	11.08(7.28)
	Test 3	.81(.18)	.28(.17)	1.75(1.18)	-.21(.32)	N/A
Pictures	Test 1	.81(.11)	.13(.13)	2.23(.82)	.16(.30)	N/A
	Test 2	.76(.14)	.18(.13)	1.85(.92)	.13(.29)	N/A
	Test 3	.72(.18)	.22(.18)	1.58(1.06)	.11(.32)	N/A

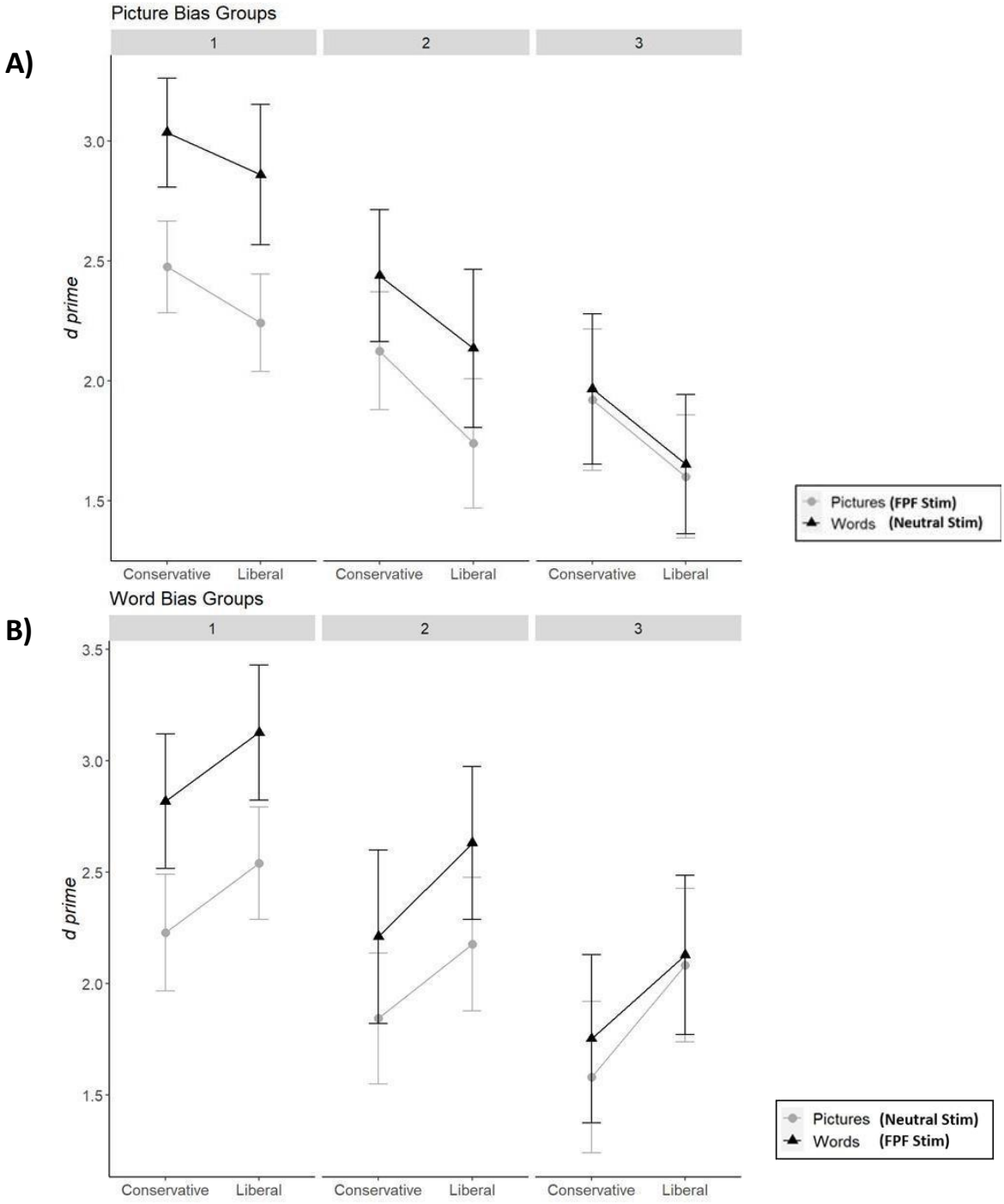
Note. The table includes corrected d' and c measures. Standard deviations in parentheses.

The Table 2 presents analogous data for Word Bias Groups, with the hit rates (HR), false alarm rates (FAR), sensitivity (*d'*), response bias (C) and FPF counts. The Figure 2B shows the sensitivity data for the Word Bias Groups for word responding and picture responding for Experiment 1. Across the three tests, it suggests that subjects were more accurate in for word stimuli, which received FPF, compared to picture stimuli within the two Bias Groups. These impressions were evaluated using a Bias Group (liberal vs conservative) by Stimulus Class (pictures vs words) by study/test Cycle (1, 2, or 3) mixed ANOVA. The ANOVA did not show a main effect of Bias Group ($F(1, 78) = 3.38$, $MSE = 4.99$, , $p = .0698$, $\eta_p^2 = .042$), but did demonstrate main effects of Stimulus Class ($F(1, 78) = 38.15$, $MSE = .43$, $p < .001$, $\eta_p^2 = .33$) and Cycle ($F(2, 156) = 91.89$, $MSE = .28$, $p < .001$, $\eta_p^2 = .54$). These effects were again conditioned with two-way interaction between Stimulus Class and Cycle ($F(2, 156) = 22.35$, $MSE = .11$, $p < .001$, $\eta_p^2 = .22$).

The interaction between Stimulus Class and Cycle was decomposed by comparing sensitivity for words and pictures at each cycle in the Word Bias Groups. During the first test, sensitivity was higher for words ($M = 2.97$, $SD = .95$) than pictures ($M = 2.38$, $SD = .81$) ($t(79) = 7.63$, $p < .001$, $d = .85$). The advantage for words ($M = 2.42$, $SD = 1.16$) continued in the second test (picture; $M = 2.01$, $SD = .94$, $t(79) = 5.58$, $p < .001$, $d = .62$). However, by the third test, sensitivity for words ($M = 1.94$, $SD = 1.16$) and pictures ($M = 1.83$, $SD = 1.09$) was similar ($t(79) = 1.62$, $p = .109$, $d = .18$). Thus, the interaction resulted because while sensitivity declined for both stimulus classes, this decline was more dramatic for words, which started at much higher sensitivity, than pictures. Critically, the Bias Groups had similar sensitivity and this factor did not interact with the other factors of the design. Hence the bias analyses reported below are not confounded with reliable differences in sensitivity because of the FPF manipulation.

Figure 2

Mean Accuracy (d prime) for Picture and Word Bias Groups in Experiment 1



Note. The figure shows the mean accuracy rate for pictures and words. Upper part compares Conservative and Liberal Bias Groups in Picture Bias Groups in Experiment 1. Lower part compares the Conservative and Liberal Bias Groups in Word Bias Groups in Experiment 1. Numbers at the top shows each test Cycle (1,2, or 3). Error bars represent 95% confidence intervals.

Response Bias (C-bias)

Potential differences in response bias (*C*) were analyzed as above, by restricting the analysis to Bias Groups either receiving FPF for picture or FPF for word. Again, the goal was to determine if biases were learned for the FPF target stimuli, and if so, whether these spread to the stimuli receiving neutral feedback. Thus, the first analysis focuses on the groups that received liberal or conservative FPF feedback for pictures in Experiment 1 using a mixed ANOVA with factors of Bias Group (liberal or conservative) and Stimulus Class (pictures [the FPF stimuli] or word [neutral feedback stimuli]) and study/test Cycle (1, 2, or 3). The DV was the C bias measure.

The Figure 3A shows the bias data for the Picture Bias Groups for picture and word stimuli for Experiment 1. Across the three tests, the Figure suggests that subjects became more conservative for pictures when they received FPF for picture probe misses (Picture-Conservative) than when they received FPF picture probe false alarms (Picture-Liberal). In contrast, decision bias for the intermixed words (all of which received correct feedback) appeared stable across the three tests. These impressions were evaluated using a Bias Group (liberal vs conservative) by Stimulus Class (pictures vs words) and study/test Cycle (1, 2, or 3) mixed ANOVA. Bias Group was a between-subjects measure. This analysis revealed main effects of Bias Group ($F(1,78) = 5.75$, $MSE = .29$, $p = .019$, $\eta_p^2 = .07$), and Stimulus Class ($F(1, 78) = 58.08$, $MSE = .15$, $p < .001$, $\eta_p^2 = .43$) but no main effect of Cycle ($p = .130$). However, these effects were conditioned by a two-way interaction between Bias Group and Cycle ($F(2,156)= 3.815$, $MSE = .09$, $p = .020$, $\eta_p^2 = .05$). Critically, the three-way interaction was also significant ($F(2,156) = 3.07$, $MSE = .03$, $p = .049$, $\eta_p^2 = .04$).

To decompose the three-way interaction, we conducted Bias Group by Stimulus Class mixed ANOVAs for each cycle.

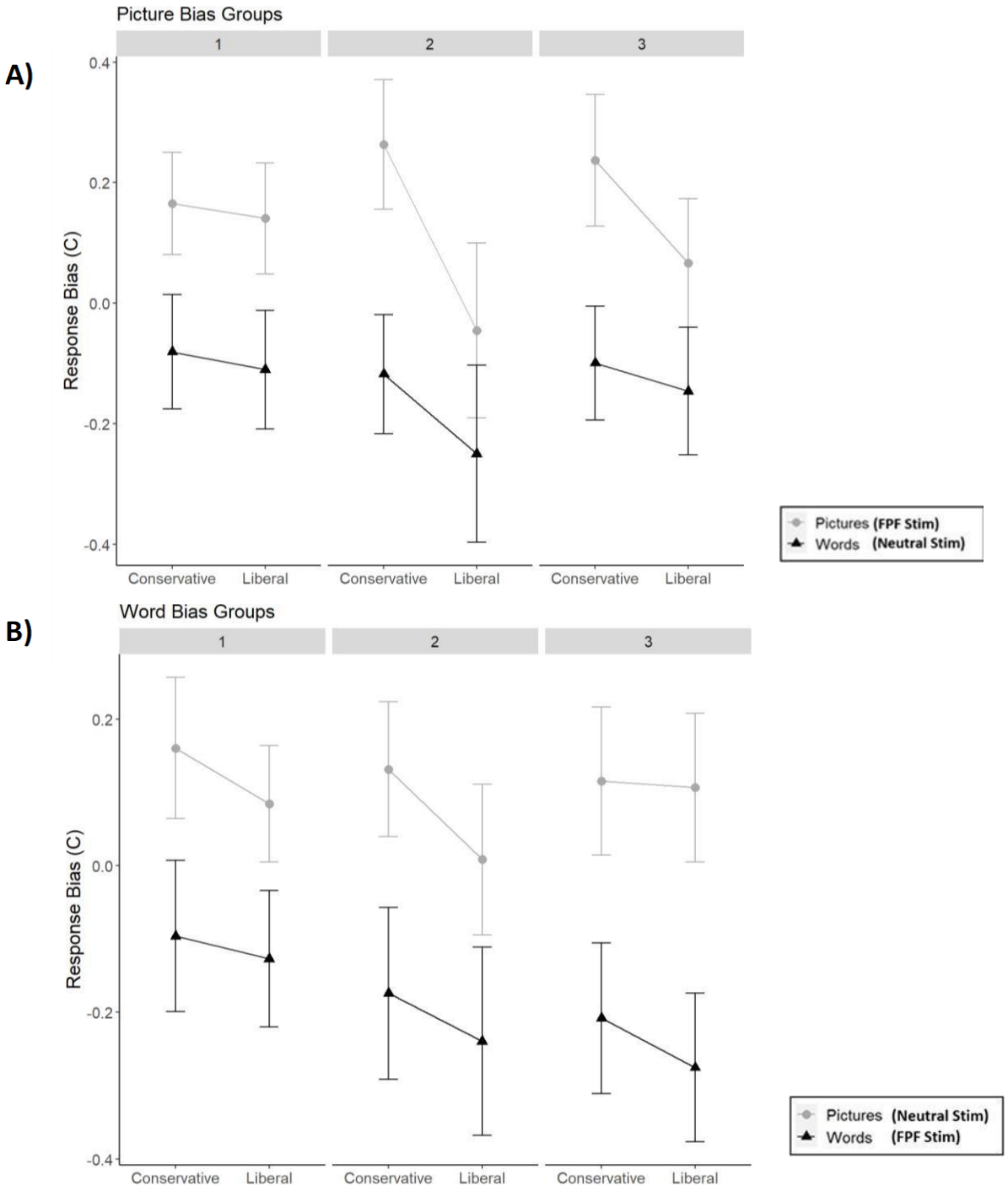
For Cycle 1 (Figure 3A) there was no main effect of Bias Group ($F(1, 78) = .329$, $MSE = .09$, $p = .568$, $\eta_p^2 = .00$) but there was a main effect of Stimulus Class ($F(1,78) = 31.49$, $MSE = .08$, $p < .001$, $\eta_p^2 = .29$). Additionally, there was no interaction between Bias Group and Stimulus Class in Cycle 1 ($F(1,78) = .003$, $MSE = .06$, $p = .957$, $\eta_p^2 = .00$). Turning to Cycle 2 the main effect of the Bias Group ($F(1,78)=8.12$, $MSE = .24$, $p = .006$, $\eta_p^2 = .09$) and the main effect of the Stimulus Class ($F(1,78)=47.33$, $MSE = .24$, $p < .001$, $\eta_p^2 = .38$) were significant. Critically, the interaction between Stimulus Class and Bias Group was also significant ($F(1,78)=4.31$, $MSE = .07$, $p = .041$, $\eta_p^2 = .05$) with pairwise comparisons showing that the interaction resulted because the Liberal and Conservative Bias Groups for differed for pictures, $t(78)=3.46$, $d = .77$, $p < .001$, but not for words ($p = .135$). Turning to Cycle 3, which contained no feedback, the main effect of Bias Groups approached significance, $F(1,78)=3.09$, $MSE = .15$, $p = .083$, $\eta_p^2 = .04$, and the main effect of Stimulus Class was significant ($F(1,78)= 51.29$, $MSE = .06$, $p < .001$, $\eta_p^2 = .40$). However, there was no interaction occurred between Stimulus Class and Bias Groups ($F(1,78)= 2.61$, $MSE = .06$, $p = .110$, $\eta_p^2 = .03$). Despite the failure to observe an interaction, I nonetheless performed pairwise comparisons to see if the pattern observed in Cycle 2 recurred in Cycle3. These pairwise comparisons a significant difference between the Liberal and Conservative Bias Groups for FPF for pictures ($t(78)= 2.26$, $d = .51$, $p = .027$), but no difference for words ($p = .507$). These results suggested that there was a response bias across the Bias Groups that was restricted to the picture stimuli that received FPF. This this effect did not transfer to the word stimuli that received wholly correct feedback.

Overall, the results indicate that there was a class-specific learned FPF bias developed for picture recognition judgments that did not transfer to word recognition judgments, and this learned bias persisted even when feedback was completely removed in the final test.

The Figure 3B shows the bias data for the Word Bias Groups for word and picture responding for Experiment 1. These data were analyzed with the analogous ANOVA used for the Picture Groups, consisting of the factors of Bias Group, Stimulus Class, and study/test Cycle (1, 2, or 3). It demonstrated a main effect of Stimulus Class ($F(1,78) = 90.53$, $MSE = .11$, $p < .001$, $\eta_p^2 = .54$) with subjects being more conservative for pictures than words and the main effect of the Cycle ($F(2,156) = 3.54$, $MSE = .08$, $p = .031$, $\eta_p^2 = .43$). It also demonstrated Stimulus Class by Cycle interaction ($F(2,156) = 4.02$, $MSE = .04$, $p = .020$, $\eta_p^2 = .05$) which occurred because the stimulus based biases modestly increased across the cycles (Figure 3B). Critically, however, the Bias Group factor was not significant ($F(1,78) = 1.68$, $MSE = .27$, $p = .198$, $\eta_p^2 = .02$), nor did it interact with the other two factors. Hence, unlike the picture groups, subjects receiving FPF for words failed to demonstrate any reliable FPF-induced biases. In Experiment 2, I attempt to address this null finding by reducing the performance/sensitivity of word recognition based on the assumption that high performance for these materials precluded sufficient FPF.

Figure 3

Mean Response Bias (C) for Picture and Word Bias Groups in Experiment 1



Note. The figure shows the mean response bias rate for pictures and words. Upper part compares Conservative and Liberal Bias Groups in Picture target groups in Experiment 1. Lower part compares the Conservative and Liberal Bias Groups in Picture target groups in Experiment 1. Numbers on the top shows each test Cycle (1,2, or 3). Error bars represent 95% confidence intervals.

Awareness Questionnaire

The awareness questionnaire was given at the end of Cycle 3 and designed to probe subjects' awareness of the influence of feedback on their performance. The open-ended responses regarding the influence of feedback are available online (OSF; osf.io/u45k7). Both authors inspected these for any evidence of awareness of the biasing effects of the feedback (an illustration of several is provided in Table 4). Neither author found any evidence that subjects thought the feedback caused them to either reduce or increase their tendency to claim items were recognized for either pictures, words, or the materials in general. Instead, most subjects appeared to believe that the feedback was present to either increase the accuracy of responding or influence the confidence with which responses were given. Nonetheless, because a bias effect may be difficult to describe, the follow-up forced-choice questions provide a more direct test of awareness. This analysis is restricted to the liberal and conservative groups receiving FPF for pictures (Figure 3A) since only these demonstrated learned biases.

Table 1 shows the response counts for the funnel forced choice questions. For part I of the accuracy question, and the responses were clearly not randomly distributed across the three options ($\chi^2(2) = 52.23$, $p < .001$). Instead, most subjects (71%) incorrectly concluded that the feedback made their recognition decisions more accurate which is reliably higher than the chance rate of 33% ($\chi^2(2) = 51.75$, $p < .001$) and consistent with the open-ended responses.

Turning to the bias question, responses again were not distributed randomly (Table 3; $\chi^2(2) = 22.90$, $p < .001$), with a majority of subjects (58%) indicating that the feedback had no influence on their tendency to respond old, again different from a chance rate of 33% ($\chi^2(1) = 21.02$, $p < .001$). This is of course, generally incorrect, as the ANOVAs and Figure 3A shows the influence effect of FPF on decision bias for the picture stimuli. For the 34 subjects that indicated

that the feedback influenced their tendency to respond 'old', only 17 correctly selected the option consistent with their FPF Bias Group assignment (i.e., the more option for the liberal group and the less option for the conservative group), which is exactly the percentage that would be achieved by chance responding. Of these 17, we further examined their response to the part II question regarding whether the bias extended to pictures, words, or both types of stimuli. Only 5 of these 17 correctly indicated that the feedback was designed to selectively bias responses to pictures, with 6 incorrectly endorsing words and the remaining 6 endorsing both stimuli as biased. This distribution of responses is consistent with chance selection ($\chi^2(2) = 00.12$, $p = .943$). Moreover, the initial open-ended responses of the subjects correctly selecting pictures as biased (Table 4) do not suggest they initially thought the feedback was biased.

Table 3

Response counts for the Part I and Part II Questions for Picture Bias Groups in Experiment 1

Part I Questions (Purpose of the feedback?)		Part II Questions (Target class of the feedback?)		
Accuracy Question		**Words**	**Pictures**	**Both**
Increase accuracy	57	6	5	46
Decrease accuracy	9	0	0	9
No effect	14	NA	NA	NA
Bias Question		**Words**	**Pictures**	**Both**
Increase old responses	22	7	7	8
Decrease old responses	12	2	6	4
No effect	46	NA	NA	NA

Note. Part I Questions consist of a blend of accuracy and bias questions, which were presented in a randomized order. Part II Questions inquire about the specificity of the given feedback with respect to stimulus classes.

Thus, only 5 of the 80 subjects selected answers to the bias questions that would indicate awareness of the current behavioral effects (see Table 4). When we removed these five subjects from the picture Bias Groups during the analysis of FPF effects on picture responses (Figure 3A) the Bias Group by Cycle interaction was replicated ($F(2,146) = 5.78$, $MSE = .06$, $p = .004$, $\eta_p^2 = .073$) with the post-hoc pairwise analysis again demonstrating no reliable bias differences in Cycle 1 ($p = .991$), and a reliable bias difference across the Liberal and Conservative Bias Groups in Cycle 2 ($t(73) = 2.91$, $d = .67$, $p = .005$). In the final test the groups numerically trended towards different biases but the difference was not significant ($t(73) = 1.67$, $d = .39$, $p = .099$). Overall, while removing subjects may have reduced power somewhat, the data still indicate that subjects who show minimal measurable awareness, demonstrate learned FPF biases. In the General Discussion, we further consider why detecting the FPF influence would be particularly difficult.

Table 4

Open-ended Question responses of the potentially aware participants in Experiment 1

Subject	Bias Condition	Response to the open-ended question: "What was the purpose of the feedback?"
2009	Picture Liberal	To see if I am actually remembering or guessing
2007	Picture Conservative	I think the purpose was in order to guide or influence us to correct answer
2082	Picture Liberal	Whether receiving feedback helps with memory recall
2097	Picture Conservative	If feedback increases or decreases accuracy in recognition
2131	Picture Conservative	Letting me know if my confidence was accurate

Note. The table includes the responses of the potentially aware subjects who gave correct answer to Part I Questions bias question, and following Part II Question inquire about the specificity of the given feedback with respect to stimulus classes. Table only includes responses in Picture Bias Groups due to only reliable bias effect observed in those groups.

Discussion

The data for Experiment 1 suggested that FPF does induce decision biases and that these biases do not spread to intermixed materials not receiving FPF. However, we only obtained this selective effect when pictures were the targeted class. In the Word Bias Groups, where words were the targeted class, no FPF effects of any kind were observed. This may have reflected the deep processing conducted on words which generally had a higher hit rate than pictures (Table 1, and Table 2). Since FPF depends on the commission of errors, ceiling level effects in hits or correct rejections can limit the impact of the procedure. To address this, we attempted to bring word performance down to the level of picture performance by using a shallower processing task.

Based on the responses to an open-ended question in the Awareness data, it seems that most participants thought that the feedback was given to improve the accuracy of their responses or affect their confidence in answering. However, when asked explicitly about the purpose of the feedback through follow-up forced-choice questions, the answers were inconsistent. This indicates that the participants were not aware that the feedback was causing a decision bias, nor did they realize that this bias only applied to the picture stimuli.

Experiment 2

The purpose of Experiment 2 was to investigate whether the modest evidence for a stimulus selective, and implicitly acquired bias for pictures in the picture groups of Experiment 1, might be an artifact of the generally higher discrimination subjects demonstrated for words than pictures. Additionally, we wondered whether the failure to observe stimulus specific biases to words when they were the target of the FPF manipulation, might also be tied to their generally

higher discriminability compared to pictures. To address this, the word encoding task was modified in Experiment 2 to decrease recognition performance in hopes of rendering it similar to the discrimination performance for pictures.

Methods

Participants

One hundred fifty-six undergraduates between 18 and 30 (M_{age} = 19.84, *SD* = 1.36) were recruited from Washington University in St. Louis (WashU) in return for partial course credit. Informed consent was obtained as required by WashU's institutional review board. Participants were fully debriefed on the nature of the feedback after the study.

Materials and Procedure

Experiment 2 is the same as Experiment 1 with the materials and the procedure used, except the modified encoding question for word stimuli. In Experiment 2, the encoding question for words was changed from "pleasant?" to "two syllables?" with the aim of decreasing word accuracy and bringing it down to the level of picture discriminability. As expected, this modification resulted in lower *d'* for words (1.29), but unfortunately, as shown below, this was now reliably lower than *d'* for pictures (1.92).

Results

Response Accuracy (*d'*)

Because criterion differences are difficult to interpret when accuracy levels also differ, I begin by comparing accuracy across Liberal and Conservative Bias Groups, separately for groups in which FPF was applied to pictures versus words. Thus, the first analysis focuses on the groups that received FPF feedback for pictures in Experiment 2, using a mixed ANOVA with factors of Bias Group (liberal or conservative) and Stimulus Class (pictures [the FPF stimuli] or

word [neutral feedback stimuli]), and study/test Cycle (1, 2, or 3). The DV was the *d'* sensitivity measured.

Table 5

Hit rates, false alarm rates, accuracy, criterion, and number of FPF trials for both words and pictures in Picture Bias Groups during Tests 1, 2, 3 in Experiment 2

Liberal Bias Groups

Stimulus	Test	Hit Rate	FA Rate	*d'*	*C*	FPF trials
Words	Test 1	.77(.10)	.21(.12)	1.70(.58)	.05(.34)	N/A
	Test 2	.77(.09)	.34(.16)	1.23(.58)	-.15(.29)	N/A
	Test 3	.75(.10)	.38(.15)	1.04(.55)	-.20(.28)	N/A
Pictures	Test 1	.83(.10)	.10(.06)	2.42(.69)	.15(.27)	8.46(4.65)
	Test 2	.79(.12)	.18(.12)	1.90(.68)	.05(.30)	12.07(5.98)
	Test 3	.76(.14)	.20(.13)	1.71(.83)	.06(.30)	N/A

Conservative Bias Groups

Stimulus	Test	Hit Rate	FA Rate	*d'*	*C*	FPF trials
Words	Test 1	.77(.10)	.22(.12)	1.61(.49)	.03(.30)	N/A
	Test 2	.76(.12)	.33(.16)	1.26(.63)	-.14(.33)	N/A
	Test 3	.73(.15)	.37(.11)	1.02(.55)	-.16(.31)	N/A
Pictures	Test 1	.82(.10)	.13(.08)	2.19(.69)	.10(.24)	9.60(5.27)
	Test 2	.78(.14)	.19(.16)	1.92(.95)	.05(.40)	12.5(8.02)
	Test 3	.76(.15)	.19(.13)	1.77(.87)	.08(.30)	N/A

Note. The table includes corrected d' and c measures. Standard deviations in parentheses.

Table 6

Hit rates, false alarm rates, accuracy, criterion, and number of FPF trials for both words and pictures in Word Bias Groups during Tests 1, 2, 3 in Experiment 2

Liberal Bias Groups

Stimulus	Test	Hit Rate	FA Rate	*d'*	*C*	FPF trials
Words	Test 1	.78(.12)	.19(.12)	1.82(.60)	.09(.36)	12.54(5.04)
	Test 2	.78(.15)	.36(.16)	1.30(.65)	-.24(.40)	16.92(6.41)
	Test 3	.74(.17)	.40(.19)	1.02(.73)	-.22(.44)	N/A
Pictures	Test 1	.83(.11)	.13(.09)	2.27(.77)	.06(.33)	N/A
	Test 2	.76(.17)	.17(.09)	1.93(.88)	.08(.38)	N/A
	Test 3	.73(.18)	.18(.11)	1.74(.93)	.18(.31)	N/A

Conservative Bias Groups

Stimulus	Test	Hit Rate	FA Rate	*d'*	*C*	FPF trials
Words	Test 1	.74(.12)	.17(.09)	1.69(.60)	.16(.27)	13.67(5.52)
	Test 2	.70(.16)	.26(.13)	1.29(.60)	.07(.36)	15.72(6.37)
	Test 3	.70(.18)	.32(.18)	1.15(.85)	-.02(.40)	N/A
Pictures	Test 1	.85(.10)	.11(.06)	2.48(.68)	.07(.26)	N/A
	Test 2	.77(.15)	.18(.12)	1.83(.79)	.08(.28)	N/A
	Test 3	.74(.18)	.19(.16)	1.78(1.03)	.14(.35)	N/A

Note. The table includes corrected d' and c measures. Standard deviations in parentheses.

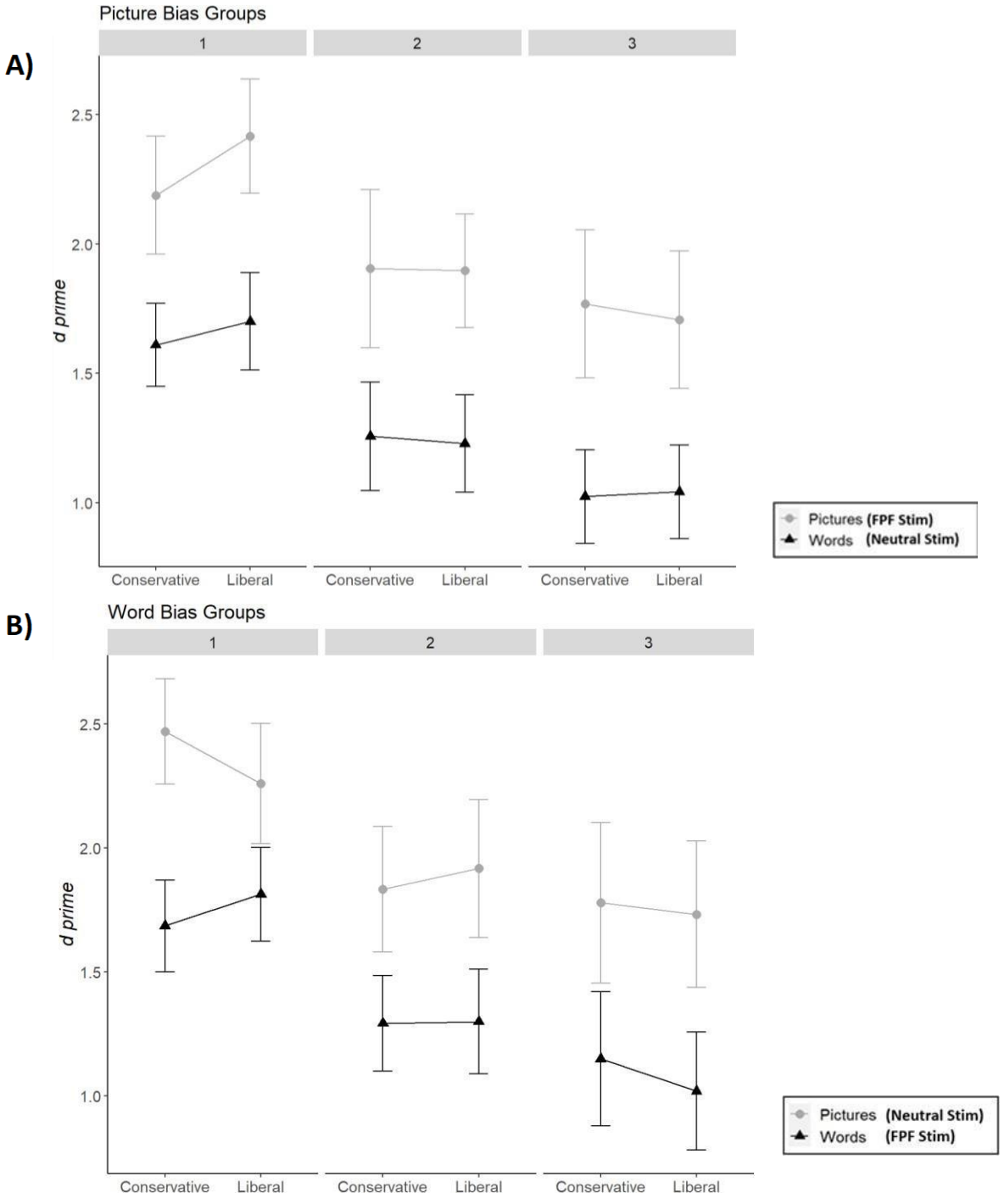
Table 5 presents data for Picture Bias Groups, with the hit rates (HR), false alarm rates (FAR), sensitivity (*d'*), response bias (*C*) and FPF counts. The Figure 4A shows the sensitivity data for the Picture Bias Groups for the picture and word stimuli for Experiment 2. The Figure suggests that picture recognition accuracy was higher than the accuracy for the intermixed words (all of which received correct feedback) through three cycles, in both Picture Conservative and Picture Liberal Bias Groups. These impressions were evaluated using a Bias Group (liberal vs conservative) by Stimulus Class (pictures vs words) and study/test Cycle (1, 2, or 3) mixed

ANOVA. Bias Group was a between-subjects measure. This analysis revealed main effects of Stimulus Class ($F(1,75) = 119.17$, $p < .001$, MSE = .43, $\eta_p^2 = .61$) and main effect of Cycle ($F(2,150) = 74.80$, $p < .001$, MSE = .19, $\eta_p^2 = .50$) but no main effect of Bias Group ($F(1,75) = .106$, $p = .746$, MSE = 1.76, $\eta_p^2 = .001$). There were no reliable interactions.

Table 6 presents data for Word Bias Groups, with the hit rates (HR), false alarm rates (FAR), sensitivity (d'), response bias (C) and FPF counts. An analogous ANOVA was conducted for the Word Bias Groups (Figure 4B). This ANOVA did not show a main effect of Bias Group ($F(1, 77) = .041$, $MSE = 2.27$, $p = .841$, $\eta_p^2 = .00$), but did demonstrate main effects of Stimulus Class ($F(1, 77) = 94.23$, $MSE = .49$, $p < .001$, $\eta_p^2 = .55$) and Cycle ($F(2, 154) = 68.58$, $MSE = .25$, $p < .001$, $\eta_p^2 = .47$). Critically, these main effects were conditioned with three-way interaction between Stimulus Class, Bias Group and Cycle ($F(2, 154) = 5.49$, $MSE = .10$, $p = .005$, $\eta_p^2 = .07$). Figure 4B suggests the three-way interaction resulted because in the first Test accuracy for words was barely higher for the liberal versus conservative groups (at least numerically). However, the reverse was true for the picture groups where accuracy appeared higher for the conservative versus liberal groups. This pattern was not present in the second test, nor in the third test.

Figure 4

Mean Accuracy (d prime) for Picture and Word Bias Groups in Experiment 2



Note. The figure shows the mean accuracy rate for pictures and words. Upper part compares Conservative and Liberal Bias Groups in Picture target groups in Experiment 2. Lower part compares the Conservative and Liberal Bias Groups in Word target groups in Experiment 2. Numbers on the top shows each test Cycle (1,2, or 3). Error bars represent 95% confidence intervals.

For Cycle 1 there was no main effect of Bias Group ($F(1, 77) = .12$, $MSE = .57$, $p = .734$, $\eta_p^2 = .00$), but there was a main effect of Stimulus Class ($F(1, 77) = 56.33$, $MSE = .27$, $p < .001$, $\eta_p^2 = .42$) with higher accuracy for pictures. Confirming the impressions above, there was an interaction between Bias Group and Stimulus Class ($F(1,77) = 4.20$, $MSE = .27$, $p = .044$, $\eta_p^2 = .05$. However, the post-hoc pairwise comparisons of the liberal and conservative groups for pictures was not significant ($t(77) = 1.32$, $p = .191$) , nor was the comparison of the liberal and conservative groups for words ($t(77) = .98$, $p = .331$). Given this, we conclude that the interaction in Cycle 1 was likely spurious. In Cycle 2 and 3 there was no evidence for a Bias Group by Stimulus Class interaction ($F(1,77) = .275$, $MSE = .22$, $p = .601$, $\eta_p^2 = .0036$; $F(1,77) = .318$, $MSE = .21$, $p = .575$, $\eta_p^2 = .0041$).

Overall, these data suggest that there were no consistent accuracy differences across bias conditions in the data and hence the consideration of bias effects below is not confounded by sensitivity differences across bias conditions.

Response Bias (*C*- Bias)

As above, potential bias effects were considered by separately analyzing groups that received FPF to pictures (Figure 5A) and groups that received FPF to words (Figure 5B) Again, the goal was to determine if biases were learned for the FPF stimuli, and if so, whether these spread to the stimuli receiving neutral feedback.

The Figure 5A shows the bias data for the Picture Bias Groups for picture and word responding. It suggests similar biases for the liberal and conservative groups regardless of Stimulus Class or Cycle. These impressions were evaluated using a Bias Group (liberal vs conservative) by Stimulus Class (pictures vs words) and study/test Cycle (1, 2, or 3) mixed ANOVA. Bias Group was a between-subjects measure. This analysis revealed the main effects of

Stimulus Class ($F(1, 75) = 27.28$, $MSE = .13$, $p < .001$, $\eta_p^2 = .27$) and a main effect of Cycle ($F(1, 150) = 13.71$, $MSE = .07$, $p < .001$, $\eta_p^2 = .16$) but no main effect of Bias Group ($F(1, 75) = .001$, $MSE = .21$, $p = .978$, $\eta_p^2 = .00$). Additionally, there was an Stimulus Class by Cycle interaction ($F(1, 150) = 7.814$, $MSE = .04$, $p = .001$, $\eta_p^2 = .09$). Figure 5A suggest that this occurred because subjects became increasingly liberal for words (the non FPF stimuli) compared to pictures across the three tests. Pairwise comparisons of the bias for words versus pictures for each test confirmed this. During Cycle 1, this comparison approached significance ($t(77) = 1.98$, $p = .051$, $d = .23$). However, it became more robust in Cycle 2 ($t(77) = 4.54$, $p < .001$, $d = .52$) and Cycle 3 ($t(77) = 6.01$, $p < .001$, $d = .69$). Overall, because there was no main effect of Bias Group and this factor did not interaction with the other factors, the data indicate that FPF applied to picture stimuli was ineffective in instilling decision biases. Thus, we failed to replicate the bias effect observed in Picture Bias Groups in Experiment 1.

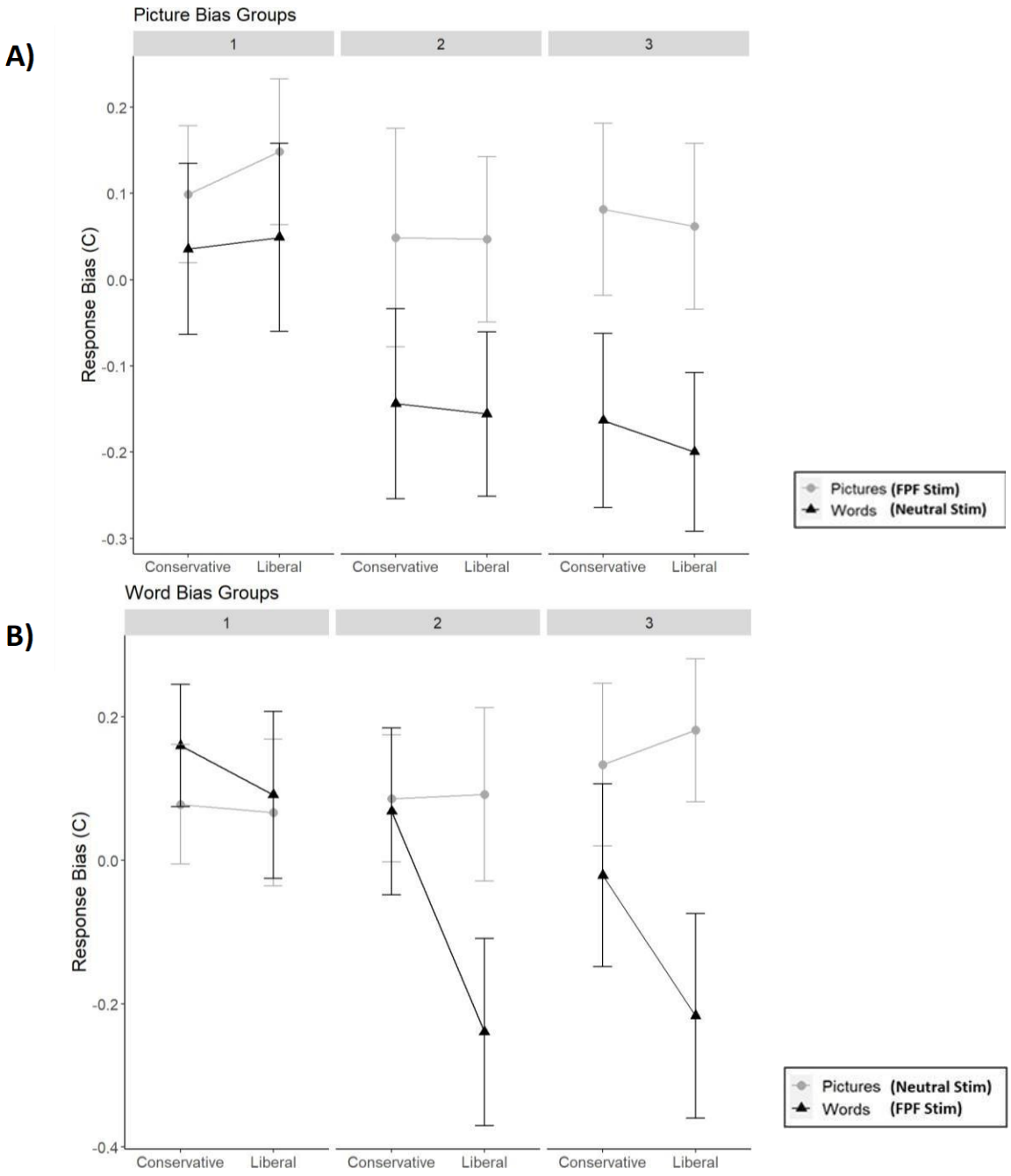
Turning to the Word Bias Groups shown in the Figure 5B, however, it appears that subjects developed a learned bias that was restricted to the class receiving FPF. This was confirmed via Bias Group (liberal vs conservative) by Stimulus Class (pictures vs words) by study/test Cycle (1, 2, or 3) mixed ANOVA. This analysis yielded main effects of Stimulus Class ($F(1, 77) = 11.83$, $MSE = .17$, $p = .001$, $\eta_p^2 = .13$) and Cycle ($F(2, 154) = 6.01$, $MSE = .07$, $p = .003$, $\eta_p^2 = .07$), but no main effect of Bias Group ($F(1, 77) = 2.66$, $MSE = .35$, $p = .107$, $\eta_p^2 = .03$). It also yielded two-way interactions of Bias group by Stimulus Class ($F(1, 77) = 7.19$, $MSE = .17$, $p = .009$, $\eta_p^2 = .09$) and Stimulus Class by Cycle ($F(2, 154) = 34.92$, $MSE = .03$, $p < .001$, $\eta_p^2 = .31$). Critically however, it also yielded a three-way interaction between Bias Group, Stimulus Class, and Cycle ($F(2,154) = 5.423$, $MSE = .03$, $p = .005$, $\eta_p^2 = .07$).

To decompose the three-way interaction, we conducted separate two-way ANOVAs with the factors of Bias Group by Stimulus Class ANOVAs for each cycle.

Beginning with the Cycle 1, this analysis yielded no main effect of Bias Group ($F(1,77)$ = .61, MSE = .10, p =. 437, η_p^2 = .01) or Stimulus Class ($F(1, 77)$ = 1.43, MSE = .08, p = .236, η_p^2 = .02), nor any interaction between Bias Group and Stimulus Class ($F(1,77)$ =.38, MSE = .08, p = .526, η_p^2 = .01). Hence for the Cycle 1, there were no bias differences across the Liberal and Conservative in neither FPF stimuli nor NF stimuli. In contrast, the same two-way ANOVA for the Cycle 2 yielded a main effect of Bias Group ($F(1,77)$ =5.54, MSE =.16, p =.021, η_p^2 =.07) and Stimulus Class ($F(1,77)$ = 12.81, MSE = .09, p =.001, η_p^2 =.14). Additionally, there was a Bias Group by Stimulus Class interaction ($F(1,77)$ = 10.55 , MSE = .09, p =.002, η_p^2 =.12). Follow-up pairwise comparisons showed that this interaction resulted because the response bias across Liberal and Conservative Bias Groups differed for the FPF stimuli (words), $t(77)$ = 3.57, d = .80, p <.001, but not for NF stimuli (pictures) (p = .938,). Turing to Cycle 3, there was no main effect of the Bias Group ($F(1,77)$ =.98, MSE =.22, p =.326, η_p^2 = .01), but the main effect of Stimulus Class ($F(1,77)$ =45.73, MSE =.07, p <.001, η_p^2 = .37) and the interaction between Bias Group and Stimulus Class ($F(1,77)$ = 8.99, MSE =.07, p =.004, η_p^2 = .11) were significant. Despite the lack of feedback at Cycle 3, pairwise comparisons showed a response bias difference between the Liberal and Conservative Bias Groups for the FPF stimuli (words), $t(77)$= 2.07, d = 47, p = .042. However, this did not occur for NF stimuli (pictures), (p=.524). These results suggest that, unlike Experiment 1, here the selective bias is demonstrated for word stimuli in Word Bias Groups, not picture stimuli.

Figure 5

Mean Response Bias (C) for Picture and Word Bias Groups in Experiment 2



Note. The figure shows the mean response bias rate for pictures and words. Upper part compares Conservative and Liberal Bias Groups in Picture target groups in Experiment 1. Lower part compares the Conservative and Liberal Bias Groups in Picture target groups in Experiment 2. Numbers on the top shows each test Cycle (1,2, or 3). Error bars represent 95% confidence intervals.

Awareness Questionnaire

Identical to Experiment 1, the awareness questionnaire was given at the end of Cycle 3 and designed to probe subjects' awareness of the influence of feedback on their performance. The open-ended responses regarding the influence of feedback are available online (OSF; osf.io/u45k7). The majority of subjects appeared to believe that the feedback was present to either increase the accuracy of responding or influence the confidence with which responses were given. Nonetheless, because a bias effect may be difficult to describe, the follow-up forced-choice questions provide a more direct test of awareness.

The analysis of awareness is restricted to the Liberal and Conservative Word Bias Groups since these were the only groups demonstrating learned FPF effects (Figure 5B). Table 7 shows the response counts for part I of the accuracy question, and the responses were clearly not randomly distributed across the three options ($\chi^2(2) = 62.68$, $p < .001$). Instead, a majority of subjects (75%) incorrectly concluded that the feedback made their recognition decisions more accurate, which is reliably higher than the chance rate of 33% ($\chi^2(2) = 60.79$, $p < .001$) and consistent with the open-ended responses.

Turning to the bias question, responses again were not distributed randomly (Table 7; $\chi^2(2) = 21.44$, $p < .001$). However, now a majority of subjects (51%) indicated they believed the feedback had no influence on their tendency to respond old, which is different from a chance rate of 33% ($\chi^2(2) = 7.75$, $p = .021$).

Table 7

Response counts for the Part I and Part II Questions for Picture Bias Groups in Experiment 2

Part I Questions (Purpose of the feedback?)		Part II Questions (Target class of the feedback?)		
Accuracy Question		**Words**	**Pictures**	**Both**
Increase accuracy	59	11	6	42
Decrease accuracy	5	1	1	3
No effect	15	NA	NA	NA
Bias Question		**Words**	**Pictures**	**Both**
Increase old responses	30	16	6	8
Decrease old responses	8	4	1	3
No effect	41	NA	NA	NA

Note. Part I Questions consist of a blend of accuracy and bias questions, which were presented in a randomized order. Part II Questions inquire about the specificity of the given feedback with respect to stimulus classes.

Of the 38 subjects indicating the feedback was biased, 24 correctly identified the direction of the putative bias. This rate does not exceed the chance expectation of 50% ($\chi^2(1) = 2.63$, $p = .105$). For these 24 subjects, 14 correctly identified words as the biased stimulus class, with 3 selecting pictures, and 7 indicating both stimulus classes were biased. This distribution differs from chance across the options ($\chi^2(2) = 12.25$, $p < .001$), and the 14 subjects correctly selecting the word option is above the chance expectation of 1/3rd ($\chi^2(1) = 6.75$, $p = .009$). Inspection of their open-ended responses in Table 8, however, does not suggest that they explicitly believed the feedback was selectively biased for word stimuli at the outset of the questionnaire.

Table 8

Open-ended Question responses of the potentially aware participants in Experiment 2

Subjects	Bias Group	Response to the open-ended question: "What was the purpose of the feedback?"
3007	Word Liberal	If it affects the choices I make in the future
3008	Word Liberal	Does categorizing words by their syllabus and images, if they are alive, make them easy to remember
3026	Word Liberal	I think that the purpose of the feedback was so we knew if we were on the right track with the correct memorization.
3031	Word Liberal	To guide you to change your answers
3035	Word Liberal	To see I my confidence level changed based on the feedback
3081	Word Conservative	Try to be more accurate/prompt u
3085	Word Liberal	To allow us to know if we were losing focus
3087	Word Liberal	Your confidence would change depending on if you received feedback
3091	Word Liberal	Sometimes, my mind engages in default mode processing, and getting an alert that I got something wrong got sometime thing wrong made me more aware.
3094	Word Liberal	To build confidence in guessing or try to improve memory if the answers were consistently wrong
3120	Word Liberal	I think it was supposed to motivate us and make us more alert.
3125	Word Liberal	The purpose of feedback was to show whether or not my initial instinct/judgment was correct.
3150	Word Liberal	To see how people's decisions change when they do/do not get confirmation of correct or incorrect responses.
3152	Word Liberal	To determine how receiving feedback on accuracy affects memory

Note. The table includes the responses of the potentially aware subjects who gave correct answer to Part I Questions bias question, and following Part II Question inquire about the specificity of the given feedback with respect to stimulus classes. Table only includes responses in Word Bias Groups due to only reliable bias effect observed in those groups.

As in Experiment 1, we removed potentially aware subjects to see if bias effects remained in the data. Hence the 14 potentially aware subjects were removed from the Word Bias Groups during the analysis of FPF effects on word responses (Figure 5B). When doing so the Bias Group by Cycle interaction only approached significance ($F(2,126) = 2.315$, $MSE = .05$, $p = .103$, $\eta_p^2 = .04$). However, post-hoc pairwise comparisons across tests (justified by the main analysis findings) demonstrated no reliable bias differences in Cycle 1, a reliable bias difference across the Liberal and Conservative Bias Groups in Cycle 2 ($t(63) = 2.31$, $d = .59$, $p = .024$), but no evidence of a remaining bias in Cycle 3 ($p = .401$). Overall, the awareness data from Experiment 2 are more equivocal than those from Experiment 1. Nonetheless, the significant bias finding in Cycle 2 for subjects not demonstrating any evidence of manipulation awareness, supports the conclusion that FPF biases can be acquired in the absence of manipulation awareness. Moreover, the less robust statistics likely reflect the reduction of power when removing 14 subjects between groups analyses. In the General Discussion, we further consider why detecting the FPF influence would be particularly difficult.

Discussion

The results of Experiment 2 again suggested that subjects could develop a class-specific response bias that is not transferred to recognition judgments of non-target classes. This bias was specific to the FPF target class. As opposed to Experiment 1; this time, bias only occurred for groups whose word stimuli were the FPF target class. This bias was learned through three study and test cycles, and even when feedback was removed, the bias persisted. More clearly, when FPF was given for missed word probes, subjects became more conservative in their judgments compared to when FPF was given for false alarm word probes. However, there was no effect on the criterion bias or non-target class, the pictures, which always received veridical feedback.

Therefore, differences in FPF-induced biases emerged between the Liberal and Conservative Word Bias groups for word probe trials but not for picture probe trials.

The Awareness Questionnaire data also showed very similar results to Experiment 1. Answers to an open-ended question suggest that the majority of subjects appeared to believe that the feedback was present to either increase their accuracy of response or influence the confidence with which they responded. Nonetheless, even with the follow-up forced-choice questions explicitly stating the possibilities about the purpose of the feedback, the subjects' answers were not consistent. These findings suggest that subjects were unaware of the role of the feedback in inducing a decision bias and that this bias was restricted to the picture stimuli.

General Discussion

The current study had two aims. The first was to examine the domain specificity of recognition decision biases learned via FPF. The second was to thoroughly examine the subjects' awareness of any biases acquired through FPF.

Using two experiments, we examined the domain specificity of recognition decision biases by administering FPF to only one class of intermixed test stimuli: in this case, words, or pictures. The goal was to examine whether FPF applied to one stimulus class yielded decision biases that remained restricted to the class or spread to the intermixed class of items that received fully correct feedback.

Prior work demonstrated that a 70% FPF manipulation yielded robust recognition biases, resulting in liberal responding when applied to false alarms and conservative responding when applied to misses (Han & Dobbins, 2009). The current study reduced this probability to 50% for each error, in hopes of making the manipulation even more subtle and difficult to explicitly detect by the subjects. In contrast to Han and Dobbins (2009) we also extensively probed the

subjects' awareness of the link between the feedback and their recognition decisions to see if any subjects appeared to understand how the feedback altered their recognition response tendencies. This was accomplished via an open-ended question and then a funnel procedure in which the specific nature of the manipulation could be selected in a forced-choice format.

Experiment 1 showed that that FPF manipulation produced response biases, but this was only reliable for the groups that received FPF targeting picture stimuli. Within these groups, the different biases of the liberal and conservative groups were only evident for the picture stimuli. They were not present in the intermixed word stimuli (which received veridical feedback), In contrast, when FPF was delivered to words in the two Word Bias Groups, there was no evidence for an induced bias. Thus, the findings suggested a selectively learned bias when pictures were targeted but we were not able to show the same phenomenon when words were targeted.

Consideration of the response questionnaire data suggested that the picture groups, which demonstrated a picture-selective decision bias, did not seem to be aware of how the feedback had altered their decision tendencies. More specifically, the responses to the open-ended question suggested that they almost unanimously thought the purpose of the feedback was to improve accuracy. Moreover, their performance on the targeted forced-choice questions demonstrated that the number of participants correctly identifying the nature of the feedback manipulation was no better than chance. Thus, for the groups where FPF targeted pictures, the results suggest a selectively learned decision bias of which the subjects are largely unaware.

This raises the questions of why the same FPF manipulation applied to words failed to produce any bias effects. We speculated that this failure might be due to the generally higher discrimination performance subjects demonstrated for words versus pictures in Experiment 1. More specifically, in Experiment 1, words *d'* was 2.29, while the picture *d'* was 1.98. Looking at

the respective error rates, which are the events that trigger FPF, misses were lower for words than pictures although false alarm rates were comparable ($Miss_{word}$: .12 vs., $FAs_{word:}$ .19, and $Miss_{picture}$: .21 vs $FAs_{picture}$: .16). Thus, it may have been the case that word groups simply did not have enough error trials to induce a decision bias. In an attempt to bring down word performance to the level of pictures in Experiment 2, we switched the encoding question for words from "pleasant?' to "two syllables?" without any change in the picture encoding task. This change indeed reduced the performance difference by producing lower *d'* for words (1.29), but unfortunately, it resulted in a reversed pattern of sensitivity with pictures now being significantly more discriminable (1.92). Thus, we were unable to equate performance across stimuli. Looking at the respective error rates, which are the events that trigger FPF, both misses and false alarms were rarer for pictures than words in Experiment 2 ($Miss_{word}$: .25 vs., FA_{word}: .30, and $Miss_{picture}$: .21 vs $FA_{picture}$: .16). This differential performance may have been responsible for the outcomes we discussed next, in which the pattern of selective biases that resulted from FPF were now reversed.

In Experiment 2, we observed selective FPF induced decision biases for the Word Bias Groups that did not spread to intermixed pictures recognition stimuli. Again, this learned bias persisted even when feedback was completely removed in the final test, supporting that it is due to incremental learning. However, the Awareness Questionnaire data were more equivocal. Although they generally suggested the subjects believed that feedback was designed to facilitate accuracy, more subjects chose the correct options on the forced-choice questions. Indeed, the majority of subjects selecting the correct forced-choice questions nonetheless indicated on the previous open-ended question that the feedback was designed to improve accuracy. Additionally, the questionnaire technique is conservative with respect to awareness because subjects might

only realize the nature of the feedback when answering the questions, which occur after testing. That is, even a subject who correctly indicated the nature of the induced bias in their group, may not have been aware of this during the testing procedure.

More critically, the FPF induced bias that was observed for the Word Bias groups in Experiment 2, was not observed for the Picture Bias groups. Thus, we were able to demonstrate selectively learned biases across both experiments but were unable to show this selectively for both classes of stimuli in either experiment. The consideration of error rates above does not strongly support the idea that this selectivity occurred because extremely low error rates in one versus the other class of stimuli because in Experiment 2 error rates were well above floor. Presently we can only speculate on the cause of this pattern, but it may stem from the fact that across both experiments, it was only the class of stimulus that was most difficult to discriminate that yielded the selectively learned bias when FPF was applied. This raises the possibility that FPF learning may largely depend not only upon committing errors, but that it may be potentiated for stimulus classes that the subjects subjectively believe are difficult to recognize. This may in turn cause FPF successes to be much more salient for the more difficult class. Related to this speculation, it may also be the case that subjective confidence in tasks using intermixed stimuli depends upon the perceived **relative** difficulty of judging the two classes. Under this interpretation, the efficacy of FPF to one stimulus class could be increased by simply improving the discrimination performance of the other class. In future research, as discussed below, we suggest that tailoring the FPF events specifically to low confidence reports may be more effective at instilling biases.

Domain Specificity of Reinforced Recognition Biases

As noted in the introduction, recognition decision biases can vary in terms of their origin. Certain biases may be implicitly acquired through environmental feedback, while others may be strategically employed in response to explicit instructions and reflect conscious goals to maximize rewards or correct responding across the test. Although we predicted that FPF induced biases would be difficult for observers to detect, which was the case, we also predicted that they would result in general biases that spread across the classes of stimuli within the test. The latter did not occur and this weighs against the idea that observers use an abstracted recognition signal during decision making that spans stimulus classes. That is, the finding of selectively suggests that what is learned during FPF is not to be generally cautious or liberal when interpreting recognition evidence in a particular test, but instead to be generally cautious or liberal for different types of recognition experiences; namely, those linked to the recognition of pictures versus words.

Generalized Signals of Familiarity or Likelihood Ratio Information

The history of the Theory of Signal Detection began with the application of the Neyman-Pearson statistical decision theory to human choice decision-making (Wixted, 2020). Under this approach, decisions are not based on raw mnemonic experiences but instead, reflect an abstracted evidence variable involving the comparison of the likelihood of two hypotheses, namely that the stimulus was studied or novel. The larger likelihood, typically represented by the value of the ratio of likelihood, dictates the choice of the observer. In its modern form, researchers have asserted that it is the use of such an abstracted likelihood ratio decision variable that leads to the regular patterns of recognition behavior across stimuli and various manipulations (Glanzer et al., 1993, 2009). This framework can be taken to predict that FPF would lead to a general bias spanning intermixed stimuli because it could reflect the learning of

how feedback relates to the utility of abstracted likelihood ratio information; information which is in a scale that is the same for all classes of stimuli. The benefit of such a decision variable is that it is universal for all memorial experiences. Of course, one could postulate that observers use likelihood ratio evidence, while also maintaining different evidence variables for every possible type of stimulus probe. However, this would undermine the utility of the likelihood transformation.

Stimulus Specific Learned Decision Biases

The stimulus specific biases demonstrated in the current manuscript suggest that decisions may instead be performed on an a less abstract evidence variable, that perhaps is available very early in the processing of the stimulus. For example, the complementary learning systems model of Norman and O'Reilly (Norman & O'Reilly, 2003) is potentially compatible with different familiarity signals for different stimulus classes, as is the notion that medial temporal lobe regions often associated with mnemonic processing may also directly contribute to perceptual experience (Barense et al., 2005). Under these approaches the perceived familiarity of different stimulus classes might be akin to different perceptual or quasi perceptual experiences, and it would be these experiences that serve as the foundation of operant learning. In short, observers would learn that different intensities of the different experiences are predictive of positive reinforcement and this learning would be selective. This learning would be similar to the type of operant learning of recognition decision biases proposed by Wixted and Gaitan (2002) for the pigeon, but it further assumes that the animals might be able to acquire biases restricted to specific stimulus classes. Of course, these ideas remain to be tested as the current study is the first to suggest that humans acquire class specific recognition decision biases under situations where they appear to be largely unaware of how these biases are acquired.

Awareness of Learned Biases

The current data provide the most compelling evidence to date that the biases acquired through FPF are opaque to the subjects. Although Han and Dobbins (2008) informally questioned subjects about the role of the feedback in shaping their recognition decisions, concluding that subjects did not seem to be aware of the manipulation, they did not conduct a funnel-style questionnaire as was done here. The current questionnaire data strongly indicate that subjects do not believe that the feedback has biased their recognition decisions generally, and consequentially further do not think they have acquired a bias selective to one or other class of stimuli. Indeed, the modal belief of the subjects appears to be that the purpose of the feedback was to improve accuracy and or alter their confidence in decisions.

The difficulty of detecting FPF likely arises because it is tied to errors, which are usually quite uncertain. When one considers that these FPF events are occurring in the midst of uncertain (but correct) responses for the targeted class, and uncertain responses (both correct and incorrect) for the intermixed class, it seems clear why FPF events are hard to detect. Indeed, in Experiment 1, not only did the open-ended responses suggest no awareness of the biasing purpose of the feedback, the responses to the more specific force-choice questions were consistent with chance responding. The awareness findings of Experiment 2 were more equivocal but also generally did not indicate high levels of manipulation awareness. Moreover, the awareness questionnaire procedure is, in general, conservative for two reasons. First, it considers subjects to be correct when they responded correctly to the forced-choice questions, even if their open-ended responses conveyed no indication that the feedback was designed to change response tendencies. Second, it is also possible that subjects might only become aware of the purpose of the feedback in hindsight, when the forced-choice questions are given. For example, when asked if the feedback

made them more or less likely to respond ‘old’ they might think back to the test and remember that they generally responded ‘old’ more often than ‘new’. Critically, however, this would not mean that during testing they realized that the FPF was steering them in this manner.

The Adaptive Significance of Shaped Recognition Biases

The current study presents evidence of a response bias that develops through reinforcement histories of two types of stimuli. Furthermore, these biases may be selective, meaning that when certain errors are encouraged for one type of stimulus (picture or word), subjects may adjust their decision criteria for that specific type of stimulus. This type of learning, if it operates broadly, would be highly adaptive because it would mean that the observers, without the use of fixed capacity, goal-based reasoning, would be able to acquire tuned recognition decision biases for a myriad of potential stimuli. Of course, there are operant learning phenomena that are tuned in this manner but the idea that individuals’ explicit decisions about their recognition experiences, may in part be determined as a function of the selective reinforcement histories of different classes of memoranda has not been proposed or considered to date. The current report suggests this may in fact occur, but caution is warranted because we were not able to demonstrate selective biases for both classes of stimuli within the same experiment. **Conclusion**

The findings of our study reveal that a mild form of FPF induces biased responses. Interestingly, this bias appears to be selective to the particular class of stimuli that received the FPF. This selective bias may reflect highly tuned learning mechanisms that are feature-based but the reason why we only observed selective FPF effects for one class of stimuli in Experiments 1 and 2 remains unclear.

Limitations and Future Directions

Recognition errors are often accompanied by low confidence (Yonelinas, 2001). Therefore, the current study aimed to manipulate feedback for only error trials to influence biases during low-confidence errors. However, the feedback manipulation was not tailored to low-confidence trials specifically and was instead randomly applied to each class of errors 50% of the time. Since recent recognition research suggests that a surprising number of errors may be accompanied by high confidence (Roediger & Tekin, 2022) it may also be the case that a sizeable portion of our FPF events were triggered by either high or medium confidence errors. This is not ideal under learning accounts that assume that the amount of learning is tied to the degree prediction error (Schultz & Dickinson, 2000) because subjects presumably expect to be told they are correct when confidence increases, rending the FPF events on these trials ineffective in altering behavior. To achieve more consistent results, future research should use the same feedback manipulation but restrict it to errors with low confidence. This should produce a stronger effect if prediction error guides performance.

References

Aminoff, E. M., Clewett, D., Freeman, S., Frithsen, A., Tipper, C., Johnson, A., Grafton, S. T., & Miller, M. B. (2012). Individual differences in shifting decision criterion: A recognition memory study. *Memory & Cognition*, *40*(7), 1016–1030. https://doi.org/10.3758/s13421-012-0204-6

Azimian-Faridani, N., & Wilding, E. L. (2006). The Influence of Criterion Shifts on Electrophysiological Correlates of Recognition Memory. *Journal of Cognitive Neuroscience*, *18*(7), 1075–1086. https://doi.org/10.1162/jocn.2006.18.7.1075

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, *74*(2), 81–99. https://doi.org/10.1037/h0029531

Barense, M. D., Bussey, T. J., Lee, A. C. H., Rogers, T. T., Davies, R. R., Saksida, L. M., Murray, E. A., & Graham, K. S. (2005). Functional Specialization in the Human Medial Temporal Lobe. *The Journal of Neuroscience*, *25*(44), 10239–10246. https://doi.org/10.1523/JNEUROSCI.2704-05.2005

Bowen, H. J., Marchesi, M. L., & Kensinger, E. A. (2020). Reward motivation influences response bias on a recognition memory task. *Cognition*, *203*, 104337. https://doi.org/10.1016/j.cognition.2020.104337

Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of Standardized Stimuli (BOSS) Phase II: 930 New Normative Photos. *PLoS ONE*, *9*(9), e106953. https://doi.org/10.1371/journal.pone.0106953

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review*, *3*(1), 37–60. https://doi.org/10.3758/BF03210740

Estes, W. K., & Maddox, W. T. (1995). Interactions of stimulus attributes, base rates, and feedback in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1075–1095. https://doi.org/10.1037/0278-7393.21.5.1075

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

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review*, *100*(3), 546–567. https://doi.org/10.1037/0033-295X.100.3.546

Glanzer, M., Hilford, A., & Maloney, L. T. (2009). Likelihood ratio decisions in memory: Three implied regularities. *Psychonomic Bulletin & Review*, *16*(3), 431–455. https://doi.org/10.3758/PBR.16.3.431

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*(4), 703–715. https://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*(3), 469–474. https://doi.org/10.3758/PBR.16.3.469

Macmillan, N. A., & Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes. *Psychological Bulletin, 107*(3), 401–413. https://doi.org/10.1037/0033-2909.107.3.401

Macmillan, N. A., & Creelman, C. D. (2004). *Detection Theory* (0 ed.). Psychology Press. https://doi.org/10.4324/9781410611147

Mintzer, M. Z., & Snodgrass, J. G. (1999). The picture superiority effect: Support for the distinctiveness model. *The American Journal of Psychology, 112*(1), 113–146.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review, 110*(4), 611–646. https://doi.org/10.1037/0033-295X.110.4.611

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

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*(2), 305–320. https://doi.org/10.1037/0278-7393.33.2.305

Roediger, H. L., & Tekin, E. (2022). Can signal detection theory explain everyday amnesia (high confident misses)? *Neuropsychologia, 166*, 108115. https://doi.org/10.1016/j.neuropsychologia.2021.108115

Schultz, W., & Dickinson, A. (2000). Neuronal Coding of Prediction Errors. *Annual Review of Neuroscience, 23*(1), 473–500. https://doi.org/10.1146/annurev.neuro.23.1.473

Turner, B. M., Van Zandt, T., & Brown, S. (2011). A dynamic stimulus-driven model of signal detection. *Psychological Review*, *118*(4), 583–613. https://doi.org/10.1037/a0025191

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 582–600. https://doi.org/10.1037/0278-7393.26.3.582

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition*, *35*(2), 254–262. https://doi.org/10.3758/BF03193446

White, K. G., & Wixted, J. T. (1999). PSYCHOPHYSICS OF REMEMBERING. *Journal of the Experimental Analysis of Behavior*, *71*(1), 91–113. https://doi.org/10.1901/jeab.1999.71-91

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, *20*(1), 6–10. https://doi.org/10.3758/BF03202594

Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(2), 201–233. https://doi.org/10.1037/xlm0000732

Wixted, J. T., & Gaitan, S. C. (2002). Cognitive theories as reinforcement history surrogates: The case of likelihood ratio models of human recognition memory. *Animal Learning & Behavior*, *30*(4), 289–305. https://doi.org/10.3758/BF03195955

Yonelinas, A. P. (2001). Consciousness, control, and confidence: The 3 Cs of recognition memory. *Journal of Experimental Psychology: General*, *130*(3), 361–379. https://doi.org/10.1037/0096-3445.130.3.361

Appendix A

Subjective Awareness Questionnaire

Open ended question

What was the purpose of the feedback?

Funnel Questionnaire

Do you think the feedback made you:
a) more accurate
b) less accurate
c) had no effect on accuracy

Do you think the feedback affected your accuracy for:
a) words
b) pictures
c) both words and pictures

Do you think the feedback made you:
a) more likely to respond old
b) less likely to respond old
c) had no effect on tendency to respond old

Do you think the feedback affected your tendency to respond old for:
a) words
b) pictures
c) both words and pictures