Summer 9-1-2014

Examining an Implicit Mechanism of Recognition Criterion Regulation

Justin Christopher Cox
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Psychology

Dissertation Examination Committee:
Ian G. Dobbins, Chair
Dave Balota
Todd Braver
Carl Craver
Henry L. Roediger III

Examining an Implicit Mechanism of Recognition Criterion Regulation

by

Justin Christopher Cox

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2014

St. Louis, Missouri

Table of Contents

List of Figures....iv

List of Tables....vi

Acknowledgements....vii

Introduction....1

Chapter 1: Memory As Decision-Making....1

The Signal Detection Model of Recognition Decisions....1

A Review of Criterion Setting....4

Explicit Influences....4

Base Rates....5

Feedback....8

Lingering Questions....11

Chapter 2: Implicit Learning: Learning Without Intent....15

Classification Learning....16

Motor Sequence Learning....21

General Trends in the Acquisition of Implicit Knowledge....23

Chapter 3: Implicitly Learning to Evaluate Explicit Memories....25

Chapter 4: Specific Aims and General Methods....27

Chapter 5: Experiment 1....32

Procedure....35

Results....37

Discussion....55

Chapter 6: Experiment 2....60

Procedure....62
Results....64
Discussion....72
Chapter 7: Experiment 3....76
Procedure....79
Results....81
Discussion....92
Chapter 8: Evaluating a Temporal Difference Account of Feedback Influence....96
Procedure....97
Results....99
Discussion....102
Chapter 9: General Discussion and Conclusions....104
The Basal Ganglia and Memory Decisions....111
Conclusions....114
References....115
Appendix A....126

List of Figures

Figure 1: Basic Signal Detection Model. Studied and novel items evoke overlapping degrees of evidence. The distance between the two distributions corresponds to the discriminability (d') of the two item types (further spaced distributions are less confusable). The vertical line between the two distributions represents the decision criterion.

Figure 2: As average memory strength increases, a statistically optimal observer should become more conservative in order to maintain a likelihood ratio of 1.

Figure 3: Rule-based versus information-integration perceptual categorization tasks. In the rule-based task (left), stimuli are categorized based on one stimulus dimension (orientation). In the information-integration task, stimuli are categorized based on a linear combination of line length and line orientation.

Figure 4: Hypothetical experimental results. The top panel (predicted results) shows hypothetical results if participants are unable to ignore the influence of FPF; criterion should not differ between Use and Ignore groups. The bottom panel shows hypothetical results if participants ARE able to ignore the influence of FPF; in this case, criterion should differ between Use and Ignore groups that receive the same feedback.

Figure 5: Experimental Design (Experiment 1). Participants were divided into four groups. All groups received veridical feedback during Test 1. Groups received either Liberal or Conservative FPF and were told to "Use" or "Ignore" this feedback during Tests 2 and 3. A surprise subsequent memory test followed Test 3, testing over particular items drawn from Test 2.

Figure 6: Effects of Feedback and Instructions on criterion in Tests 2 and 3 (Experiment 1). Instructions to ignore feedback do not appear to eliminate its effects (vertical bars represent 95% confidence intervals).

Figure 7: Hypothetical results. The top panel (predicted results) shows hypothetical results if the FPF effect transfers to stimuli that do not receive reinforcement (faces); criterion should not differ for words and faces. The bottom panel shows hypothetical results if the FPF effect does not transfer to stimuli that never receive reinforcement; criterion for faces should not change appreciably from baseline.

Figure 8: Examples of face stimuli for Experiment 2

Figure 9: Differences in c become difficult to interpret when d' differs. What appears to be the same criterion on the memory evidence axis would result in a different value of c for the two stimulus types.

Figure 10: Criterion diverges across tests based on feedback delivered (Experiment 2). Importantly, criterion appears to generalize between words and faces (vertical bars represent 95% confidence intervals).

Figure 11: Hypothetical results. The top panel (predicted results) shows hypothetical results if the FPF effect is context-sensitive (i.e., does not transfer to a new context); criterion should return to baseline during the final test for the Shift Context groups. The bottom panel shows hypothetical results if the FPF effect is NOT context-sensitive; the FPF effect should remain in the new testing context for the Shift Context groups.

Figure 12: Effects of context change on criterion (Experiment 3). Dotted line indicates when all groups left the original testing room. No Shift participants returned to the original testing room to complete Test 4, while Shift participants remained in the new context and completed Test 4. Changing context did not appear to eliminate the effects of FPF (vertical bars represent 95% confidence intervals).

Figure 13: Logistic regression betas for Liberal participants. FA+ indicates false alarms that received FPF, while FA- indicates false alarms that received negative feedback.

Figure 14: Logistic regression betas for Conservative participants. M+ indicates misses that received FPF, while M- indicates misses that received negative feedback.

Supplemental Figure 1: Three-way interaction between Test, Feedback Group, and Volition on Believed Accuracy (vertical bars denote 95% confidence intervals).

List of Tables

Table 1: Hit rates, false alarm rates, accuracy, criterion, and number of FPF trials received for the four groups across all tests (Experiment 1). Standard deviations in parentheses.

Table 2: Conditional hit rates for Subsequent Memory items. Standard deviations in parentheses.

Table 3: Subjective Awareness Questionnaire data, Experiment 1.

Table 4: Hit rates, false alarm rates, accuracy, criterion, and number of FPF trials for both words and faces during Tests 1, 2, 3, and 4 (Experiment 2). Standard deviations in parentheses.

Table 5: Subjective Awareness Questionnaire data, Experiment 2.

Table 6: Hit rates, false alarm rates, accuracy, criterion, and number of FPF trials for Tests 1, 2, 3, and 4 (Experiment 3). Standard deviations in parentheses.

Table 7: Subjective Awareness Questionnaire data, Experiment 3.

Supplemental Table 1: Individual difference measures. Standard deviations in parentheses

Acknowledgments

Firstly, I would like to thank my faculty adviser, Professor Ian G. Dobbins for his mentorship and guidance during my graduate career. I owe my success in graduate school to his patience and supervision. Next, I would like to acknowledge my dissertation committee for their helpful comments and assistance during the preparation of this dissertation. Finally, I would like to thank my family and my partner Joshua Carlson for their support during my graduate career.

Introduction

Episodic recognition is the judgment that a stimulus has been encountered previously. Although researchers have traditionally focused on elucidating the number and nature of memory signals that directly guide recognition, numerous other aspects of the environment may also provide information that should be relevant for recognition judgment; particularly under situations in which the memory signals are weak or ambiguous. For example, determining whether a face in a crowd is someone you know well enough to approach and speak to should be heavily influenced by whether one is in a highly familiar or unfamiliar environment such as a local grocery store, versus a market in a foreign country; or other factors such as the potential social embarrassment from making the wrong decision, and whether one has recently embarrassed oneself by recently and accidentally warmly greeting a stranger. Under this conceptualization, people do not make decisions about their memories in a vacuum and multiple factors should play a role in biasing memory decisions. This dissertation looks at the role of reinforcement histories in shaping how recognition judgments are biased towards conclusions of novelty or familiarity.

Chapter 1

Memory As Decision-Making

The Signal Detection Model of Recognition Decisions

A vast majority of recognition memory models – both single and dual-process models – assume that some aspect of memory decisions is based on a continuous dimension of recognition

strength or intensity, and this dimension is most often modeled via Signal Detection Theory (Banks, 1970; Wixted & Mickes, 2010; Yonelinas, 2002). Under this approach, different classes of stimuli (e.g., novel and familiar faces) are assumed to evoke different distributions of memory evidence – with novel stimuli forming a distribution lower on the evidence axis than familiar stimuli (Figure 1). Critically, because the distributions are continuous and overlapping, and observers are forced to make dichotomous judgments, an observer's task is to divide this evidence continuum into values diagnostic of novelty or familiarity using a cutoff known as a decision criterion, or decision bias. Items that evoke evidence values less than the criterion are judged as novel, whereas items that evoke evidence values greater than the criterion are judged as familiar.

If the observer is assumed to be statistically ideal, then the evidence axis is not one of raw memory strength signals, but a translation of strength into statistical likelihoods of the two possible judgments (Glanzer & Adams, 1990; Glanzer, Hilford, & Maloney, 2009; Pastore, Crawley, Berens, & Skelly, 2003; Turner, Van Zandt, & Brown, 2011). Under this conception, decisions are based on the likelihood a given strength would have been encountered if sampled from the pool of old items versus the likelihood that the same strength value would have been encountered if the item were instead sampled from the pool of new items. These likelihoods are indicated by the relative heights of the normal distributions at that point on the strength axis. The statistically ideal observer then selects the decision with the greater likelihood. This decision process is most efficiently represented by combining the two likelihoods in a ratio [p(old)/p(new)], with values greater than one favoring an old conclusion, and those less than one favoring the new conclusion; this decision rule is a statistically optimal approach to distinguishing between two alternative hypotheses (Neyman & Pearson, 1992).

Such a description denotes a decision-making system that seems highly optimal and largely explicit. Critically, the likelihood decision rule, if feasible, enables observers to make statistically optimal decisions under a host of different scenarios (Criss & McClelland, 2006; Turner, Van Zandt, & Brown, 2011). For example, if an observer knows a recognition test contains an equal number of old and new items, he can set the criterion to the ideal likelihood ratio of 1. Alternatively, if the number of old items outnumbers the number of new items by 3:1, the ideal observer establishes the criterion at a likelihood ratio of 3 (Macmillan & Creelman, 2004). Under either case, a statistically ideal observer will maximize the number of correct responses by modifying his responding to accommodate the relative preponderance of old and new items. This brief review examines some of the influences on criterion placement in recognition.

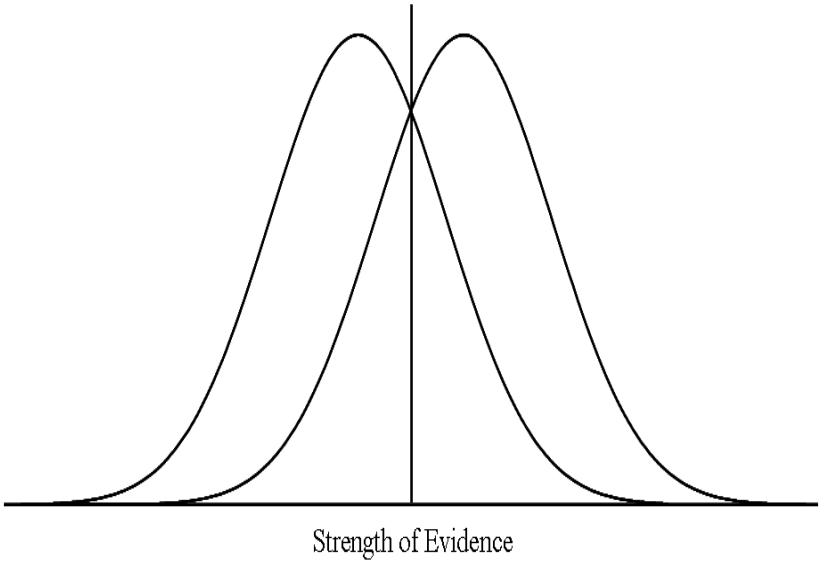


Figure 1: Basic Signal Detection Model. Studied and novel items evoke overlapping degrees of evidence. The distance between the two distributions corresponds to the discriminability (d') of the two item types (further spaced distributions are less confusable). The vertical line between the two distributions represents the decision criterion.

A Review of Criterion Setting

Under the likelihood ratio version of Signal Detection theory, observers can judiciously regulate their criteria in order to maximize some desired outcome. For instance, when informed of the target to lure ratio (i.e., the base rates of targets and lures) they should update their criteria accordingly to maximize the proportion of correct responses. Such a conceptualization denotes a decision-making mechanism that is arguably largely strategic and controlled. For instance, several researchers have described criterion shifting as strategically controlled or requiring "cognitive effort" (Benjamin & Bawa, 2004; Dobbins & Kroll, 2005; Rhodes & Jacoby, 2007; Stretch & Wixted, 1998). Many of the manipulations discussed below generally support the notion of controlled, explicit criterion-setting. However, as will be evidenced by the review below, the precision of these shifts is often vastly improved by performance feedback, suggesting that active feedback-based learning can fine-tune an initially coarse strategically placed criterion.

Explicit Influences

Manipulations that fall under the banner of “explicit influences” are those that use some sort of explicit instruction or environmental cue to encourage participants to adopt a more liberal or conservative criterion than he or she would naturally employ. Critically, in these situations, observers are likely able to state how they should proceed in order to improve overall accuracy; for example, by being more stringent with respect to old materials since they form a minority of items in the test list. Manipulations of this sort are typically quite effective in influencing criterion placement. A key factor that unites many of these manipulations is the assumption that

observers can exert a measure of control and precision over the placement of the criterion in order to maximize some desired outcome.

Task instructions are one such example of explicitly influencing the criterion (Healy & Jones, 1975; Hirshman, 1995; Miller, Handy, Cutler, Inati, & Wolford 2001) with researchers simply verbally encouraging observers to respond in a certain manner (e.g., "be very cautious when calling an item old"), and participants adjusting their criteria accordingly. Other researchers have used trial-by-trial hints (or "cues") in order to dynamically affect criterion placement (Jaeger, Cox, & Dobbins, 2012; Jaeger, Lauris, Selmeczy, & Dobbins, 2011; O'Connor, Han, & Dobbins, 2010; Selmeczy & Dobbins, 2013). In a typical cueing experiment, participants are given hints on a subset of trials during a recognition memory test. Participants are typically informed of the validity of the cues (e.g., "cues will be correct 75% of the time"), and are instructed to use the cues to improve their performance. As would be predicted, participants are able to incorporate these cues to adjust their responding (e.g., participants are more likely to respond "old" when given a cue that an item is "likely old"); that is, participants shifted their criteria in accordance with an explicit instruction that had a known validity.

Base Rates

One variable that has traditionally been manipulated when trying to influence observers' decision criteria is the base rates of targets to lures. For example, participants in an experiment can be told that an upcoming test is composed of 25% targets and 75% lures, and should be able to adjust the decision criterion and their relative response frequencies to approximate the item frequencies. Indeed, base rate manipulations have been fairly successful in that observers do seem to incorporate base rate information when adjusting the criterion (Estes & Maddox, 1995;

Healy & Jones, 1975; Healy & Kubovy, 1978; Ratcliff, Sheu, & Gronlund, 1992; Van Zandt, 2000; Van Zandt & Maldonado-Molina, 2004).

Van Zandt (2000) informed participants of the relative proportion of studied items and found that participants adjusted their criteria in accordance with the informed base rates. Similarly, Ratcliff, Sheu, and Gronlund (1992) informed their participants of the relative proportions of studied and novel items on an upcoming test, and found that decision criteria tended to track this information. Finally, Kantner and Lindsay (2010) manipulated base rates prior to a recognition test, but did not inform participants about the test make-up. However, participants were able to learn this information and respond accordingly when trial-wise performance feedback was provided at test.

Although the studies reviewed above seem to indicate a general effectiveness of base rate manipulations on the criterion, examples of extreme base rate manipulations in the recognition memory literature have painted a markedly different picture of how well observers can incorporate base rate information (Cox & Dobbins, 2011; Ley & Long, 1987, 1988; McKelvie, 1993; Strong & Strong, 1916; Underwood, 1972; Wallace, 1978, 1982; Wallace et. al., 1978). In these instances, one class of items is wholly removed from the test environment, and participants are either correctly informed or uninformed about the exact construction of the test list. Wallace demonstrated remarkable similarities in the endorsement rates of studied items regardless of whether studied items were presented with or without novel items intermixed during testing. This similarity was present regardless of whether participants were aware of the manipulation (Wallace et. al., 1978) or not (Wallace, 1982). Likewise, Cox and Dobbins (2011) found that participants failed to shift the criterion optimally during these so-called “pure list” tests (tests composed entirely of studied or entirely of novel items). In fact, they found that participants

were more liberal (more likely to respond "old") during a test composed entirely of novel items (a target-free test), demonstrating an unwillingness to use the criterion to compensate for wildly different base rates of items. Furthermore, Cox and Dobbins (2011) demonstrated that not only were overall target hit rates similar for pure and mixed lists, but that the distribution of confidence was also quite similar when participants were uninformed of the prior probabilities, further suggesting limited awareness of anything aberrant about the test.

As opposed to the base-rate literature reviewed previously, the pure-list literature suggests that observers are unable – or at least unwilling – to use base rate information as an impetus for opportunistically shifting the decision criterion, at least when making recognition memory decisions. Cox and Dobbins (2011) highlighted two potential key differences between traditional base rate and pure-list manipulations, and argued that opportunistic shifts require observers to both a) detect a large scale difference in base rates (if not explicitly told the proportions of each item type) and b) be encouraged to capitalize on this information. For example, Healy and Kubovy (1978) reported large criterion shifts as a function of changing base rates. However, in their experiments, participants were informed of the base rates prior to initiating a test block, and were given trial-to-trial performance feedback. Likewise, in her first reported experiment, Van Zandt (2000) informed participants what proportion of an upcoming test list would be composed of studied items and provided points and feedback to encourage an appropriate criterion shift. Finally, Estes and Maddox (1995) manipulated base rates and the presence of performance feedback, and found that participants only appreciably shifted their response criteria when feedback was provided at test. Cox and Dobbins (2011) argued that observers might "not spontaneously detect such manipulations, or if they do, they require some

further encouragement or impetus to opportunistically shift the decision criterion," with this impetus typically tied to the provision of feedback and performance summaries.

Supporting this, Heit, Brockdorff, and Lamberts (2003) examined how observers could dynamically adjust their criteria using a paradigm dubbed the "response signal technique." The response signal technique requires participants to respond to a recognition stimulus at some varying time interval after the stimulus appears (i.e., respond after hearing a tone, and the onset of the tone varies pseudorandomly). The authors varied the base rates associated with the timing of the signal (e.g., an early signal was associated with a 9:1 ratio in favor of the item being new). In this case, the timing of the tone served as a cue for the item's status, but this fact was never stated to participants. Critically, summary feedback was provided at the end of each test. Response criterion appeared to track changes in the base rates across signal timing – in essence, participants had learned an association between response timing and the base rates, even without trial-by-trial feedback or specific instructions. The authors suggest that participants are able to strategically adjust the criterion to reflect the changes in base rates associated with response timing, although the authors did not specifically query participants about whether they had any explicit awareness of the relation between signal timing and their own response patterns. Critically, the summary feedback may have rendered the relationship between signal timings and base-rates more explicit to the participants such that they then used the former to strategically bias responding.

Feedback

As discussed above, feedback often plays a role in studies of criterion placement. Consider the power of embarrassing social feedback from falsely recognizing someone; such a

situation provides salient information about the caution an observer should exercise when evaluating his own recognition memory. Further, this type of feedback can serve as a sort of "alerting tool" by providing information about the memory environment an observer encounters; for instance, the likelihood of an observer encountering faces he will recognize (i.e., base rate information). As alluded to earlier, feedback is typically most useful when some aspect of the task demands is initially opaque to the observer (e.g., different base rates, Kantner & Lindsay, 2010). Indeed, it has been suggested that “feedback is a tool for optimizing or controlling bias” in recognition memory (Rotello & Macmillan, 2008). Thus feedback may be important in incrementally adjusting and fine-tuning the placement of the criterion.

There are a fair number of studies demonstrating how feedback can influence decision criterion placement (Carterette, Friedman, & Wyman, 1966; Estes & Maddox, 1995; da Silva & Sunderland, 2010; Han & Dobbins, 2008, 2009; Kantner & Lindsay, 2010; Rhodes & Jacoby, 2007). Supporting this idea, da Silva and Sunderland (2010) demonstrated that trial-by-trial feedback reduced the variability of the recognition criterion in older adults (i.e., older adults demonstrated more stable criteria). Further, Kantner and Lindsay (2010) investigated whether feedback could improve recognition memory performance. They found that while feedback did not improve overall discriminability, it encouraged participants to adopt more appropriate decision criteria when unequal base rates were used (Experiment 2); critically, participants were not aware of the differences in base rates prior to the test, but were able to pick up on this information using the performance feedback. Likewise, Rhodes and Jacoby (2007) correlated the probability that a test item had been previously studied with its location on the screen during test presentation (e.g., items presented on the left half of the screen had a 66% chance of being previously studied) and provided performance feedback (Experiment 3). They found that

feedback moderated the effects of awareness of the manipulation on the criterion: even when participants claimed to be aware of the different base rates, they were not able to capitalize on this information to the same degree without the presence of reinforcing feedback.

Verde and Rotello (2007) attempted to encourage dynamic, within-test criterion shifts using seamless changes in the average memory strength of items within a single test context. When average memory strength changes (e.g., mean of the target distribution moves closer to the lure distribution), a statistically ideal observer should shift the criterion such that the likelihood ratio is still 1 (Figure 2). Verde & Rotello (2007) manipulated the average strength of targets seamlessly across two halves of a test list such that, for example, lures were intermixed with strongly encoded targets during the first half whereas lures were intermixed with weakly encoded targets in the second half of the subjectively same test list. That is, there was no break or environmental cue given to observers when transitioning from the portion with strong targets to the portion with weak targets. If observers could optimally utilize memory strength itself as a cue to update the decision criterion, they should adopt different criteria for each half of the test when average difficulty of discrimination changes across test halves. Although Verde and Rotello (2007) found that initial criterion placement differed depending upon whether subjects started with a strong or weak test half, they did not alter the adopted criterion when transitioning to the second half of the test. It was not until summary feedback was included partway through each test (Experiment 5) that participants were able to appreciate the changes in average memory strength. As Verde and Rotello (2007) point out, observers likely do not directly utilize the average strength of encountered materials to modulate and update recognition criteria, and may instead rely upon cues not directly tied to memory representations (such as information provided by error feedback; or expectations, McCabe & Balota, 2007).

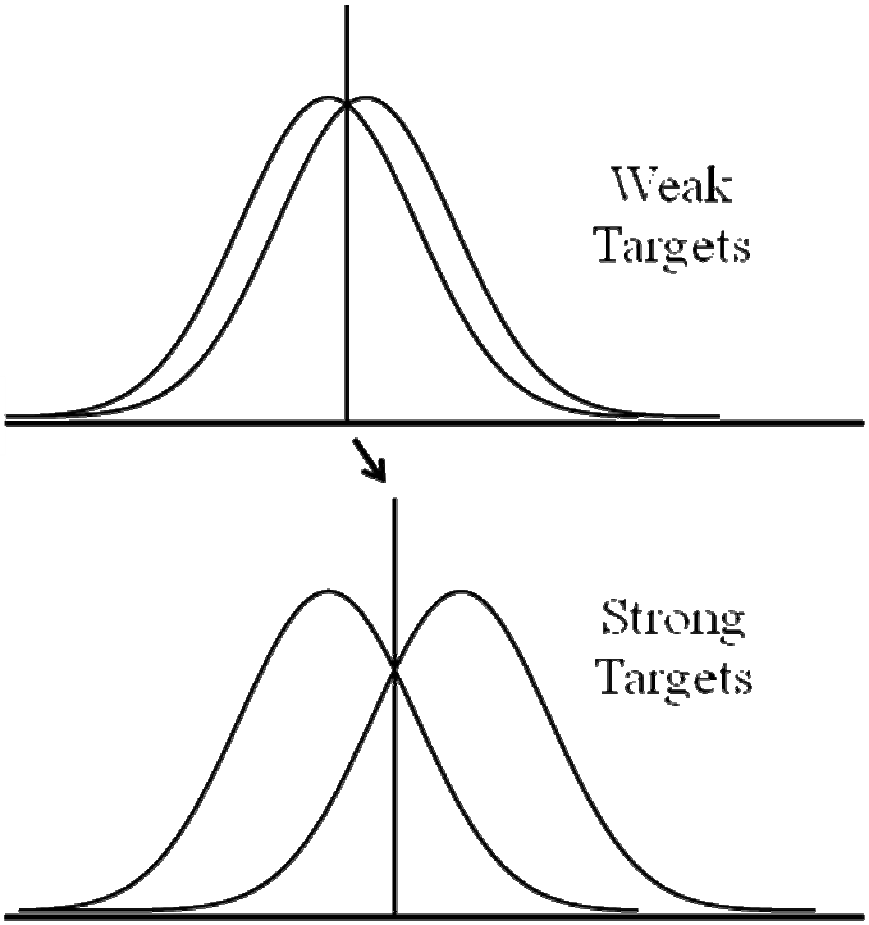


Figure 2: As average memory strength increases, a statistically ideal observer should become more conservative in order to maintain a likelihood ratio of 1.

Lingering Questions

Overall, the evidence reviewed above paints a general picture of a recognition decision criterion that lies under the volitional control of the observer (although the optimality of this control is a separate question). Observers appear to be able to generally adjust the criterion to account for instructions, payoffs, and base rates, all of which suggest a reasonable degree of volitional control over criterion placement. However, in addition to all of the evidence pointing toward an explicitly set and controlled criterion, several other studies suggest the criterion can be

influenced incrementally via trial-by-trial feedback (Han & Dobbins, 2008, 2009; Kantner & Lindsay, 2010; Rhodes & Jacoby, 2007). To elaborate, these studies together demonstrate dynamic criterion shifts that are, additionally, not always transparent to the observer; that is, feedback may sometimes alter the decision tendencies even though the observer has not used the feedback to adopt an explicit response strategy.

Further evidence suggesting a role for feedback-based implicit learning governing criterion placement comes from studies that used probabilistically biased feedback procedures to differentially reinforce one response type over another. In one set of experiments, Han and Dobbins (2009) utilized a biased feedback schedule to influence the decision criterion. In Experiment 1, participants probabilistically received positive feedback about one class of error (i.e., being told an incorrect "old" response was correct; henceforth referred to as *false positive feedback [FPF]*). Across the experiment, large criterion differences were observed depending on which response type was differentially reinforced – for example, participants tended to respond "old" more often when FPF was given for incorrect "old" reports. When the tests were broken down into sub-blocks, it was revealed that criterion differences became larger across sub-blocks. Finally, criterion differences persisted across two later tests, even when feedback was removed. Interestingly, participants who demonstrated the effects seemed unaware of the anomalous nature of the feedback (Han & Dobbins, 2008). Part of what makes this design interesting is that the feedback manipulation only appears on error trials – trials in which observers likely have very little diagnostic information and are apt to be subjectively guessing. Because of this, the anomalous nature of the feedback often remains apparently unknown to observers, so any changes in decision-making is unlikely to reflect top-down strategies in response to performance feedback (e.g., Rhodes & Jacoby, 2007).

The FPF paradigm suggests that feedback can be used as more than just an alerting tool in recognition decision-making. In this case, feedback appears to drive a slow, incremental sort of reinforcement-based learning. Others have proposed that this sort of learning is responsible for the surface resemblance that humans and other animals bear to statistically optimal likelihood ratio decision-makers (Wixted & Gaitan, 2002). While it is reasonable to assume that humans can place and control a strategic and explicit criterion (somewhat akin to a likelihood ratio decision axis), it seems odd to assume the same approach on the part of animals such as pigeons. Nonetheless, these animals will demonstrate criterion shifts when given appropriate reinforcement, providing further evidence for the role of feedback in recognition criterion setting. Taken together, this suggests evidence for both explicitly controlled criterion setting and feedback-based criterion learning operating during recognition, at least for humans (c.f. Poldrack & Packard, 2003) with these two mechanisms potentially operating in parallel. This parallel operation is clear in studies where feedback amplifies some explicit response strategy (Rhodes & Jacoby, 2007; Verde & Rotello, 2007). While explicitly controlled criterion setting has been well-established by the extant literature, the notion of a feedback-based criterion learning mechanism remains relatively uncharacterized. For feedback-based criterion learning, it is evident that observers do not necessarily acquire a global strategy, it is (perhaps obviously) heavily feedback-dependent, and it can persist in the absence of continued reinforcement (Han & Dobbins, 2008, 2009; Kantner & Lindsay, 2010). These descriptions are consistent with other forms of feedback-based implicit learning such as habit formation or procedural learning (Dayan & Daw, 2008; Yin & Knowlton, 2006).

The current dissertation proposes a framework of recognition criterion setting whereby feedback can influence the recognition criterion in one of two ways. The first is through alerting

the observer to enact some useful response strategy (e.g., capitalizing on differing base rates in a test; Kantner & Lindsay, 2010, Rhodes & Jacoby, 2007). The second is through incremental adjustment, akin to sorts of reinforcement learning commonly seen in pigeons and other animals (Wixted & Gaitan, 2002). This work focuses on the latter avenue of feedback-based criterion learning. The next chapter will present a brief introduction to implicit learning, focusing on how it may relate to the current proposed framework.

Chapter 2

Implicit Learning: Learning Without Intent

Implicit learning constitutes knowledge or associations acquired from the environment whose functional properties are held largely outside of consciousness (Reber, 1989). In other words, in contrast to most explicit learning, implicit learning is typically acquired without conscious attempt to do so, and the associations acquired are typically unavailable to conscious access. Observers are unlikely to be able to state what features of a task they are learning. Implicit learning also tends to occur gradually and automatically through multiple encounters with a stimulus or situation. (e.g., learning to draw an image while only viewing it in a mirror). The following presents a brief introduction to several implicit learning paradigms. This is by no means an exhaustive review; rather, these sections should serve to introduce these concepts in order to relate them to the current work.

This section will focus selectively on implicit *learning* paradigms; that is, paradigms that typically involve acquisition of some skill or habit. This is in contrast to implicit *memory* paradigms, whereby the learning is typically expressed in terms of priming (Jacoby, 1991; McDermott & Roediger, 1994). Although the two topics fall under the general rubric of nondeclarative memory, several nuances warrant separation of the two topics for the purpose of this discussion. One critical difference concerns the state of the knowledge being assessed. Implicit *learning* involves the incidental construction of new knowledge, whereas implicit *memory* concerns the incidental activation of existing knowledge structures (Pothos, 2007). Because the work discussed here presumably involves gradual learning across many trials, it falls more in line with the description of "implicit learning" described above.

Classification Learning

Classifying and categorizing stimuli into groups is a useful skill for navigating the world. For example, categorizing a pan as a sauté pan or wok determines what foods the items is best used to prepare. Likewise, categorizing an individual as familiar or novel determines whether you decide to engage or interact with them. In many cases, the rule(s) for classifying something are verbalizable and available to conscious knowledge. These are in contrast to other cases when the classification rule(s) may be difficult or impossible to verbalize or totally unavailable to conscious introspection. This section will focus on three major types of classification learning task that generally fall into the latter category: artificial grammar learning, information-integration perceptual categorization, and probabilistic classification (e.g., the weather prediction task).

Artificial grammar learning is one of the most common paradigms used to study implicit learning. An artificial grammar consists of a set of rules that are used to generate sequences of symbols (strings). Strings are formed by iterating through the rules and generating output. Strings are considered grammatical if they conform to the rules of the grammar. For example, one could define a rule that strings that start with the letter K must always be immediately followed by the letter Q. Under this grammar, "KQZZF" may be considered grammatical, but "KZQ" would not. A typical artificial grammar task consists of a learning phase and a testing phase. During the learning phase, participants observe several strings generated by the grammar without any knowledge that the strings are arising from a finite set of rules; often they are told to recall letter strings to some performance criterion as an incidental learning task. During the test, participants are informed that the letter strings arise from a complex set of rules, and are asked to classify novel strings as grammatical or not. Participants are able to discriminate grammatical

from non-grammatical strings above chance, despite often reporting no explicit knowledge of the rules or task parameters (Pothos, 2007; Reber, 1968). Interestingly, learning often occurs without feedback; thus observers abstract implicit knowledge during the initial exposure phase incrementally without trial-and-error learning.

As stated above, one hallmark of artificial grammar learning is that observers learn the underlying grammar without conscious attempt to specifically do so; grammar learning is a byproduct of processing the stimuli in some other manner (e.g., studying and recalling specific strings). In fact, the abstraction acquired appears to be a construct largely outside of explicit memory (Knowlton, Ramus, & Squire, 1992; Knowlton & Squire, 1994). In one study by Knowlton, Ramus, and Squire (1992), amnesiac participants and healthy controls performed equally well on artificial grammar classification, but amnesiac participants performed worse than controls when their recognition memory for studied exemplars was tested. In a separate task, Knowlton, Ramus, and Squire (1992) trained participants on an artificial grammar but gave them instructions that encouraged an explicit recall strategy during the classification task (e.g., "say 'yes' if the item seems familiar or reminds you of one of the items you saw earlier"). When the instructions were phrased explicitly, amnesiac participants performed worse than healthy controls at grammar classification, in spite of their prior equivalent performance (see also Reber, 1976 for a similar example in healthy controls).

In artificial grammar tasks it appears that observers learn the featural regularities of the items without necessarily being explicitly aware of said regularities. In this sense, artificial grammar learning resembles another type of classification learning paradigm known as information-integration categorization. In information-integration categorization, like artificial grammar learning, an observer must identify stimuli as belonging to one of several (typically

two) distinct classes. In this case, the stimuli are composed of several orthogonal perceptual dimensions: for example, line segments that vary in length and orientation of tilt. In information-integration tasks, an observer must classify stimuli into one of two classes based on the integration of multiple orthogonal dimensions (see Figure 3). The categorization rule is typically difficult or impossible to verbalize because the perceptual category is defined by a continuous integration of the two orthogonal attributes. This harkens back to the definition set forth by Reber (1989) that implicit learning contrasts with explicit learning in the level of conscious awareness of the regularities underlying the learning. Considering the example in Figure 3, one might be able to describe the category bound as, "for every unit increase in orientation, increase one unit in length; different classes fall on either side of this bound." Such a description would be exceedingly difficult to spontaneously acquire and verbalize by the casual observer. Nevertheless, observers can gradually learn to adequately classify stimuli such as these, but the learning is heavily feedback-dependent and difficult to verbalize (Ashby & Maddox, 2005). However, in both tasks, once learning has occurred, new exemplars are then successfully categorized.

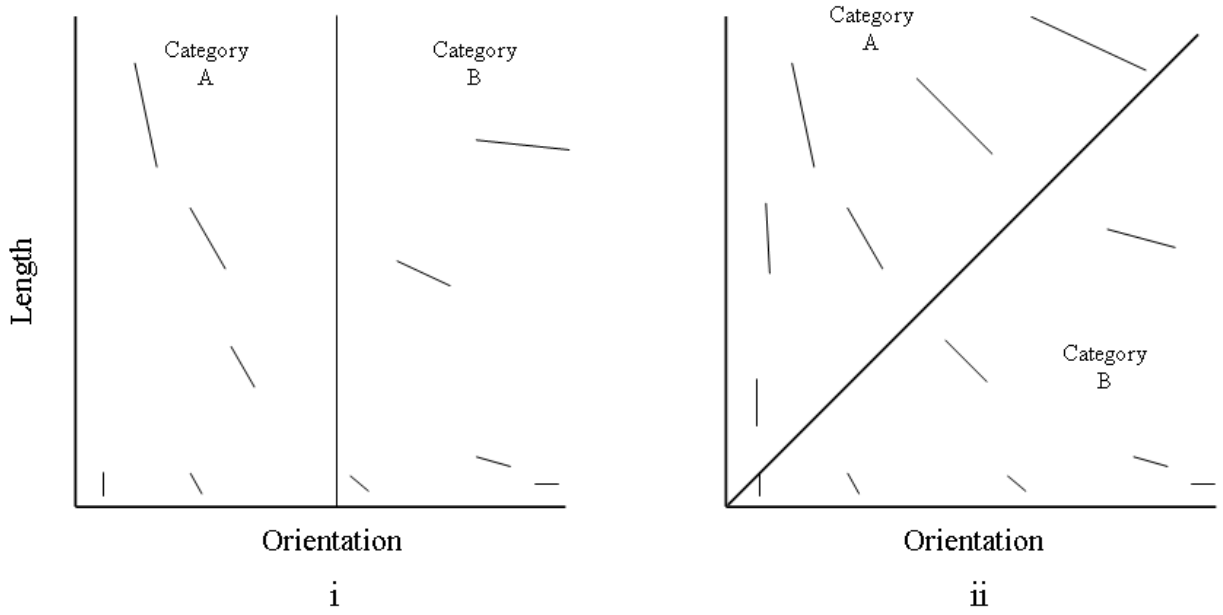


Figure 3: Rule-based versus information-integration perceptual categorization tasks. In the rule-based task (left), stimuli are categorized based on one stimulus dimension (orientation). In the information-integration task, stimuli are categorized based on a linear combination of line length and line orientation.

In contrast to information-integration tasks, some forms of perceptual category learning can occur with the support of explicit, verbalizable strategies. For example, in rule-based categorization tasks (categorization tasks that typically rely on explicit knowledge systems and are likely NOT implicit; see Ashby & Maddox, 2005), observers can adopt and use a simple categorization rule if they are supplied one. Provided this rule does not significantly hamper performance (i.e., violations of the rule are not too egregious), observers will continue to classify according to this rule (Allen & Brooks, 1991). In contrast, in information-integration tasks, explicit knowledge does not aid in categorization performance, and learning to properly categorize stimuli is driven by feedback-based trial-and-error learning. For example, Ashby, Queller, and Berretty (1999) tried to use explicit instructions to aid observers in solving an

information-integration categorization task. In the task, participants had to classify line stimuli into one of two arbitrary categories based on the length and orientation of the line segments (a la Figure 3). Participants were instructed that perfect performance was only possible if both sources of information were used. Nevertheless, such instructions alone were not sufficient for participants to optimally classify the stimuli; high classification performance was only achieved when error feedback was provided during training. This study underscores the dichotomy between explicit and implicit systems: even with conscious volition and the explicit knowledge of how to properly perform the task, proper learning did not occur without the aid of feedback.

In the case of information-integration and artificial-grammar learning, the mapping between features and appropriate categories is deterministic. That is, the same combination of features always belongs to the same category across exposures. In contrast, in so-called probabilistic classification tasks the outcomes are not deterministic; that is, a given stimulus will not always predict the same decision outcome (i.e., responding 'category A' to a given stimulus will be correct only 75% of the time). Because the outcomes are not deterministic, explicit strategies are usually not helpful; responding is driven largely by feedback probabilistically reinforcing a given decision. Like information-integration categorization, probabilistic classification is learned gradually across many trials through trial-and-error. One example of such a task is the so-called “weather prediction task,” a probabilistic categorization task. During the weather prediction task, participants are shown one or more cards with unique geometric designs, and are asked to decide whether the combination of cards represents a “sunny” or “rainy” forecast (Poldrack et. al., 1999). Each card is probabilistically associated with a given weather outcome, and participants gradually learn to associate the most probable weather outcome with each card combination throughout the course of the task. Performance often hovers

around chance for the first several trials and gradually increases across several blocks. Although it is often considered to be an exclusively implicit task, observers will often report a variety of generally viable strategies when approaching this task. Interestingly, observers who claim to be guessing perform just as well as those who are able to verbalize some kind of strategy (Gluck, Shohamy, & Myers, 2002), suggesting that both cognitive strategies and implicit learning can play a role in this type of categorization task, in contrast to information-integration categorization.

This point leads to a general criticism of some of the classification learning literature. Although incidental encoding (e.g., artificial grammar learning) leads to tacit knowledge and improved performance in these tasks, explicit strategies can often lead to above chance performance as well. For example, memorizing stimulus-outcome pairings in the weather prediction task can lead to adequate classification performance (Gluck, Shohamy, & Meyers, 2002). In artificial grammar learning, if participants are provided with a structured version of the grammar (e.g., presenting like grammatical strings together during training), they appear to explicitly learn the grammar as well as those who attempt the task in an implicit manner (Reber, Kassin, Lewis, & Cantor, 1980). Thus it is unclear that performance in implicit classification tasks is always driven by implicit learning - some observers could be using explicit knowledge systems to the same effect without experimenter knowledge. This would be most likely to occur in scenarios where stimuli are repeated often enough to lead to memorization of explicit examples, as in the weather prediction example described above.

Motor Sequence Learning

Motor sequence learning is another often-used putatively implicit learning task. In a typical motor sequence learning task, participants are asked to press specific keys after receiving seemingly random cues on screen (for example, press the '1' key when a '1' appears on screen). Unbeknownst to the observer, a sequence is embedded within randomly sorted cues; this sequence follows a complex but predictable succession. Participants demonstrate learning in this task via faster and more accurate responses on sequenced versus unsequenced stimuli following training.

One critical characteristic of this sort of learning is that it is heavily tied to the manner of responding. That is, observers learn to issue a particular motor plan to a particular cue rather than an underlying structure to the cues. Critically, observers are not mapping a specific motor *response* onto a particular stimulus (e.g., "index finger on left hand after seeing a blue square") but rather a specific *distal response location* (e.g., "press blue button after seeing a blue square"). Thus, switching distal response locations should disrupt learning in this sort of implicit learning task (Willingham, 1999). This was demonstrated by Willingham, Wells, Farrell, and Stemwedel (2000). In their experiment (Experiment 2), participants responded to cues presented on screen by pressing a corresponding button on a response box. Participants first trained on a sequence of cues. During transfer, participants either responded using the same fingers but new response locations (Fingers condition) or new fingers but the same response location (Locations condition). Participants in the Locations condition showed significantly better transfer when presented with the originally learned sequence as evidenced by faster responses to sequenced cues as compared to participants in the Fingers condition. These findings support the notion outlined earlier that participants learn a fairly abstract series of response locations during this task rather than specific motor sequences. A given cue (or sequence of cues) is associated with

the specific motor plan; disrupting that motor plan serves to disrupt learning. This idea will be tested in the current work in Experiment 3 (see Chapter 7).

Because the learning appears to be an abstract series of response locations, it is independent of the actual stimuli. In other words, motor sequence learning transfers robustly when the surface features change but the underlying sequence remains intact. Willingham (1999) demonstrated just such a property. During training, the digits 1-4 appeared in a single box on the center of the screen, the given digit indicating which of four buttons to press (e.g., '1' on screen, press first button). During transfer, four individual boxes on screen corresponded to each of the four buttons. The presence of an asterisk within a box indicated that corresponding button should be pressed. Critically, the sequence remained the same during transfer despite the grossly different surface features. Participants showed faster reaction times when responding to the sequence than to random stimuli during transfer, despite the visual details of the stimuli changing vastly. This suggests, like artificial grammar learning, this form of implicit learning results in a fairly abstract representation of what is learned.

General Trends in the Acquisition of Implicit Knowledge

Some general assertions can be made that are germane to the experiments presented later. Generally, explicit intention has either limited (Ashby, Queller, & Berretty, 1999) or sometimes detrimental effect on implicit learning (Knowlton, Ramus, & Squire, 1992; Reber, 1976). Implicit learning sometimes transfers to novel stimuli (Reber, 1967; Willingham, 1999), while other times the learning is confined to the particular stimuli encountered (Ashby & Waldron, 1999; Seger, 2008; Poldrack & Packard, 2003); this likely represents the structure of the actual learning (e.g., abstract knowledge vs. stimulus-response associations). Finally, some forms of

learning tend to be particularly sensitive to contextual cues (Graybiel, 2008; Neal, Wood, Labrecque, & Lally, 2012). How do these general trends relate to the notion that recognition criteria can be implicitly influenced via feedback outlined earlier? It is clear that although the paradigms may differ substantially in many core aspects, all seem to suggest that implicit learning produces knowledge that is fairly abstract with respect to the stimuli and decisions upon which it operates. This sort of abstraction is important to reflect upon when considering whether similar learning mechanisms may regulate the mappings between memory representations and memory judgments. The next chapter will outline a framework for understanding how such mechanisms may operate within the domain of recognition decisions.

Chapter 3

Implicitly Learning to Evaluate Explicit Memories

The brief review of implicit learning above highlights several regularities that also occur in false positive feedback (FPF) studies discussed briefly earlier (Han & Dobbins, 2008, 2009). Namely, these shifts appeared to occur in an incremental fashion and largely outside of observer awareness. Before proceeding to the methodology underlying the FPF criterion shifting paradigm designed by Han and Dobbins (2009), I briefly note some of the characteristics of more formal math models of reinforcement learning mechanisms that will be relevant to the predictions of the current experiments. One family of reinforcement learning models relies on computational algorithms known as temporal difference algorithms (Dayan & Daw, 2008; Sutton & Barto, 1981). Roughly speaking, these algorithms require an organism's decision-making system to predict the value of an action-linked reward. If that prediction does not match the outcome, the prediction is adjusted to better approximate the obtained reward. Learning occurs because specific actions – and stimuli associated with that action – begin to be associated with more accurate reward predictions.

Under the temporal difference framework, learning is driven by errors of prediction (thus, "prediction error"). Larger prediction errors lead to larger adjustments of expectations (and behavior), eventually honing in on the optimal behavior for a given situation. How does this relate to the proposed notion of implicit criterion learning and the procedures used to induce it? It is interesting to note that under the temporal difference framework described above, learning does not occur unless a prediction error is made – that is, a *new* stimulus-response contingency will never be formed if the system makes accurate predictions about the outcome of said contingency. Within the FPF paradigm, participants receive unexpected positive feedback for a

subset of one type of recognition error (e.g., false alarms or misses). Because these classifications are typically made with low confidence, it is reasonable to assume that many observers expect these decisions to be incorrect – in other words, observers are expecting to receive negative feedback for a majority of these decisions based on the quality of the memory evidence they have. Thus, the unexpected positive feedback results in a prediction error that should serve to drive an association between a given level of memory evidence and a certain classification tendency – the end result is a change in evidence-to-decision mappings.

The following experiments sought to better understand a putatively implicit learning process that governs how observers attribute given levels of familiarity to recognition decisions. The experiments use the FPF procedure designed by Han and Dobbins (2008, 2009) and are based on questions naturally arising from the similarity between the FPF paradigm and the implicit learning research and prediction error framework discussed above – namely, that positive feedback for errors should produce an unexpected positive outcome, akin to a prediction error in reinforcement learning. Using the FPF paradigm, the following experiments focused on several questions: can observers volitionally inhibit the influence of FPF? Does FPF produce subsequent memory benefits for items that received unexpectedly positive feedback? Does implicit criterion learning generalize across different types of stimuli? Does implicit criterion learning transfer across different testing contexts and response procedure? Finally, in accordance with a temporal difference framework of reinforcement, do unexpected positive outcomes drive responding in this paradigm more than expected positive outcomes?

Chapter 4

Specific Aims and General Methods

The specific aims of the three experiments are listed below. These will be repeated and expanded upon in the relevant chapters.

1) Can observers inhibit the effects of FPF on the decision criterion, or does FPF influence the criterion despite an individual's volition (Chapter 5)?

2) Are items that receive FPF subsequently remembered better than items receiving veridical feedback (Chapter 5)?

3) Would FPF-based criterion learning on one stimulus type generalize to other types of stimuli (Chapter 6)?

4) Would a radical shift of context eliminate an implicitly acquired response bias (Chapter 7)?

5) Do unexpected positive outcomes drive responding more than expected positive outcomes (Chapter 8)?

GENERAL METHODS

Before turning to the specific methods used in Experiment 1, I briefly outline those that are general across all of the presented studies.

Participants

Participants earned either $15 or partial credit toward fulfillment of a course assignment for participating. Informed consent was obtained for all participants, as required by the university's institutional review board. Due to the minor deception involved in providing false positive feedback, participants were fully debriefed on the nature of the feedback after the study and given the option to withdraw their data if they desired. All participants chose to have their data included.

Materials

All experiments were generated via The Psychophysics Toolbox (version 3.0.8) (Brainard, 1997; Pelli, 1997) implemented in Matlab using a standard keyboard and PC. Unless otherwise noted, stimuli were drawn from a pool of 1216 words selected from the MRC Psycholinguistic Database (Wilson, 1988). Words in the pool had an average of 7.09 letters, 2.35 syllables, and an average log HAL frequency of 7.74 log. This word pool was used for all experiments. Target (studied test items) and lure (novel test items) lists were formed by randomly sampling from this word pool for each participant. All recognition tests contained an equal number of targets and lures.

Feedback

When present, feedback was presented following the individual's confidence report. When positive feedback was provided, the word "CORRECT" flashed in green on screen. When

negative feedback was provided, the word “INCORRECT” flashed in red on screen. When veridical feedback was provided, positive feedback followed hits and correct rejections (correct “old” and “new” reports, respectively), and negative feedback followed false alarms and misses (incorrect “old” and “new” reports, respectively). During the FPF manipulation, hits and correct rejections always receive correct, veridical feedback; it is the nature of the feedback given to errors that constitutes the manipulation. When “liberal” FPF was provided, a percentage of false alarms were incorrectly signaled as correct responses (70%) whereas all misses were correctly identified as errors. When “conservative” FPF was provided, the feedback percentages for errors were reversed with 70% of misses being incorrectly signaled as correct whereas all false alarms were veridically signaled as errors. As noted in the introduction, the purpose of restricting the manipulation to errors and giving it probabilistically is to increase the likelihood that the feedback manipulation does not become apparent to the observer. Since observers are typically not confident during error trials, they should ascribe occasional positive feedback as simply reflecting ‘lucky guessing.’

Subjective Awareness Questionnaire

The Subjective Awareness Questionnaire consisted of several questions to gauge participants’ subjective experiences that may correspond to bias, accuracy, and awareness of the purpose of feedback. Parts of this questionnaire were administered following each recognition test, and the full questionnaire was administered following all recognition tests (see specific Experiments for more detailed information). More specifically, participants were asked the following:

1. The relative proportion of studied items in the test(s) taken (0% - 100%).

2. Personal test performance (percentage of correct responses overall)
3. The believed purpose of the feedback (open response).
4. Whether there was anything anomalous about the feedback (yes or no).
    a. If yes, what (open answer).
5. Level of influence of positive feedback following an "old" response (6 point Likert scale).
6. Level of influence of positive feedback following a "new" response (6 point Likert scale).
7. Level of influence of negative feedback following an "old" response (6 point Likert scale).
8. Level of influence of negative feedback following a "new" response (6 point Likert scale).
9. Please select one of the following options regarding the feedback you received during the experiment:
    a. The feedback was occasionally inaccurate when I correctly responded "old."
    b. The feedback was occasionally inaccurate when I correctly responded "new."
    c. The feedback was occasionally inaccurate when I incorrectly responded "old."
    d. The feedback was occasionally inaccurate when I incorrectly responded "new."
    e. I don't agree with any of the above statements.

Personality Measures

Several personality measures were collected during each experiment in order to evaluate potential individual differences in feedback influence. These measures were chosen for their relation to reward and punishment sensitivity: the BIS/BAS scales (Carver & White, 1994); the Regulatory Focus Questionnaire (RFQ; Higgins et. al., 2001); and the Generalised Reward and Punishment Expectancy Scales (GRAPES; Ball & Zuckerman, 1990). The BIS/BAS scales were also chosen specifically because prior work using the FPF paradigm showed the BIS and Reward Responsiveness subscales both correlated with degree of criterion change in the FPF paradigm (Han, 2009). Thus it was hoped to replicate and extend these findings using other measures of reward sensitivity.

Chapter 5

Experiment 1

1) Can observers inhibit the effects of FPF on the decision criterion, or does FPF influence the criterion despite an individual's volition?

Evidence suggests that criterion shifts achieved through the FPF procedure are not apparent to participants (Han & Dobbins, 2009) and thus may rest on a largely implicit reinforcement learning phenomenon. If correct, observers should have little to no conscious control over the effects of FPF on the criterion – that is, even when instructed to ignore the potential influence of feedback, observers should still demonstrate robust criterion shifts in response to the manipulation because the effect does not rest on the formulation of an explicit judgment strategy (see Figure 4).

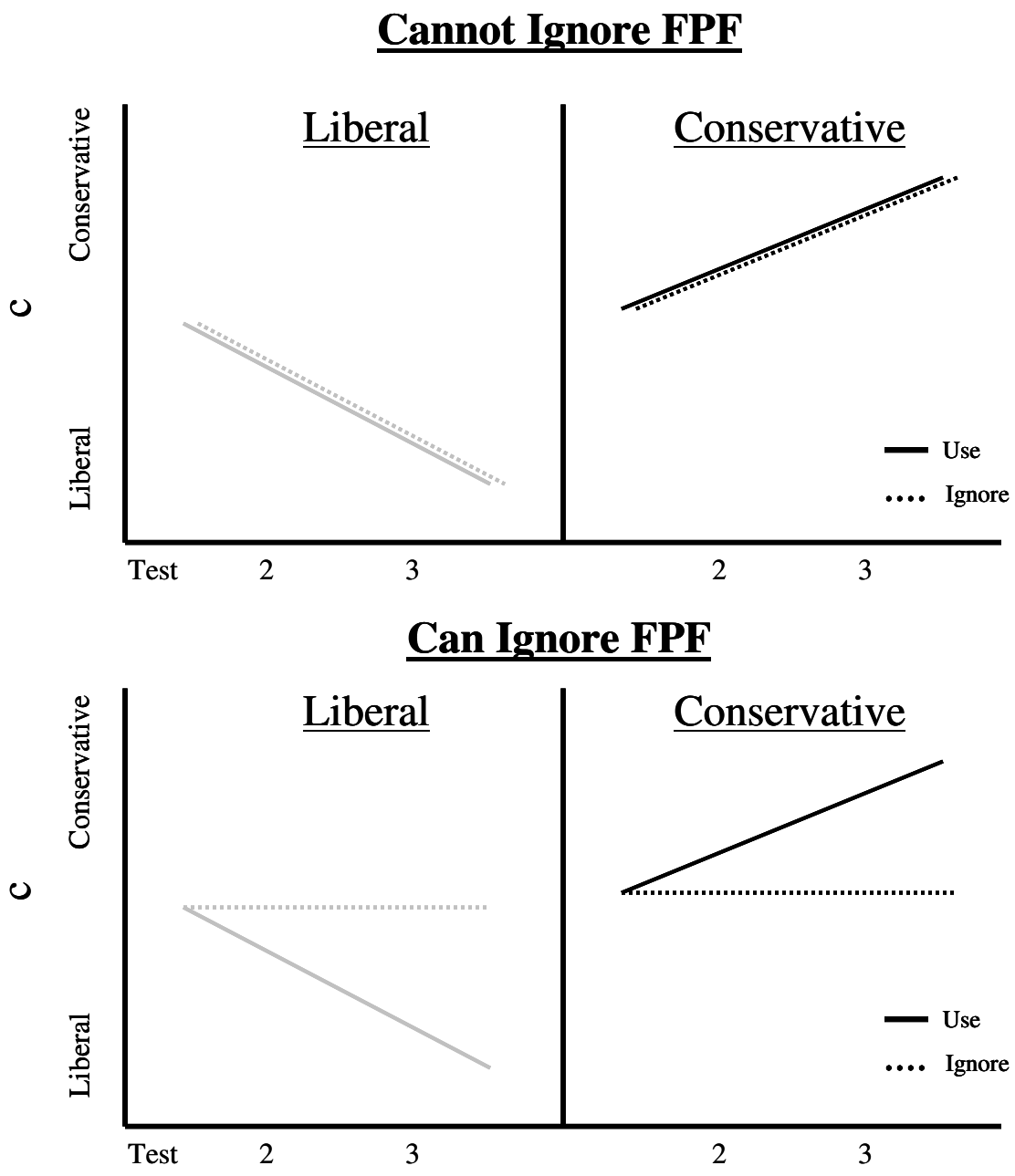


Figure 4: Hypothetical experimental results. The top panel (predicted results) shows hypothetical results if participants are unable to ignore the influence of FPF; criterion should not differ between Use and Ignore groups. The bottom panel shows hypothetical results if participants ARE able to ignore the influence of FPF; in this case, criterion should differ between Use and Ignore groups that receive the same feedback.

2) Are items that receive FPF subsequently remembered better than items receiving veridical feedback?

If the FPF effect reflects reinforcement learning it should demonstrate certain hallmarks. From a neurobiological perspective, reinforcement learning is often held to critically depend on prediction error signals generated within the dopaminergic system, with increased learning on trials in which outcomes and predictions surprisingly diverge (Dayan & Watkins, 2001; Schultz & Dayan, 1997; Schultz & Dickinson, 2000). A similar cognitive phenomenon is referred to as the 'hypercorrection effect' (Butterfield & Metcalfe, 2006) in which errors issued with high confidence that receive immediate corrective feedback yield durable changes in subsequent judgments. Error confidence tends to positively correlate with later veridical memory; in other words, correcting a mistake made with high confidence produces better memory for the correct answer than correcting a low confidence mistake. Because hypercorrection effects are linked to highly confident initial errors they may reflect dopaminergic modulation of episodic memories for large prediction error outcomes, and there is some initial fMRI evidence that reward-related activity within the dopaminergic system yields subsequent memory benefits for reward-related stimuli (e.g., Adcock et. al., 2006). If surprising mistakes are remembered well, it stands to reason that surprising correct responses (falsely reinforced errors) should also appreciate a mnemonic benefit.

If the effects of FPF reflect an implicit learning phenomenon that gradually alters the subject's mapping of internal evidence signals onto overt outcomes, it is anticipated that subjects will not be able to inhibit the influences of FPF on their measured decision criterion. Furthermore, because the FPF effect relies on providing feedback during trials in which confidence is low (viz. errors) it may be the case that feedback falsely signaling to the subject

that he or she is correct is sufficiently surprising to yield dopaminergic modulations of memory. That is, subjects may find these trials, and hence the stimuli, reliably more memorable in subsequent testing.

Participants

103 individuals participated in return for partial course credit. Participants were excluded if their performance on any single test was at or below chance (i.e., if d' $\leq$ 0). Using this criterion, two participants were removed from the analysis, making the effective sample size 101 (26 participants in the Liberal/Use group, 25 participants in all others).

Procedure

The experiment consisted of three study/test cycles, followed by a fourth surprise recognition test where subsequent memory for the prior test materials was assessed. During study, 100 words were serially presented and participants rated the number of syllables in each. Participants had two seconds to respond to each item; if this time was exceeded, the words 'TOO SLOW' appeared on screen and the computer moved onto the next trial. During test phases, 100 targets and 100 lures were randomly intermixed and presented serially; participants first indicated whether each item was old or new, followed by a confidence report ("low," "medium," or "high" confidence). Test responding was self-paced. Each test was followed by questions 1-2 of the Subjective Awareness Questionnaire. Following the final test, participants completed the full Subjective Awareness Questionnaire and the BIS/BAS at their testing computers. Finally, participants completed paper copies of the GRAPES and RFQ before finishing the experiment.

Feedback was present during the first three recognition tests. The first test was used to assess baseline recognition characteristics in all observers and during this test all participants received veridical feedback after every test response. The second and third test examined the FPF effect under two different instructions conditions (Use versus Ignore) crossed with two different bias directions (Liberal versus Conservative FPF) manipulated between groups (creating a 2×2 between-subjects factorial design). Instructions for the Use and Ignore groups were identical other than a few key sentences. More specifically, participants in the Use groups were told:

> "We are particularly interested in how well you can incorporate this feedback. The feedback provided may improve your performance. Please try to incorporate the feedback as best you can when responding."

In contrast, participants in the Ignore groups were told:

> "We are particularly interested in how well you can ignore this feedback. The feedback provided may impair your test performance. Please try to IGNORE the feedback as best you can and do not let it affect your responding."

During test, the words "USE FEEDBACK" or "IGNORE FEEDBACK" were always present at the top of the screen (depending on the participant's particular instructions).

Participants were given a surprise final recognition test following the third test phase. Target items for this test were drawn from the second recognition test phase (items were drawn from the same test to control for average memory strength of test items); specifically, any items for which the participant could have received FPF. For participants who received Liberal FPF, targets for the fourth test consisted of the 100 prior lures from the second test (since false alarms were reinforced); for participants who received Conservative FPF, targets for the fourth test

consisted of the 100 prior targets from the second test (since misses were reinforced). 100 novel items were used for lures for all groups. Participants were instructed that, "any item seen before in this experiment, regardless of which test and how [they] classified it, should be judged old." Feedback was not provided during this final test, and participants were informed that they would no longer be receiving feedback (see Figure 5 for design).

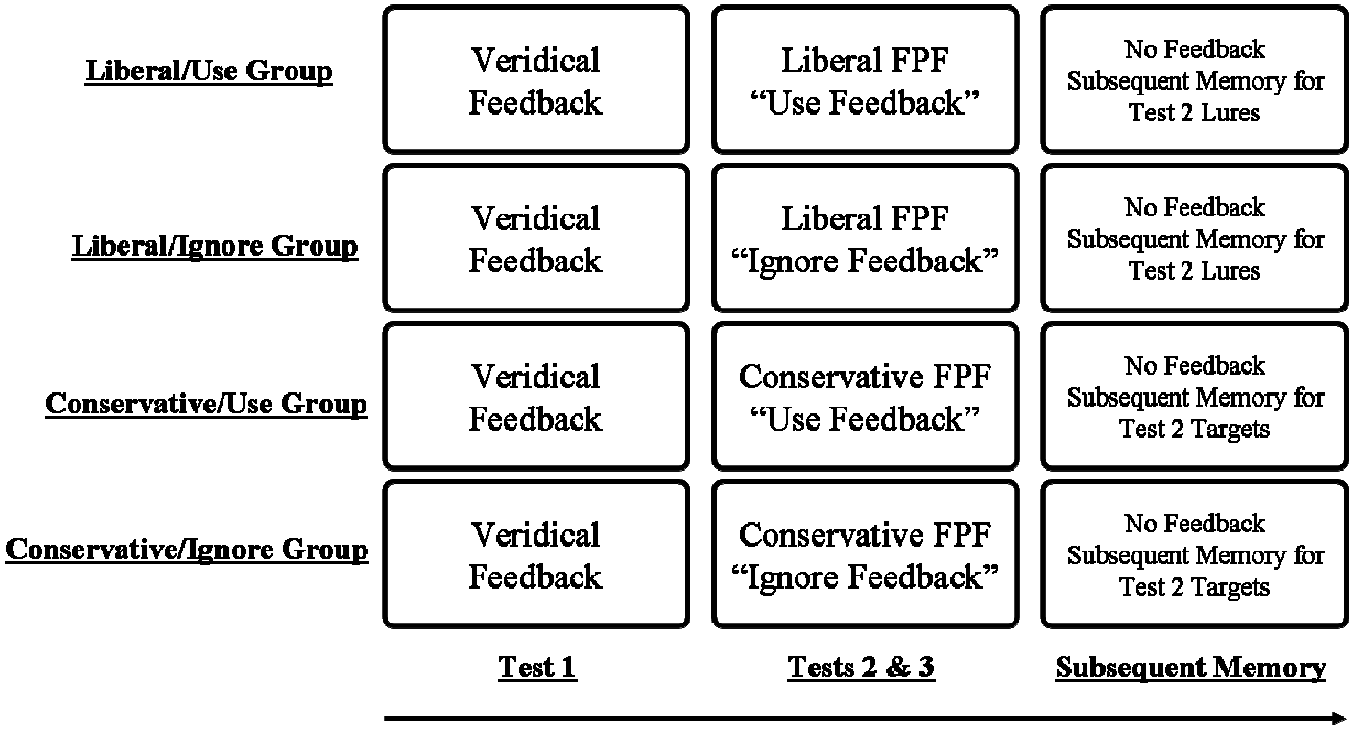


Figure 5: Experimental Design. Participants were divided into four groups. All groups received veridical feedback during Test 1. Groups received either Liberal or Conservative FPF and were told to "Use" or "Ignore" this feedback during Tests 2 and 3. A surprise subsequent memory test followed Test 3, testing over particular items drawn from Test 2.

Results

Hit rates, FA rates, d', criterion (c), and average number of FPF trials are presented in Table 1. Since the groups were not treated differently during Test 1, analyses of data from Test 1

are simply to confirm that the groups had similar accuracy and bias prior to the implementation of the key experimental manipulations in Tests 2 and 3. The data from Tests 2 and 3 will take advantage of the 2 × 2 factorial design crossing Instructions and Feedback. The Subsequent Memory Test will be separately analyzed to see if FPF influences subsequent memory separately for each group defined by the factorial design in Tests 2 and 3.

	Liberal/Use			N = 26	
Test	Hit Rate	FA Rate	d'	c	FPF trials
Test 1	.74 (.088)	.22 (.077)	1.47 (0.33)	0.082 (0.22)	N/A
Test 2	.81 (.081)	.37 (.14)	1.28 (0.34)	-0.28 (0.31)	26.42 (10.57)
Test 3	.81 (.079)	.44 (.17)	1.09 (0.31)	-0.39 (0.38)	31.38 (13.36)
Subsequent Memory	.76 (.10)	.43 (.12)	0.95 (0.42)	-0.30 (0.30)	N/A

	Conservative/Use			N = 25	
Test	Hit Rate	FA Rate	d'	c	FPF trials
Test 1	.76 (.084)	.21 (.099)	1.62 (0.49)	0.061 (0.23)	N/A
Test 2	.73 (.12)	.19 (.096)	1.61 (0.60)	0.13 (0.27)	18.88 (8.58)
Test 3	.67 (.15)	.22 (.11)	1.34 (0.57)	0.18 (0.34)	22.32 (11.11)
Subsequent Memory	.82 (.12)	.22 (.10)	1.83 (0.62)	-0.079 (0.34)	N/A

	Liberal/Ignore			N = 26	
Test	Hit Rate	FA Rate	d'	c	FPF trials
Test 1	.71 (.091)	.20 (.088)	1.44 (0.40)	0.15 (0.23)	N/A
Test 2	.73 (.093)	.28 (.12)	1.25 (0.40)	-0.014 (0.26)	19.48 (8.82)
Test 3	.74 (.12)	.38 (.17)	1.04 (0.52)	-0.17 (0.34)	27.12 (12.36)
Subsequent Memory	.70 (.16)	.35 (.16)	1.01 (0.45)	-0.080 (0.41)	N/A

	Conservative/Ignore			N = 26	
Test	Hit Rate	FA Rate	d'	c	FPF trials
Test 1	.75 (.11)	.23 (.094)	1.48 (0.44)	0.041 (0.23)	N/A
Test 2	.68 (.16)	.22 (.11)	1.32 (0.38)	0.15 (0.39)	22.72 (11.29)
Test 3	.60 (.18)	.21 (.11)	1.13 (0.50)	0.31 (0.38)	28.92 (14.14)
Subsequent Memory	.72 (.18)	.23 (.15)	1.54 (0.56)	0.068 (0.52)	N/A

Table 1: Hit rates, false alarm rates, accuracy, criterion, and number of FPF trials received for the four groups across all tests (Experiment 1). Standard deviations in parentheses.

Accuracy

Because this experiment primarily focuses on criterion, it is important to first establish that there are no significant differences in initial accuracy among the groups because differences in c become difficult to interpret when accuracy also differs across comparisons (Pastore, Crawley, Berens, & Skelly, 2003).

Accuracy (d') during Test 1 was analyzed using a one-way ANOVA on the four groups that would constitute the 2 × 2 design of Tests 2 and 3. There were no significant differences in accuracy among the groups [$F(3,97) < 1$].

Next, accuracy during the second and third tests was analyzed. Test was included as a factor, but since it did not interact with the other two factors, I present a simplified analysis collapsing across this factor using a 2×2 factorial ANOVA examining between-subjects factors of Feedback Group (Liberal vs. Conservative) and Instructions (Use vs. Ignore). This analysis revealed a significant main effect of Feedback type on accuracy [$F(1,97) = 4.52$, $MSe = 0.378$, $p < .05$, $\eta^2 = .044$], which indicated that participants who received Liberal feedback performed worse than did participants who received Conservative feedback [$M_L = 1.17$, $SE_L = 0.061$; $M_C = 1.35$, $SE_C = 0.061$]; this likely represents how accuracy was calculated and will be discussed in more detail in the Discussion section of this chapter. The main effect of Instructions was not significant [$F(1,97) = 2.78$, $MSe = 0.38$, $p = .10$, $\eta^2 = .028$]. The two-way interaction did not

approach significance [$F(1,97) = 1.52$, $MSe = 0.38$, $p = .22$, $\eta^2 = .015$]. None of the interactions among the factors approached significance (all p's > .22).

Performance during the final test will be covered in the "Subsequent Memory Effects" section.

Criterion

The main purpose of this experiment was to investigate whether observers could volitionally inhibit the effects of FPF on the decision criterion. However, before examining the effects of FPF on the criterion, it is important to establish that the criterion was similar for all groups prior to the manipulation. To examine this, criterion during Test 1 was subjected to a one-way ANOVA amongst the (future) groups. This analysis was not significant [$F(3,97) = 1.03$, $MSe = 0.052$, $p = .38$, $\eta^2 = .031$].

Tests 2 and 3 were examined using a 2×2×2 mixed model ANOVA examining factors of Test (Test 2 vs. Test 3), Feedback Group (Liberal vs. Conservative), and Volition (Use vs. Ignore feedback). If observers were able to inhibit the influence of the FPF, this analysis should demonstrate an interaction with the Volition factor, such that changes in criterion are reduced (or eliminated) for participants told to Ignore the feedback. On the other hand, if observers are unable to ignore the influence of the FPF, this factor should not interact (i.e., no differences in criterion change as a function of instructions).

There was no main effect of Test [$F(1,97) < 1$] on overall criterion. There was a robust main effect of Feedback Group [$F(1,97) = 44.29$, $p < .0001$, $MSe = 0.19$, $\eta^2 = .35$], indicating that participants who received Liberal FPF were more liberal than participants who received Conservative FPF [$M_L = -0.21$, $SE_L = 0.05$; $M_C = 0.19$, $SE_C = 0.04$]. There was also a significant

main effect of Volition [$F(1,97) = 6.58$, $p < .05$, $MSe = 0.19$, $\eta^2 = .06$], indicating that participants in the Use groups were significantly more liberal than participants in the Ignore groups [$M_U = -0.089$, $SE_U = 0.043$; $M_I = 0.068$, $SE_I = 0.044$]. There was a robust Test×Feedback Group interaction [$F(1,97) = 19.19$, $p < .0001$, $MSe = 0.039$, $\eta^2 = .16$] that occurred because the effects of FPF increased across the two tests. Across tests, participants who received Liberal FPF became more liberal [$M_{T2} = -0.15$, $SE_{T2} = 0.043$; $M_{T3} = -0.28$, $SE_{T3} = 0.051$; $p < .01$, Tukey's HSD], whereas participants who received Conservative FPF became more conservative [$M_{T2} = 0.14$, $SE_{T2} = 0.044$; $M_{T3} = 0.25$, $SE_{T3} = 0.051$; $p < .05$, Tukey's HSD]. There was no Test × Volition interaction [$F(1,97) < 1$]. Critically the Feedback Group × Volition interaction was not significant [$F(1,97) = 1.87$, $p = .18$, $MSe = 0.19$, $\eta^2 = .019$] which demonstrates that the biasing effects of the FPF manipulation are not eliminated for the groups instructed to ignore the feedback. The three-way interaction was not significant [$F(1,97) = 2.31$, $p = .13$, $MSe = 0.039$, $\eta^2 = .023$].

Although neither the Volition × Feedback Group nor the Volition × Feedback Group × Test interactions were significant, I nonetheless conducted follow-up interaction analyses given the prediction that subjects would not be able to effectively ignore the biasing feedback. The full data are shown in Figure 6 and I separately consider Volition and Test factors for the Liberal FPF manipulation (left panel) and the Conservative FPF manipulation (right panel). Focusing first on the groups exposed to Liberal FPF, a Volition × Test ANOVA yielded a main effect of Volition [$F(1,49) = 8.80$, $MSe = 0.17$, $p < .01$, $\eta^2 = .15$] which indicated the Use group was more liberal than the Ignore group [$M_{LU} = -0.34$, $SE_{LU} = 0.057$; $M_{LI} = -0.09$, $SE_{LI} = 0.058$]. The main effect of Test was also significant [$F(1,49) = 10.00$, $MSe = 0.045$, $p < .01$, $\eta^2 = .17$] which indicated that both groups became more liberal across tests [$M_{T2} = -0.15$, $SE_{T2} = 0.040$; $M_{T3} = -$

0.28, SE_{T3} = 0.051]. Thus, both groups demonstrated an increasingly liberal criterion as the manipulation continued across tests. However, the Use group was generally more liberal than the Ignore group, suggesting that the feedback had a more prominent effect on their behavior. Turning to the groups exposed to Conservative FPF, a second Volition × Test ANOVA yielded no effect of Volition [$F(1,48) < 1$], but a significant effect of Test [$F(1,48) = 9.35$, $MSe = 0.032$, $p < .01$, $\eta^2 = .16$] which indicated that both groups became more conservative across tests [M_{T2} = 0.14, $SE_{T2} = 0.048$; $M_{T3} = 0.25$, $SE_{T3} = 0.051$]. This asymmetry in the effects of Volition explain why it did not survive the omnibus ANOVA above. In this case, when feedback reinforced false alarms, instructions to use or ignore feedback influenced criterion placement. However, the lack of a Volition × Test interaction in the Liberal feedback groups suggests a gradually shifting criterion that shifts at the same rate for both groups.

This is further reinforced by an analysis of Feedback Group and Test focused solely on the groups told to Ignore Feedback (dotted lines Figure 6). This analysis revealed a main effect of Feedback [$F(1,48) = 13.14$, $MSe = 0.20$, $p < .001$, $\eta^2 = .21$] and a significant interaction [$F(1,48) = 15.78$, $MSe = 0.042$, $p < .001$, $\eta^2 = .25$; see dotted lines in Figure 6]. Thus, even when told to ignore feedback, both groups appear to still learn from in it in a gradual manner as evidenced by the criterion differences between Liberal/Ignore and Conservative/Ignore groups.

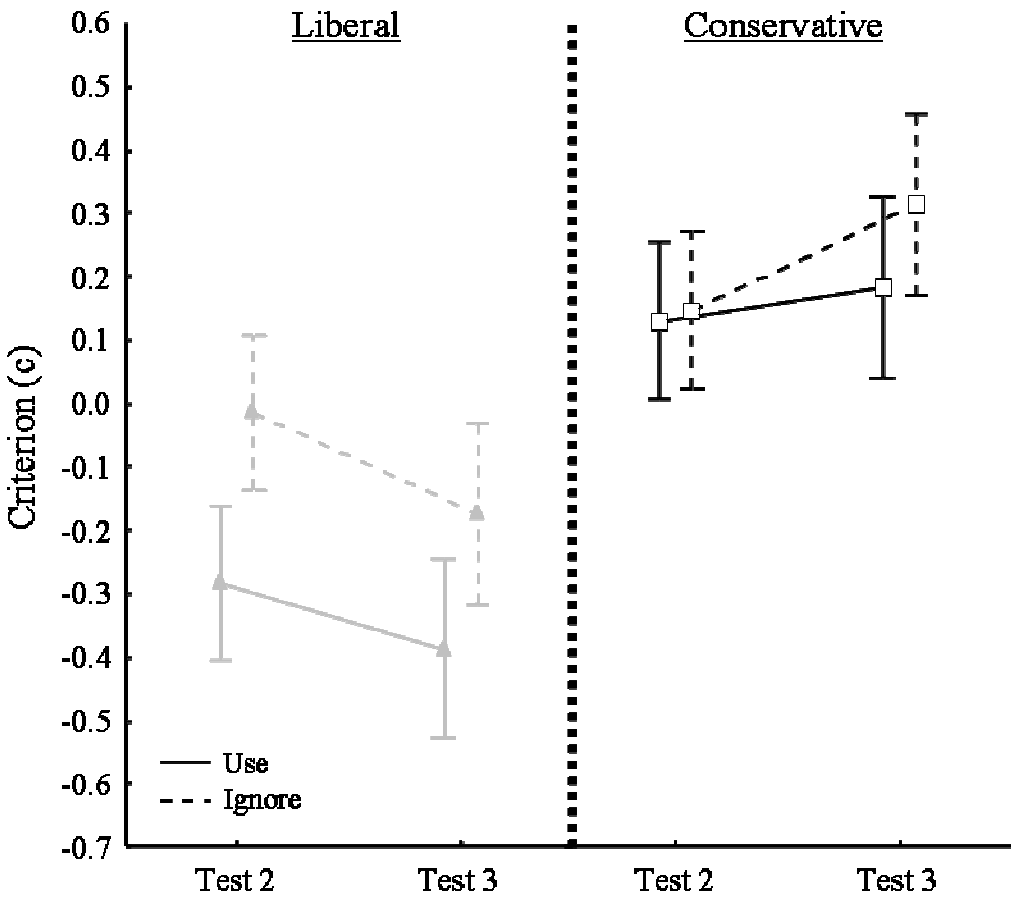


Figure 6: Effects of Feedback and Instructions on criterion in Tests 2 and 3 (Experiment 1). Instructions to ignore feedback do not appear to eliminate its effects (vertical bars represent 95% confidence intervals).

Subsequent Memory Effects

It was hypothesized that participants might have better later memory for items that receive FPF than for items that receive negative feedback due to the unexpected positive outcome associated with FPF. This hypothesis was tested by examining the hit rates for the Subsequent Memory items (which were drawn from Test 2), separated out by their prior feedback status – prior errors that received negative feedback and prior errors that received (false) positive feedback. These hit rates were calculated as the conditional probability that an

item is called “old” during the Subsequent Memory Test given its prior outcome [e.g., p(“old”|prior FPF FA)]. Thus for example, consider subjects who were told to use the feedback and exposed to the Liberal FPF manipulation. For these subjects new materials from Test 2 that yielded false alarms were re-presented during the Subsequent Memory Test as recognition targets, and I contrasted the tendency to correctly recognize items that were previously correctly identified as false alarms versus those that were falsely identified as hits (viz. received FPF). In other words, if 20 false alarms had received prior positive feedback during Test 2, and an observer identified 10 of those items as previously encountered during this final test, his hit rate for prior false positive errors would be .50. Analogous contrasts were conducted for the three remaining groups of subjects. Table 2 shows the relevant conditional hit rates for each group.

Hit rates were contrasted for each of the four groups. While none of the individual contrasts survived (all p’s > .11), each group demonstrated a small benefit for Prior FPF Errors as compared to Prior Negative FP Errors (see Table 2). It should be kept in mind however that the analysis is necessarily restricted to prior errors, which limits the item counts available for the participants. Due to the more exploratory nature of this analysis and to increase power, I collapsed across all four groups and contrasted the two subsequent hit rates. In this case, the hit rate was significantly higher for prior errors that received false positive feedback as compared to prior errors that received veridical negative feedback [$M_{NFE} = .69$, $M_{PFE} = .73$; $t(100) = -2.28$, $p < .05$]. This finding will be discussed in more detail in the Discussion.

Subsequent Memory Conditional Hit Rates

Group	Prior Negative FB Error	Prior FPF Error
Liberal/Use	.79 (.17)	.82 (.12)
Conservative/Use	.66 (.26)	.71 (.19)
Liberal/Ignore	.74 (.22)	.79 (.17)
Conservative/Ignore	.56 (.32)	.61 (.26)

Table 2: Conditional hit rates for Subsequent Memory items. Standard deviations in parentheses.

Duration of Learned Biases

As noted in Han and Dobbins 2008 and 2009, biases acquired via FPF tend to persevere even when feedback is removed. Here I provide another test of this claim using data from the Subsequent Memory Test. Because the hit rates are comprised of items of two different classes (namely prior errors to old items for Conservative Groups and prior errors to new items for Liberal Groups) I did not use c to examine whether biases instilled in the prior tests carried over into the Subsequent Memory Test. Instead, I used the false alarm rates, which are perfectly matched for all four groups in terms of prior exposure, consisting of novel items shown for the first time in the experiment (See Table 1). The false alarm rates were analyzed using a 2 × 2 factorial ANOVA, which examined factors of Prior Feedback Group (Liberal vs. Conservative) and Prior Volition (Use vs. Ignore). There was a robust effect of Prior Feedback [$F(1,97) = 38.10$, $MSe = 0.018$, $p < .0001$, $\eta^2 = .28$], which indicated that participants who previously received Liberal feedback were more liberal (i.e., false alarmed more often) than those who previously received Conservative feedback [$M_L = .39$, $SE_L = .019$; $M_C = .22$, $SE_C = .019$]. The effect of Volition was not significant, but the interaction approached significance [$F(1,97) = 2.91$, $MSe = 0.018$, $p = .09$, $\eta^2 = .029$]; this reflected the fact that the Liberal/Use participants were more liberal than the Liberal/Ignore participants (see Table 1), as in Tests 2 and 3. These

findings confirm that FPF effects persevere once feedback is removed and indeed this savings is even evident for observers who were instructed to ignore the previous feedback.

Confidence

Analysis of confidence was restricted to correct reports. Confidence during Test 1 was analyzed using a 4×2 mixed model ANOVA examining a between-subjects factor of Group, and a within-subjects factor of Response Type (Hit vs. Correct Rejection). The effect of Group was not significant [$F(3,97) < 1$]. There was a robust main effect of Response Type on confidence [$F(1,97) = 40.77$, $MSe = 0.040$, $p < .0001$, $\eta^2 = .30$], which indicated that hits were rendered with more confidence than correct rejections [$M_H = 2.44$, $SE_H = 0.032$; $M_{CR} = 2.26$, $SE_{CR} = 0.041$]. The Group × Response Type interaction was not significant [$F(3,97) < 1$].

Confidence during Tests 2 and 3 was analyzed using two separate 2×2×2 mixed model ANOVAs examining a between-subjects factor of Volition (Use vs. Ignore), and within-subjects factors of Response Type (Hit vs. Correct Rejection) and Test (Test 2 vs. Test 3). Liberal and Conservative groups were analyzed separately in order to simplify the resulting analyses. Looking first to the Liberal groups, the only main effect was Response Type [$F(1,49) = 36.71$, $MSe = 0.076$, $p < .0001$, $\eta^2 = .43$] indicating that hits were more confident than correct rejections [$M_H = 2.52$, $SE_H = 0.046$; $M_{CR} = 2.28$, $SE_{CR} = 0.064$]. No other main effects or interactions were significant (all p's > .50).

Turning to the Conservative groups, there was a significant main effect of Volition [$F(1,48) = 5.28$, $MSe = 0.58$, $p < .05$, $\eta^2 = .10$], indicating that participants told to Use the feedback were more confident than participants told to Ignore the feedback [$M_U = 2.48$, $SE_U = 0.076$; $M_I = 2.23$, $SE_I = 0.076$]. There was also a main effect of Response Type [$F(1,48) = 4.99$

MSe = 0.11, $p < .05$, $\eta^2 = .094$] which indicated that hits were more confident than correct rejections [$M_H = 2.41$, $SE_H = 0.061$; $M_{CR} = 2.30$, $SE_{CR} = 0.056$]. Turning to the interactions, the only significant interaction was the Response Type × Test interaction [$F(1,48) = 4.05$, MSe = 0.020, $p < .05$, $\eta^2 = .078$]. This interaction indicated that the difference between hit and correct rejection confidence was larger for Test 2 [$M_H = 2.42$, $SE_H = 0.062$; $M_{CR} = 2.28$, $SE_{CR} = 0.058$; $p < .001$, Tukey's HSD] than for Test 3 [$M_H = 2.39$, $SE_H = 0.062$; $M_{CR} = 2.32$, $SE_{CR} = 0.059$; $p = .11$, Tukey's HSD]. No other main effects or interactions were significant (all p's > .31).

Confidence during the Subsequent Memory Test was analyzed using a 2×2×2 mixed model ANOVA examining between-subjects factors of Prior Feedback Group (Liberal vs. Conservative) and Prior Volition (Use vs. Ignore), and a single within-subjects factor of Response Type (Hit vs. Correct Rejection). The only significant effect was a main effect of Response Type [$F(1,97) = 64.75$, MSe = 0.083, $p < .001$, $\eta^2 = .40$], which indicated that hits were more confident than correct rejections. The effect of Prior Volition trended toward significance [$F(1,97) = 3.42$, MSe = 0.28, $p = .067$, $\eta^2 = .034$], which indicated that participants previously told to Use the feedback were still generally more confident than those previously told to Ignore the feedback [$M_U = 2.38$, $SE_U = 0.052$; $M_I = 2.24$, $SE_I = 0.053$]. No interactions approached significance (all p's > .16).

Subjective Awareness Questionnaire

This next section concerns the questions that participants answered following each test. These data are presented in Table 3.

Believed Bias

Group	Test 1	Test 2	Test 3	Subsequent Memory
Liberal/Use	55.30 (11.98)	68.68 (15.84)	65.15 (16.26)	54.64 (13.77)
Conservative/Use	52.00 (8.60)	48.88 (14.60)	46.44 (12.69)	51.88 (8.05)
Liberal/Ignore	51.83 (11.26)	60.60 (12.69)	61.32 (14.15)	51.92 (18.63)
Conservative/Ignore	50.11 (10.20)	45.50 (15.19)	37.29 (13.36)	51.48 (13.14)

Believed Accuracy

Group	Test 1	Test 2	Test 3	Subsequent Memory
Liberal/Use	66.80 (9.94)	76.58 (7.77)	74.23 (9.76)	46.31 (12.16)
Conservative/Use	69.64 (10.27)	74.52 (14.91)	63.84 (15.05)	54.60 (16.56)
Liberal/Ignore	64.12 (11.61)	66.40 (11.57)	64.20 (16.56)	44.12 (17.18)
Conservative/Ignore	66.29 (13.90)	68.20 (17.19)	66.52 (17.44)	55.60 (18.68)

Believed Feedback Manipulation

Group	Hits	CRs	FAs	Misses	None
Liberal/Use	3	3	5	1	14
Conservative/Use	5	5	2	3	11
Liberal/Ignore	4	4	4	3	11
Conservative/Ignore	5	1	1	7	11

Feedback Influence

Group	Positive|"Old"	Positive|"New"	Negative|"Old"	Negative|"New"
Liberal/Use	4.08 (1.26)	3.77 (1.36)	4.46 (1.14)	4.15 (1.40)
Conservative/Use	3.48 (1.45)	3.60 (1.55)	4.08 (1.50)	3.72 (1.54)
Liberal/Ignore	4.04 (1.02)	3.84 (1.14)	4.40 (1.04)	3.96 (1.24)
Conservative/Ignore	4.20 (1.08)	3.96 (1.34)	4.04 (1.14)	4.20 (1.22)

Table 3: Subjective Awareness Questionnaire data, Experiment 1.

Believed Bias

Participants indicated the proportion of test items they believed were old following each test, and this value ranged from 0-100. These were analyzed to determine if the consequences of FPF are apparent to the subjects when they are subsequently questioned. For this and future

sections, this 0-100 value will be referred to as the "believed bias" as it reflects the degree to which the subject believed that that prior test contained different proportions of old relative to new items. Occasionally participants would skip this question (likely by accidently pressing enter too quickly). Thus, degrees of freedom may not match the actual number of participants in the groups.

Believed bias was examined for the first test alone, then the second and third tests together, and the final subsequent memory test alone. This provides a baseline measure, a measure potentially sensitive to the FPF influences, and a measure reflecting whether believed bias persists even when feedback is removed. Believed bias for Test 1 was analyzed using a one-way ANOVA examining differences among the groups. This analysis was not significant [$F(3,81) < 1$]. However, examining how the mean believed bias differed from 50/50, participants were slightly liberal (i.e., thought there were more than 50% old items) in their assessment of the first test's makeup ($M = 52.45$, $SE = 1.15$; $p < .05$).

Next, believed bias during Tests 2 and 3 were analyzed using a 2×2×2 mixed model ANOVA examining between-subjects factors of Feedback Group (Liberal vs. Conservative) and Instructions (Use vs. Ignore), and a within-subjects factor of Test (Test 2 vs. Test 3). This analysis revealed a robust main effect of Feedback [$F(1,94) = 64.41$, $MSe = 292.1$, $p < .0001$, $\eta^2 = .41$], indicating that participants who received Liberal feedback thought there were more old items than did those who received Conservative feedback [$M_L = 64.14$, $SE_L = 1.71$; $M_C = 44.54$, $SE_C = 1.74$]. This mirrors the differences in actual criterion among the two feedback groups. There was also a significant main effect of Instructions [$F(1,94) = 6.71$, $MSe = 292.1$, $p < .05$, $\eta^2 = .067$], indicating that participants told to Use feedback thought there were more old items than did participants told to Ignore feedback [$M_U = 57.50$, $SE_U = 1.73$; $M_I = 51.18$, $SE_I = 1.73$].

Finally, the main effect of Test was marginally significant [$F(1,94) = 3.92$, $MSe = 123.5$, $p = .05$, $\eta^2 = .040$], indicating that participants thought there were more old items in Test 2 than in Test 3 [$M_{T2} = 55.91$, $SE_{T2} = 1.48$; $M_{T2} = 52.77$, $SE_{T3} = 1.43$]. No interactions among the factors were significant (all p's > .14). Turning to mean estimates of believed bias, both Liberal/Use and Liberal/Ignore participants overestimated the proportion of old items in Test 2 (p's < .001) and Test 3 (p's < .001; see Table 3). Among the Conservative groups, the only estimate differing from 50/50 was the Conservative/Ignore's estimate of Test 3 items ($M_{CIT3} = 37.29$, $SE_{CIT3} = 2.72$, $p < .01$).

Finally, believed bias from the Subsequent Memory Test was analyzed using a 2×2 factorial ANOVA, examining factors of Prior Feedback (Liberal vs. Conservative) and Prior Instructions (Use vs. Ignore feedback). This analysis did not reveal any main effects or interactions between the factors (all p's > .56). Turning to the mean estimates of bias, none of the groups differed from 50/50 (all p's > .10). Thus, although objective measures of response bias showed a significant degree of carryover during this final test, subjective measures of response bias did not.

Believed Accuracy

Participants also indicated the proportion of test items they believed they identified correctly following each test, and this value also ranged from 0-100. FPF appeared to affect participants' subjective beliefs about the relative preponderance of old and new items; did it also alter their subjective senses of their own performance? For this and future sections, this 0-100 value will be referred to as the "believed accuracy." As with believed bias, occasionally

participants would skip this question (likely by accidently pressing enter too quickly). Thus, degrees of freedom may not always match the actual number of participants in the groups.

As with believed bias I separately considered Test 1, Tests 2 and 3 together, and the Subsequent Memory test. There was no reliable difference among the four groups during Test 1 [$F(3,95)<1$]; thus all the groups began with a similar estimation of their own performance. These values were converted to proportions (i.e., 0 to 1) and were compared to actual accuracy (percent correct ranging from 0 to 1). Participants actually *underestimated* their own performance on average (whether believed accuracy – actual accuracy was different from 0; $M = -.10$, $SE = .013$, $p < .0001$).

Next, believed accuracy during Tests 2 and 3 were analyzed using a 2×2×2 mixed model ANOVA examining between-subjects factors of Feedback (Liberal vs. Conservative) and Instructions (Use vs. Ignore), and a within-subjects factor of Test (Test 2 vs. Test 3). The main effect of Feedback Group was not significant [$F(1,97) <1$]. The main effect of Instructions was significant [$F(1,97) = 5.31$, $MSe = 337.9$, $p < .05$, $\eta^2 = .051$], with participants told to Use the feedback believing they were more accurate than those told to Ignore it [$M_U = 72.29$, $SE_U = 1.82$; $M_I = 66.33$, $SE_I = 1.84$]. There was also a significant main effect of Test [$F(1,97) = 14.41$, $MSe = 62.6$, $p < .001$, $\eta^2 = .13$], which indicated that participants thought their accuracy declined across tests [$M_{T2} = 71.42$, $SE_{T2} = 1.32$; $M_{T3} = 67.20$, $SE_{T3} = 1.49$]. Among the interactions, the only significant two-way interaction was the Instructions × Test interaction [$F(1,97) = 4.22$, $MSe = 62.6$, $p < .05$, $\eta^2 = .041$]. While the group told to Use feedback thought they performed better than the group told to Ignore feedback, this difference diminished across tests (see Table 3), although neither group difference survived post hoc comparisons (both p's > .11, Tukey's HSD). Finally, there was a significant three-way interaction among the factors [$F(1,97) = 3.95$, $MSe =$

62.6, $p < .05$, $\eta^2 = .039$]. This interaction is shown in Supplementary Figure 1. As with Test 1, participants underestimated their own performance numerically, but in this case, not significantly ($M_{T2} = -.021$, $SE_{T2} = .014$, $t(100) = -1.52$, $p = .13$; $M_{T3} = -.025$, $SE_{T3} = .016$, $t(100) = -1.57$, $p = .12$)

Finally, believed accuracy from the Subsequent Memory Test was analyzed using a series of between-group t-tests among the two Prior Volition groups within each type of Prior Feedback received (i.e., Liberal/Use vs. Liberal/Ignore) due to the differences in test makeup between Feedback Groups. Prior Volition did not affect believed accuracy during the Subsequent Memory Test (both p's > .60). As with previous tests, participants underestimated their own performance during the Subsequent Memory test. This was true for both Prior Liberal ($M = -.22$, $SE = .021$, $t(50) = -10.43$, $p < .0001$) and Prior Conservative ($M = -.22$, $SE = .021$, $t(49) = -10.43$, $p < .0001$) participants.

Feedback Influence

Recall that participants ranked how they thought different feedback outcomes influenced their responding on a 6 point Likert scale. A first pass analysis examined whether there were any mean differences in the level of influence participants felt the feedback possessed. Influence ratings were subjected to two $2 \times 2 \times 2$ mixed model ANOVAs examining a between-subjects factor of Prior Instructions (Use vs. Ignore), and within-subjects factors of Response Type (Old vs. New) and Feedback Valence (Positive vs. Negative). Prior Liberal and Prior Conservative participants were analyzed separately in order to simplify the resulting analyses. Looking first at the Prior Liberal participants, there was a significant main effect of Feedback Valence [$F(1,49) = 4.15$, $MSe = 1.20$, $p < .05$, $\eta^2 = .078$] which indicated that participants felt negative outcomes

were more influential than positive feedback [M_{Pos} = 3.93, SE_{Pos} = 0.16; M_{Neg} = 4.24, SE_{Neg} = 0.15]. There was also a significant main effect of Response Type [$F(1,49) = 5.02$, $MSe = 0.49$, $p < .01$, $\eta^2 = .17$] which indicated that influence ratings were higher for hits than for correct rejections [M_H = 4.24, SE_H = 0.13; M_{CR} = 3.93, SE_{CR} = 0.15]. No other main effects or interactions approached significance (all p's > .54). Turning to the Prior Conservative participants, the only significant effect was the three-way interaction amongst the factors [$F(1,48) = 6.26$, $MSe = 0.39$, $p < .05$, $\eta^2 = .12$]. This interaction indicated several things. First, those told to Ignore feedback generally ranked it as more influential than those told to Use the feedback, irrespective of the Response Type or Feedback Valence. Second, those told to Use the feedback generally felt negative feedback was more influential, particularly following false alarms. No other main effects or interactions from this analysis approached significance (all p's > .21).

Effects of Awareness

It is important to consider how these results might be affected by awareness of the nature of the feedback. That is, if a participant could pick out what was wrong with the feedback, might he or she show a different pattern of behavior than one who was ignorant about the feedback's nature? At the end of the experiment, participants made a forced-choice decision about how the computer was manipulating the feedback. The distribution of these responses is displayed in Table 3. A total of 19 participants selected the response option reflecting the actual manipulation in place for their group (see Table 3). One can approximate the 95% confidence intervals around these values using the normal approximation to determine if they differ significantly from chance (in this case, 20%). Neither proportion was different than one would expect from chance within

either Feedback Group (Liberal: .196 ± .109; Conservative: .173 ± .103) or either Volition Group (Use: .154 ± .098; Ignore: .216 ± .113). Additionally, a large proportion of participants (significantly above chance) in each group selected the "I don't agree with any of the above options" response (Liberal: .431 ± .136; Conservative: .481 ±.136), suggesting that an overwhelming majority of participants were unaware of the specific nature of the FPF manipulation. These data suggest that the manipulation is not very salient with the vast majority of the subjects failing to identify it on a forced choice question. Additionally, the instruction to ignore the feedback did not increase awareness of the manipulation as demonstrated by Fisher's exact test ($p > .45$, two-sided).

The analysis of criterion on Tests 2 and 3 was rerun, with participants who correctly identified the manipulation removed. None of the main effects or interactions were affected by this exclusion. This procedure was repeated, this time only examining the participants who chose "I don't agree with any of the above options." In this case, the Volition × Feedback Group × Test interaction in the omnibus ANOVA was significant [$F(1,43) = 5.20$, $MSe = 0.027$, $p < .05$, $\eta^2 = .11$]; in this case, this interaction indicated that participants told to Ignore the feedback were generally MORE influenced by it (as evidenced by slightly larger changes in criterion across Tests for those told to Ignore feedback). Regardless, the interpretation of the effect remained as above (i.e., participants appear unable to ignore the feedback as evidenced by Feedback Group differences in criterion). A similar analysis of just the "aware" participants could not be conducted due to the small sample sizes.

Personality Measures

The BIS/BAS, RFQ, and GRAPES questionnaires did not correlate with any of the measures described above.

Discussion

This experiment replicated and extended previous work using the FPF paradigm. Critically, the main hypothesis of this experiment was confirmed: observers are unable to inhibit the effects of FPF on their decision criterion. These results lend further support to the notion that implicit learning mechanisms may influence recognition memory criteria.

This experiment also replicated the stability of FPF-induced criterion shifts in the absence of further reinforcement (Han & Dobbins, 2008, 2009). Specifically, group differences in criterion persisted during the Subsequent Memory Test even though participants were no longer receiving any feedback. Interestingly, subjective senses of bias (i.e., the Believed Bias) were not different among the groups despite differences in criterion during this final test. Even participants who correctly identified their feedback manipulation showed this pattern (although this analysis does suffer from low statistical power). This harkens back to a main tenant of implicit versus explicit learning - namely, that learner should show little to no explicit knowledge that learning has occurred (Reber, 1989). In this case, even though FPF affected the Believed Bias estimates during the two FPF tests, participants likely ascribed the lopsided feedback they received to the construction of the test list rather than anything abnormal about the feedback procedure itself. Thus awareness of the actual feedback-based learning in this experiment was fairly minimal.

These results are consistent with other results demonstrating the limited influence of explicit control over implicit learning tasks (Reber, 1966, 1976). For instance, certain

categorization tasks require integration of two orthogonal dimensions to categorize stimuli, and learning to categorize in this manner presumably relies on procedural learning (information integration tasks; Ashby & Maddox, 2005). In these tasks, simply informing observers of the number of categories and encouraging them to base their categorization decisions on all the stimulus dimensions does not result in learning the categorization rule when it requires procedural learning; only when feedback is provided can participants learn to accurately categorize these sorts of stimuli (Ashby, Queller, & Berretty, 1999). Similarly, Reber (1976) encouraged participants to determine the rules underlying stimulus generation in an artificial grammar task. Those participants actually did WORSE when encouraged to look for structure. Essentially, participants were looking for rules and structure that were incorrect instead of letting feedback guide their learning. In this case, explicit strategies actually interfered with implicit learning. Indeed, when the underlying structure is relatively straightforward or when participants are given more useful strategies to follow, explicit control can aid in implicit learning (Reber, Kassin, Lewis, & Cantor, 1980).

There was modest evidence indicating that surprising feedback outcomes affected later item memory. Specifically, false positive feedback, compared to true negative feedback during Test 2 produced small but reliable benefits in later item memory during the Subsequent Memory Test. This pattern of performance is consistent with the hypothesis that unexpected positive outcomes lead to better memory. These results are also in line with prior research concerning positive mnemonic benefits associated with reward-related dopaminergic activity (Adcock et. al., 2006). Future work involving functional neuroimaging should be conducted to investigate whether regions involved with processing FPF overlap with striatal regions implicated by Adcock and colleagues (2006). However it should be noted that this experiment did not include a

true baseline with which to compare performance for prior error trials. In other words, since all trials received feedback of some kind, it is admittedly a bit odd to say that surprising feedback outcomes led to a mnemonic benefit since it is not entirely clear where baseline lies. It is entirely possible that errors that receive no feedback would be remembered as well as errors that received positive feedback, especially if "no feedback" was considered a surprising outcome. Further, estimates of memorability were noisy due to the small number of error trials that ended up being sampled for the Subsequent Memory test. Future studies should be designed to replicate this effect and include a more adequate set of baseline error trials.

One novel aspect of this experiment lies in the instructions to the participants to ignore the feedback. It seems counterintuitive to examine how observers can inhibit feedback as feedback-based learning paradigms focus exclusively on how this learning proceeds. It is not clear that any research has been done in examining how individuals might inhibit feedback during an implicit learning task. Lampinen and colleagues (2007) did examine how instructions to ignore feedback influenced performance during an eyewitness identification task. Feedback following eyewitness identification tends to influence subjective ratings about the identification without influencing the accuracy of the identification. For example, positive feedback increases subjective confidence in the veracity of the identification and the quality of the witnessing conditions (e.g., how well-lit the suspect was), while negative feedback decreases these subjective feelings. Lampinen and colleagues (2007) demonstrated mixed results when asking participants to ignore prior feedback outcomes during an eyewitness identification task. When warned they had received random feedback, participants were able to discount its effects on their metacognitive ratings. However, when simply asked to disregard the feedback they had seen previously (and not told of its erroneous nature), feedback valence still influenced metacognitive

ratings. Thus, in an explicit memory paradigm, observers can have some success inhibiting the effects of feedback on their own responding. Other than the current study, however, the literature offers no insight into how observers can successfully ignore feedback designed to implicitly influence their behavior. Presumably animals cannot be asked to "ignore" potential learning cues, so this sort of paradigm is not feasible for a large portion of the learning literature. In humans, it seems reasonable to assert that feedback learning that proceeds despite the observer's intentions is in some critical sense "implicit."

It was noted that there were significant differences in accuracy as measured by d' amongst the feedback groups. This is likely due to using the equal variance calculation of accuracy, which is technically incorrect. Under most situations, equal and unequal variance calculations of accuracy converge on the same answer. However, when criteria become extreme, equal variance estimates of accuracy are no longer independent of estimates of bias (Macmillan & Creelman, 2005). In other words, observers who become extremely liberal show a lower d' than those who become extremely conservative. In these cases, d_a could be calculated as it more accurately reflects the unequal variance of the two distributions. Reevaluating accuracy this way shows no difference amongst the feedback groups [Test 2: $t(99) = -1.60$, $p = .11$; Test 3: $t(95) = -0.89$, $p = .38$]. Regardless, such nuances in how accuracy is calculated do not affect the interpretation of the criterion data – that is, that FPF shifts criteria, and instructions to ignore it do not seem altogether effective.

Despite the clear influence of the feedback in this experiment, participants did not rate the feedback as remarkably influential to their responding as evidenced by the feedback influence questions of the Subjective Awareness Questionnaire. This is in contrast with the believed bias questions which demonstrated clear differences between the Liberal and

Conservative feedback groups. If the feedback manipulation is supposed to be relatively opaque to observers, then how does one explain this discrepancy? As mentioned above, it is likely the estimates of test construction reflect a response to receiving largely positive feedback for one type of response. That is, participants may recognize that they are saying “old” or “new” fairly often but believe that this is due to test construction rather than aberrant feedback. In fact, the biased feedback may be confirming any suspicions they may have about an unbalanced test makeup.

Chapter 6

Experiment 2

3) Would criterion learning on one stimulus type generalize to other types of stimuli?

Across some incremental reinforcement learning paradigms, learning tends to be very feature specific – in many cases, learning occurs at the level of the individual stimulus or cue (Ashby & Waldron, 1999; Seger, 2008; Poldrack & Packard, 2003) whereas other forms of learning are independent of the actual stimuli (Willingham, 1999). In typical incremental reinforcement learning paradigms, participants view the same stimuli multiple times across an experiment. The type of learning that develops depends on the type of task. In information integration tasks, participants tend to build associations between particular regions of perceptual space and particular responses (Ashby & Maddox, 2005), whereas learning in motor sequence and artificial grammar tasks is decidedly independent from the stimuli (in that surface features of the stimuli can change grossly without disrupting learning; Reber, 1969; Willingham, 1999).

The typical incremental reinforcement procedure lies in contrast with typical memory experiments, as specific stimuli are rarely repeated during a single test in standard memory experiments. Indeed during the FPF manipulation in Experiment 1, biases are effectively induced during test lists in which each item is shown exactly once. Thus this effect appears to rely on reinforcement learning, but which does not require any reinforced stimulus to actually repeat (Han & Dobbins, 2008, 2009). In other words, given the setup of the FPF paradigm, it is impossible for observers to learn a particular stimulus-response association (e.g., "APPLE - respond old"). Rather, participants are likely learning the mapping between abstracted levels of generalized memory evidence and recognition judgments, and generalizing this across perceptually distinct stimuli with the same levels of memory evidence. Thus, two stimuli with

vastly different perceptual features (such as words and faces) may show similar response patterns due to similar levels of abstracted memory evidence. In other words, provided both types of stimuli evoke signals of general memory evidence, one should show transfer of this learned mapping across fundamentally different stimulus classes (Figure 7).

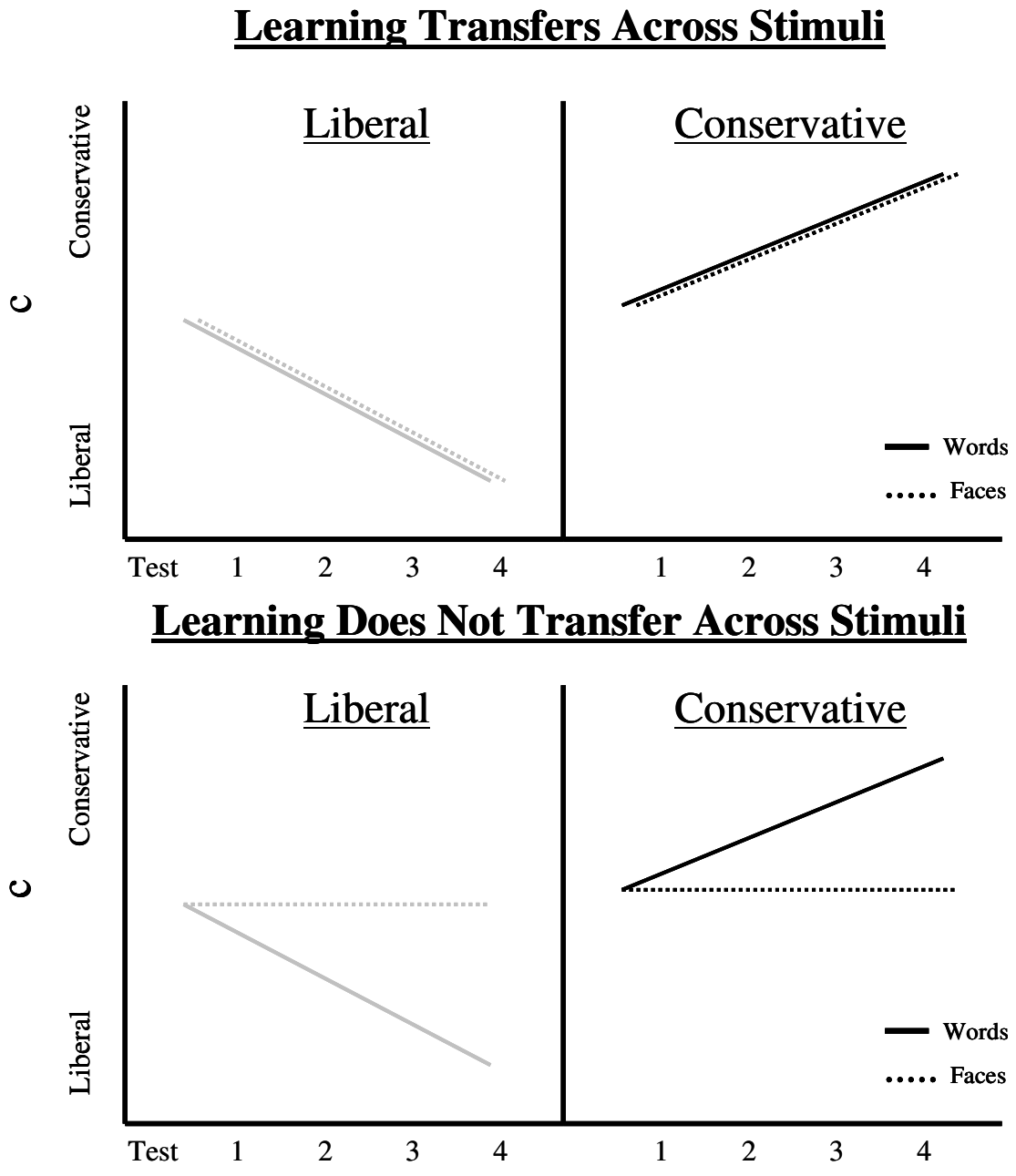


Figure 7: Hypothetical results. The top panel (predicted results) shows hypothetical results if the FPF effect transfers to stimuli that do not receive reinforcement (faces); criterion should not differ for words and faces. The bottom panel shows hypothetical results if the FPF effect does not transfer to stimuli that never receive reinforcement; criterion for faces should not change appreciably from baseline.

Participants

56 (28 per group) individuals participated in Experiment 2 in return for payment ($15).

Face Stimuli

Face stimuli were drawn from a database obtained from Endl and colleagues (1998). Stimuli consisted of black and white photographs of young Caucasian adults without distinctive facial features. The black and white photographs were carefully edited to maintain a standard brightness and contrast (Endl et. al., 1998). Examples of face stimuli are shown in Figure 8. These cropped faces were chosen in order to minimize the contribution of feature recollection to recognition.

Figure 8: Examples of face stimuli for Experiment 2

Procedure

Pilot studies were conducted in order to match accuracy (d') across words and faces. As discussed in the results section of the previous chapter, when d' differs, differences in c become difficult to interpret, as c is calculated relative to the intersection of the target and lure distributions (see Macmillan & Creelman, 2004, Verde & Rotello, 2007). It was important for

this particular experiment to match accuracy across stimulus types to interpret any potential differences in c between words and faces within a given test. For example, consider the scenario in Figure 9. In this case, an observer may require just as much evidence to identify either a word or face as old, but the calculation for c would show a more conservative criterion for faces (roughly zero) as compared to words (less than zero). Thus, pilot studies were conducted to titrate face contrast and encoding conditions to match word and face accuracy. At first word performance (d') far exceeded face performance, even with no facial blur. Thus the encoding tasks were designed to boost face encoding and impoverish word encoding.

The experiment consisted of four study/test cycles. Participants first studied 80 serially presented words and then studied 20 serially presented faces. For words, participants were asked to determine if the first and last letter of each word were in alphabetical order. For faces, participants were asked to rate how comfortable the person appeared with being photographed (on a 4 point scale ranging from "Very Uncomfortable" to "Very Comfortable"). Study was self-paced. These encoding tasks proved to match recognition performance for both stimulus types.

Following study, 160 (80 studied and 80 novel) word and 40 (20 studied and 20 novel) face stimuli were randomly intermixed during test. For each item type, participants first indicated whether the item was old or new, then indicated confidence in their old/new decision ("low," "medium," or "high" confidence). Participants were split into two groups during test phases: one group received conservative FPF, and the other received liberal FPF; FPF was provided for both groups for all four tests. Participants only received feedback following word stimuli (participants were simply informed that they would not receive feedback on every trial). This was designed to determine whether a bias instilled for words would transfer to faces that never received reinforcement. Participants answered the first two questions of the Subjective Awareness

Questionnaire following the first three tests, although they were modified slightly: instead of judging the proportion of old and new items, participants separately judged the proportions of old words and old faces and separately judged their performance for words and faces. Following the final test, participants completed the full (again, slightly modified) Subjective Awareness Questionnaire, BIS/BAS, GRAPES, and the RFQ before completing the experiment.

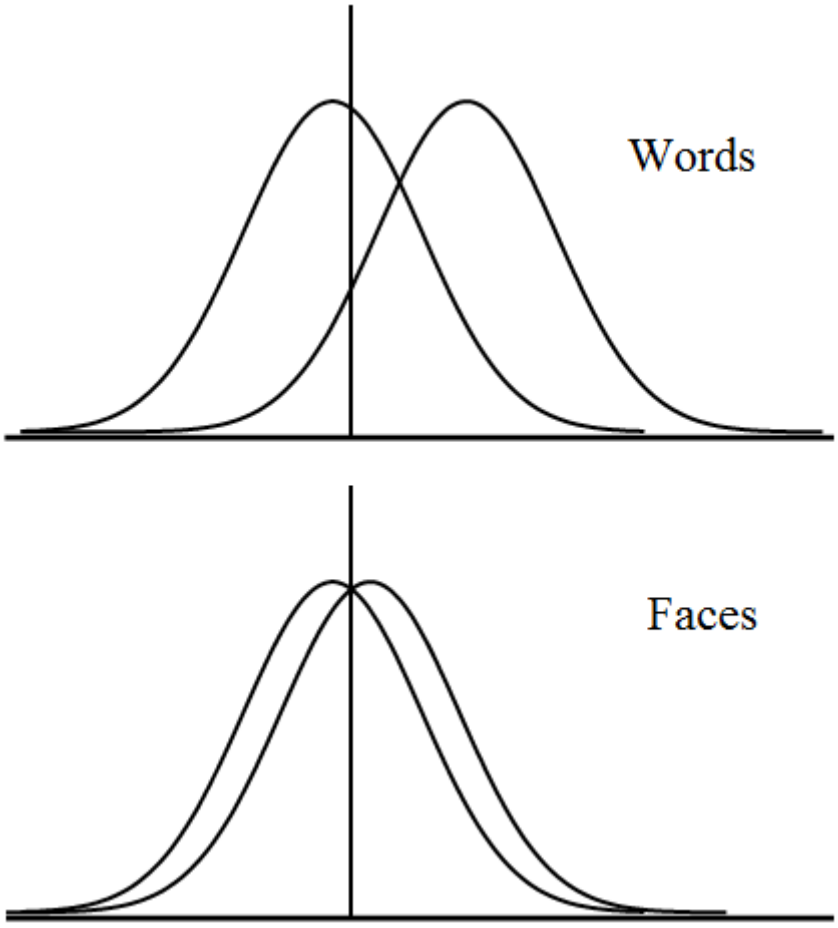


Figure 9: Differences in c become difficult to interpret when d' differs. What appears to be the same criterion on the memory evidence axis would result in a different value of c for the two stimulus types.

Results

Hit rates, FA rates, d', and c for words and faces are presented in Table 4.

Liberal

Stimulus	Test	Hit Rate	FA Rate	d'	c	FPF trials
Words	Test 1	.68 (.12)	.33 (.16)	1.01 (0.43)	-0.0087 (0.39)	17.77 (8.96)
	Test 2	.74 (.12)	.46 (.16)	0.84 (0.61)	-0.29 (0.41)	26.23 (10.02)
	Test 3	.78 (.13)	.52 (.18)	0.81 (0.54)	-0.46 (0.43)	28.15 (10.79)
	Test 4	.77 (.14)	.57 (.20)	0.64 (0.57)	-0.50 (0.47)	31.15 (10.57)
Faces	Test 1	.64 (.16)	.29 (.13)	0.94 (0.53)	0.10 (0.30)	N/A
	Test 2	.71 (.16)	.35 (.16)	1.07 (0.71)	-0.10 (0.35)	N/A
	Test 3	.68 (.21)	.36 (.15)	0.94 (0.56)	-0.065 (0.50)	N/A
	Test 4	.69 (.21)	.42 (.21)	0.73 (0.64)	-0.17 (0.52)	N/A

Conservative

Stimulus	Test	Hit Rate	FA Rate	d'	c	FPF trials
Words	Test 1	.55 (.15)	.22 (.10)	0.94 (0.36)	0.35 (0.34)	25.27 (7.74)
	Test 2	.51 (.20)	.23 (.16)	0.86 (0.35)	0.41 (0.53)	27.12 (11.74)
	Test 3	.48 (.24)	.22 (.15)	0.81 (0.43)	0.50 (0.61)	29.50 (13.32)
	Test 4	.44 (.26)	.19 (.17)	0.81 (0.46)	0.64 (0.71)	31.96 (15.60)
Faces	Test 1	.59 (.12)	.25 (.16)	1.00 (0.59)	0.25 (0.32)	N/A
	Test 2	.56 (.21)	.29 (.16)	0.79 (0.69)	0.23 (0.45)	N/A
	Test 3	.53 (.23)	.26 (.15)	0.74 (0.57)	0.34 (0.47)	N/A
	Test 4	.52 (.24)	.29 (.15)	0.69 (0.57)	0.28 (0.53)	N/A

Table 4: Hit rates, false alarm rates, accuracy, criterion, and number of FPF trials for both words and faces during Tests 1, 2, 3, and 4 (Experiment 2). Standard deviations in parentheses.

Accuracy

Accuracy (d') was analyzed using a 2 × 2 × 4 mixed ANOVA examining a between-subjects factor of Feedback (Liberal vs. Conservative) and within-subjects factors of Stimulus (Word vs. Face) and Test (Test 1 vs. Test 2 vs. Test 3 vs. Test 4). The only significant effect on accuracy came from Test [$F(3,135) = 5.52$, $p < .01$, $\eta^2 = .11$], indicating that accuracy declined across tests for both groups. No other main effects or interactions were significant (all p's > .33).

Importantly, because accuracy did not differ between the stimuli, comparisons of c for both stimuli and both groups within a given test can be made without issue.

Criterion

The purpose of this experiment was to determine whether a criterion reinforced for one set of stimuli (words) would generalize to a starkly different class of stimuli (faces) that were not reinforced, but which presumably evoked a similar range of memory evidence. Feedback group differences in criterion for faces would support this hypothesis. It should be noted that 15 participants (7 Liberal, 8 Conservative) had at least one instance of chance or slightly below chance accuracy (for either words or faces) for a given test. In a control analysis, any participant with any instances of chance performance were initially excluded. Excluding these participants did not change the conclusions of the analyses presented below. Thus the primary results are presented with all participants included.

To investigate the main hypothesis, criterion was analyzed using a $2 \times 2 \times 4$ mixed ANOVA examining a between-subjects factor of Feedback Group (Liberal vs. Conservative) and within-subjects factors of Stimulus (Word vs. Face) and Test (Tests 1-4). This analysis revealed a main effect of Feedback Group [$F(1,45) = 32.19$, $MSe = 0.95$, $p < .0001$, $\eta^2 = .42$], with the Liberal group being more liberal than the Conservative group. The main effect of Stimulus was not significant [$F(1,45) < 1$], which indicated that bias did not differ between words and faces (although this is collapsed across Feedback Group). There was a marginal effect of Test [$F(3,135) = 2.67$, $p = .05$, $\eta^2 = .056$] on criterion, indicating a slight liberal trend across tests. There was a significant Feedback × Stimulus interaction [$F(1,45) = 14.01$, $MSe = 0.10$ $p < .001$, $\eta^2 = .24$]. This interaction indicated that the difference in criterion between Feedback Groups

was more extreme for words than for faces; post-hoc tests confirmed this for both groups [L_W = -0.31 < L_F = -.053, $p < .05$; C_W = 0.49 > C_F = 0.28, $p < .001$]. There was a significant interaction between Feedback Group and Test [$F(3,135) = 14.60$, $MSe = 0.10$, $p < .001$, $\eta^2 = .24$], indicating that criterion differences diverged across tests. The Stimulus × Test interaction was not significant [$F(3,135) = 1.24$, $MSe = 0.063$, $p = .30$, $\eta^2 = .027$]. Finally, there was a significant three-way interaction among the factors [$F(3,135) = 4.41$, $MSe = 0.063$, $p < .01$, $\eta^2 = .089$], graphed in Figure 10. This interaction indicates that criterion for words diverged further and more quickly than did criterion for faces. This is unsurprising considering responses to faces were never directly reinforced. Importantly however, there were group differences in criterion for faces; that is, criterion for faces was more liberal for participants given Liberal feedback than for participants given Conservative feedback, and this effect increased across testing, as demonstrated by a significant Feedback Group × Test interaction when looking only at the face stimuli [$F(3,135) = 2.74$, $p = .046$, $\eta^2 = .057$] (dashed lines Figure 10); in summary, response bias toward faces mirrored bias toward words, despite the former never receiving direct reinforcement during the experiment.[1]

[1] It should be noted that since d' differed across tests, comparisons of c across tests should be interpreted with caution, as c represents the midpoint of the distributions, which is changing locations. Critically however, both feedback groups shift in the predicted direction, lending support to the interpretations described here.

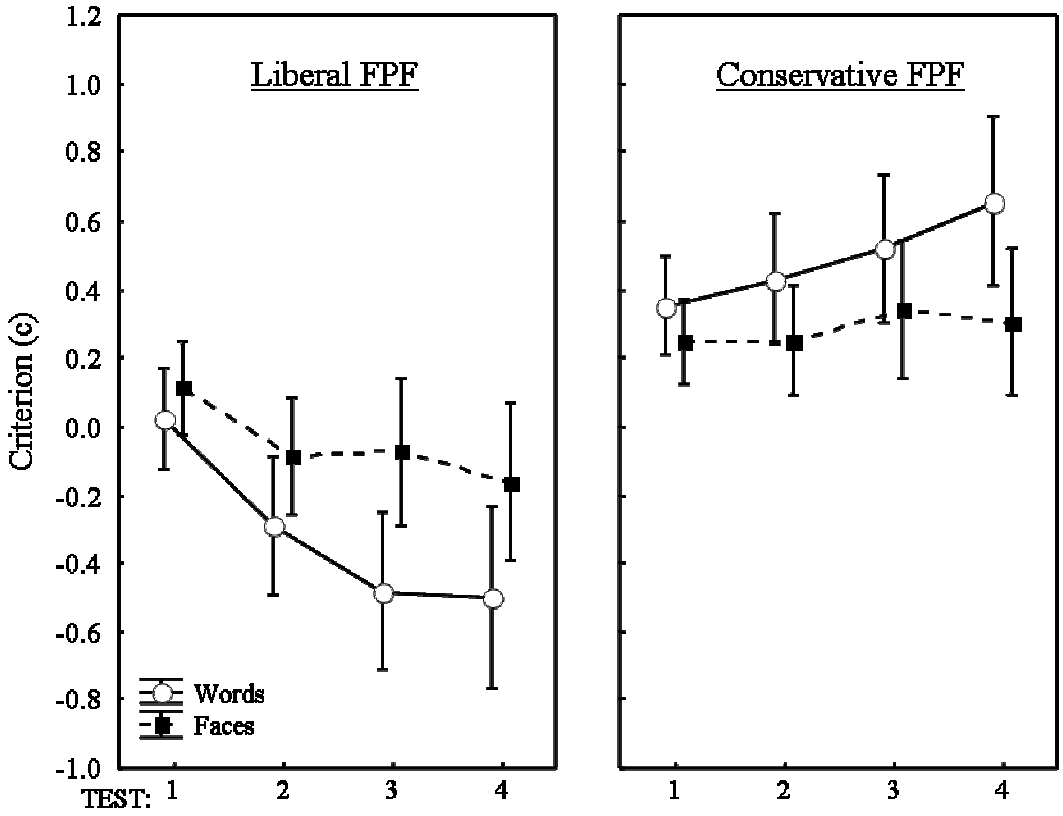


Figure 10: Criterion diverges across tests based on feedback delivered (Experiment 2). Importantly, criterion appears to generalize between words and faces (vertical bars represent 95% confidence intervals).

Confidence

Confidence for correct reports was examined using a 2 × 2 × 2 × 4 mixed ANOVA examining between-subjects factors of Feedback Group (Liberal vs. Conservative) and within-subjects factors of Stimulus (Word vs. Face), Response (Hit vs. CR), and Test (Test 1 vs. Test 2 vs. Test 3 vs. Test 4). The only significant effect from this analysis was a significant effect of Stimulus [$F(1,50) = 9.69$, $p < .01$, $\eta^2 = .16$], indicating that confidence was generally higher for faces than for words. No other main effects or interactions were significant (all p's > .11).

Subjective Awareness Questionnaire

Due to an error in the experimental script, Feedback Influence and Awareness ratings were missing for 15 participants (8 Liberal, 7 Conservative). Those data are reported for the remaining 37 participants and all data from the Subjective Awareness Questionnaire are presented in Table 5.

Believed Bias

Group	Stimulus	Test 1	Test 2	Test 3	Test 4
Liberal	Word	59.23 (12.86)	63.58 (13.78)	66.19 (13.79)	66.76 (15.82)
	Face	51.60 (19.67)	23.88 (18.47)	54.15 (17.29)	55.38 (15.74)
Conservative	Word	42.88 (15.28)	37.71 (12.24)	39.84 (15.04)	39.80 (18.40)
	Face	45.73 (16.11)	51.32 (15.98)	40.12 (17.94)	45.31 (18.23)

Believed Accuracy

Group	Stimulus	Test 1	Test 2	Test 3	Test 4
Liberal	Word	66.88 (15.76)	66.85 (18.25)	64.88 (18.89)	61.38 (17.47)
	Face	55.12 (23.24)	47.62 (19.44)	51.62 (19.98)	49.96 (20.28)
Conservative	Word	75.19 (11.44)	71.04 (11.39)	65.77 (15.34)	70.31 (15.40)
	Face	52.77 (13.63)	50.15 (14.70)	47.12 (15.97)	48.35 (15.46)

Believed Feedback Manipulation

Group	Hits	CRs	FAs	Misses	None
Liberal	5	0	2	0	11
Conservative	2	5	1	5	6

Feedback Influence

Group	Positive|"Old"	Positive|"New"	Negative|"Old"	Negative|"New"
Liberal	4.17 (1.25)	3.72 (1.32)	4.56 (1.15)	4.22 (1.22)
Conservative	4.10 (1.32)	4.32 (1.33)	4.05 (1.39)	4.05 (1.35)

Table 5: Subjective Awareness Questionnaire data, Experiment 2.

Believed Bias

Participants reported their believed proportion of old item judgments separately for both words and faces for each test and these values are listed in Table 5. As with Experiment 1, this measure was used to examine whether their subjective beliefs tracked the influence of the feedback manipulation. These values were subjected to a 2 × 2 × 4 mixed ANOVA examining a between-subjects factor of Feedback Group (Liberal vs. Conservative) and within-subjects factors of Stimulus (Word vs. Face) and Test (Tests 1 - 4). There was a main effect of Feedback Group [$F(1,46) = 42.09$, $p < .0001$, $\eta^2 = .48$] indicating that participants who received Liberal FPF reported subjectively more old items than did participants who received Conservative FPF [$SO_L = 58.80$, $SO_C = 42.56$]. There was no effect of Stimulus [$F(1,46) < 1$] nor of Test [$F(3,138) < 1$] on believed bias. There was an interaction between Feedback Group and Stimulus on believed bias [$F(1,46) = 13.22$, $p < .001$, $\eta^2 = .22$]. This interaction indicated that, like the actual criterion, bias reports were more extreme in each group for words than for faces, although this difference only survived post-hoc comparisons in the Liberal feedback group [$L_W = 63.81$, $L_F = 53.78$, $p < .05$; $C_W = 39.46$, $C_F = 45.67$, $p = .22$]. Finally, there was a significant interaction between Feedback Group and Test on subjective proportion of old items [$F(3,138) = 2.75$, $p < .05$, $\eta^2 = .056$]. Like the criterion results above, this interaction indicated that differences in believed bias for the two Feedback groups increased across subsequent tests.

Believed Accuracy

As with believed bias for this experiment, participants rated their believed accuracy separately for words and faces. These ratings were subjected to a similar 2 × 2 × 4 mixed ANOVA as above, again examining a between-subjects factor of Feedback Group (Liberal vs.

Conservative) and within-subjects factors of Stimulus (Word vs. Face) and Test (Tests 1-4). There was a robust effect of Stimulus on subjective performance estimates [$F(1,49) = 91.05$, $p < .0001$, $\eta^2 = .65$] indicating that participants thought they identified far more words correctly than they did faces [$M_W = 68.01$, $SE_W = 1.88$; $M_F = 50.46$, $SE_F = 2.23$] even though actual performance was equivalent for the two stimulus types. No other main effects or interactions were significant (all p's > .08). Next, believed accuracy was contrasted with actual performance (percent correct) for both words and faces, collapsed across Feedback Group in order to improve power. In contrast with Experiment 1, participants *overestimated* their recognition performance for words across all tests, but only significantly so in Test 2 [Test 1 diff: 3.92, $t(51) = 1.97$, $p = .054$; Test 2 diff: 5.10, $t(51) = 2.61$, $p < .05$; Test 3 diff: 2.29, $t(51) = 1.02$, $p = .31$; Test 4 diff: 4.52, $t(51) = 1.98$, $p = .054$]. Participants significantly *underestimated* their own recognition performance for faces across all tests [Test 1 diff: -13.82; Test 2 diff: -16.93; Test 3 diff: -15.35; Test 4 diff: -13.39].

Feedback Influence

Participants ranked how they thought different feedback outcomes influenced their responding on a 6 point Likert scale ranging from "not at all influenced" to "very influenced." Separate influence ratings were made for all possible feedback outcomes. A first pass analysis examined whether there were mean differences in the subjective influence of the feedback among the feedback groups. Mean influence ratings were analyzed using a $2 \times 2 \times 2$ mixed ANOVA examining a between-subjects factor of Feedback Group (Liberal vs. Conservative) and within-subjects factors of Response Type (Old vs. New) and Feedback Valence (Positive vs. Negative). There were no main effects or interactions among the factors (all p's > .11).

Effects of Awareness

As with Experiment 1, participants were loosely classified as "aware" if they correctly indicated the nature of their feedback manipulation when given five different options. By this criterion, 7 participants were deemed aware (Liberal = 2, Conservative = 5). The proportion of aware subjects did not differ from chance within each Feedback Group (Liberal: .111 ± .095; Conservative: .263 ± .121). Additionally, a large proportion of participants in both groups (Liberal: .611 ± .132; Conservative .316 ± .128) selected "I don't agree with any of the [feedback] options" (Table 5), suggesting many participants were unfazed by the manipulation. Fisher's exact test indicated no relationship between group membership and awareness of the feedback manipulation ($p > .40$, two-sided). As with Experiment 1, none of the main effects or interactions were affected when aware participants were excluded from the analyses.

Personality Measures

The BIS/BAS, RFQ, and GRAPES questionnaires did not correlate with any of the measures described above.

Discussion

The purpose of this experiment was to test the hypothesis that a criterion learned for one class of stimuli would transfer to another class of stimuli with starkly different surface features. More specifically, FPF was provided for words but no form of feedback was ever provided for faces. Both words and faces demonstrated more liberal responding for the Liberal than the Conservative feedback group overall. Further, within each stimulus class there was a Feedback

Group × Test interaction, demonstrating that criterion differences increased across repeated testing and exposure to biasing feedback.

As noted, response bias for faces was not as extreme as response bias for words. That is, participants who received Liberal feedback were more liberal for words than faces; participants who received Conservative feedback were more conservative for words than for faces. This is likely due to faces never being directly reinforced. Both stimulus classes presumably overlap to some degree on some underlying memory evidence variable. To the degree that these stimulus classes share some abstracted memory evidence, the stimulus that is never directly reinforced should show some effect of feedback. If the reinforcement schedule were flipped such that faces received feedback and words did not, one would expect the opposite pattern: a more extreme criterion for faces instead of words.

These results are consistent with some implicit learning tasks showing transfer across seemingly disparate stimuli. For example, Reber (1968) demonstrated intact learning of an artificial grammar when symbols were changed but the underlying syntactic rules remained intact. In his task, participants were first trained on an artificial grammar via memorizing several example strings. Following study, groups of participants were tested with strings from a grammar that either: used the same symbols (letters) and syntax as the training grammar; used the same symbols but different underlying syntax; used different symbols but the same underlying syntax as the training grammar; or used different symbols and underlying syntax. Reber (1969) found that groups who encountered the same underlying syntax at test correctly identified grammatical strings equally often, even when the symbols representing the grammar had changed. In contrast, participants who encountered a new syntactic structure performed worse, regardless of the identity of the symbols. In this case, implicit knowledge was not bound

to the surface features of the item; rather, observers learned a deeper representation of the grammar (see also Mathews et. al., 1989; Posner & Keele, 1968).

Why would a criterion for one set of materials transfer to another set of materials with starkly different surface features? One explanation postulates that, at some level, both stimulus types evoke a domain-general memory strength signal, akin to perceptual intensity. Basic Signal Detection models of recognition assume this – that is, observers make recognition decisions based on signals along a "memory evidence" or "familiarity" axis (Macmillan & Creelman, 2005; Yonelinas, 2002). The results of this experiment support the idea that, despite vastly different surface features, these two stimulus types are broken down into a raw memory evidence signal at some level. It may not be much of a stretch to claim that ALL types of stimuli evoke some basic memory evidence signal to be evaluated. Under this framework, FPF alters the mapping between this underlying evidence and corresponding judgments; hence the transfer across stimuli from different domains. The degree of overlap between two types of stimuli with regards to how much general memory evidence they evoke may relate to the degree of transfer. In other words, stimuli that share more features with words (e.g., nonwords) should show better transfer, as they would likely overlap more considerably in their levels of generalized memory evidence. This remains an intriguing area for future study.

It is clear that implicit criterion learning acts on some general memory evidence signal common across different types of stimuli. How might such a signal be characterized? Hippocampal and medial temporal lobe (MTL) models of recognition memory postulate that recognitions occur via pattern matching in the hippocampus and the surrounding cortex (Norman & O'Reilly, 2003). In Norman and O'Reilly's (2003) model, MTL outputs a scalar value based on the degree of overlap between a given stimulus and the contents in memory. This value could

easily read into a system that evaluates it against a criterion value and issues an old/new response. The difference between this scalar and the criterion could serve as a proxy for the prediction error; larger values should result in larger adjustments of behavior following unexpected outcomes (positive or negative). The system aims for a prediction error of zero. In the presence of veridical feedback, observers should hold generally neutral criteria (Benjamin, Diaz, & Wee, 2009; Kantner & Lindsay, 2010). In the case of biased feedback, however, positive prediction errors bias the observer toward one response over another. Finally, in the absence of continued biased feedback, observers would generally hold the same criteria they had previously learned, as demonstrated in Experiment 1 and prior work (Han & Dobbins, 2008, 2009).

However, it is relevant to speculate why this transfer is not perfect. One potential explanation is that familiarity is not simply a unitary process, but an amalgamation of multiple sources of information (Rugg & Curran, 2007). In other words, only one aspect of the familiarity process represents the actual contribution of memory evidence. The memory decision could also be influenced by factors such as conceptual or perceptual fluency. Words and faces could differ in levels of fluency; not surprising since they were encoded differently. If these sorts of extraneous sources of information are independent of the abstracted memory strength, then one might expect criterion transfer across stimulus types to be limited.

Chapter 7

4) Would a shift of context eliminate an implicitly acquired response bias?

Experiment 2 demonstrated that FPF learning generalizes across stimuli with grossly different perceptual features. This suggests a fairly abstract form of learning in which the mapping between general memory evidence and overt recognition judgments is altered. However, as noted earlier, some forms of incremental reinforcement learning tend to be highly context specific, with learning most strongly expressed within the acquisition context; this is particularly true for habits (Bouton, 2002; Yin & Knowlton, 2006). For example, rodents extinguish a learned stimulus-response habit more effectively in a different physical context from where it was learned as opposed to the same context (McDonald, King, & Hong, 2001), and recovery of an extinguished association is improved when attempted in the original learning context (Bouton & Moody, 2004; LaBar & Phelps, 2004). Concerning the current work, although the first two experiments suggest that FPF induces a shifted mapping between given levels of abstracted memory evidence and certain overt memory classifications, it remains unclear whether this could simply represent a motor preference for a given response key in a given context; a basic form of motor reinforcement learning. In other words, when memory for a given item was low, context cues might trigger a motor preference, yielding several responses that were not based on the memory decision process. Thus, two related questions remain open: is FPF-based learning context sensitive? Further, can it be explained as a simple motor preference?

FPF-based learning likely operates on general memory evidence abstracted from the stimuli, as evidenced by Experiment 2. Presumably this memory evidence is a signal generated via the medial temporal lobe (Squire, 1992), and there is evidence that the MTL projects directly onto the dopaminergic midbrain and striatum (Cohen, Schoene-Bake, Elger, & Weber, 2009; Di

Martino et. al., 2008; Haber & Knutson, 2010; Rose, Haider, Weiller, & Büchel, 2002), regions that support various forms of reinforcement learning. However, parts of the medial temporal lobe – specifically the hippocampal formation – are also important for recording spatial location (Devan, Goad, & Petri, 1996; Eichenbaum, 2000; Nadel, 1991; Nadel, Hoscheidt, & Ryan, 2013; Squire, 1992; Stella et. al., 2012). Further, basic forms of implicit association learning in which observers gradually learn to attend to different spatial locations during visual search tasks based on contextual cues have been shown to be MTL dependent (Chun & Jiang, 2003; Chun & Phelps, 1999). Rapid stimulus-response learning that sometimes occurs during repeated semantic classification tasks has also been shown to depend on the MTL (Schnyer, Dobbins, Nicholls, Schacter, & Verfaellie, 2006). As a collection, these works suggest that even basic forms of learning that appear implicit in nature (as the FPF effect appears to be) may nonetheless be linked to MTL processes, and thus may be highly contextually-specific.

The context-specificity of some explicit memory judgments is well-established, (Godden & Baddeley, 1975, 1980; though c.f. Smith, Glenberg, & Bjork, 1980, Smith & Vela, 2001 for further discussion on controversy), but the above research suggests that a variety of implicit learning phenomena may also be bound tightly to the learning context. This type of sensitivity might constrain the FPF effect, which Experiments 1 and 2 converge in suggesting is an implicitly acquired remapping of underlying memory evidence to overt memory judgments. If this effect were highly context specific, or simply reflected the repeated reinforcement of a given behavioral action, one would not expect it to generalize outside of this specific laboratory paradigm. On the other hand, if the learning were more context-general it would suggest that FPF engenders a genuine change in how memory evidence is considered. Thus, Experiment 3 was designed to test the context-specificity of the FPF effect by manipulating both A) the spatial

context of testing, and B) the motor response used to issue recognition judgments (i.e., computer versus writing). Figure 11 shows predicted results.

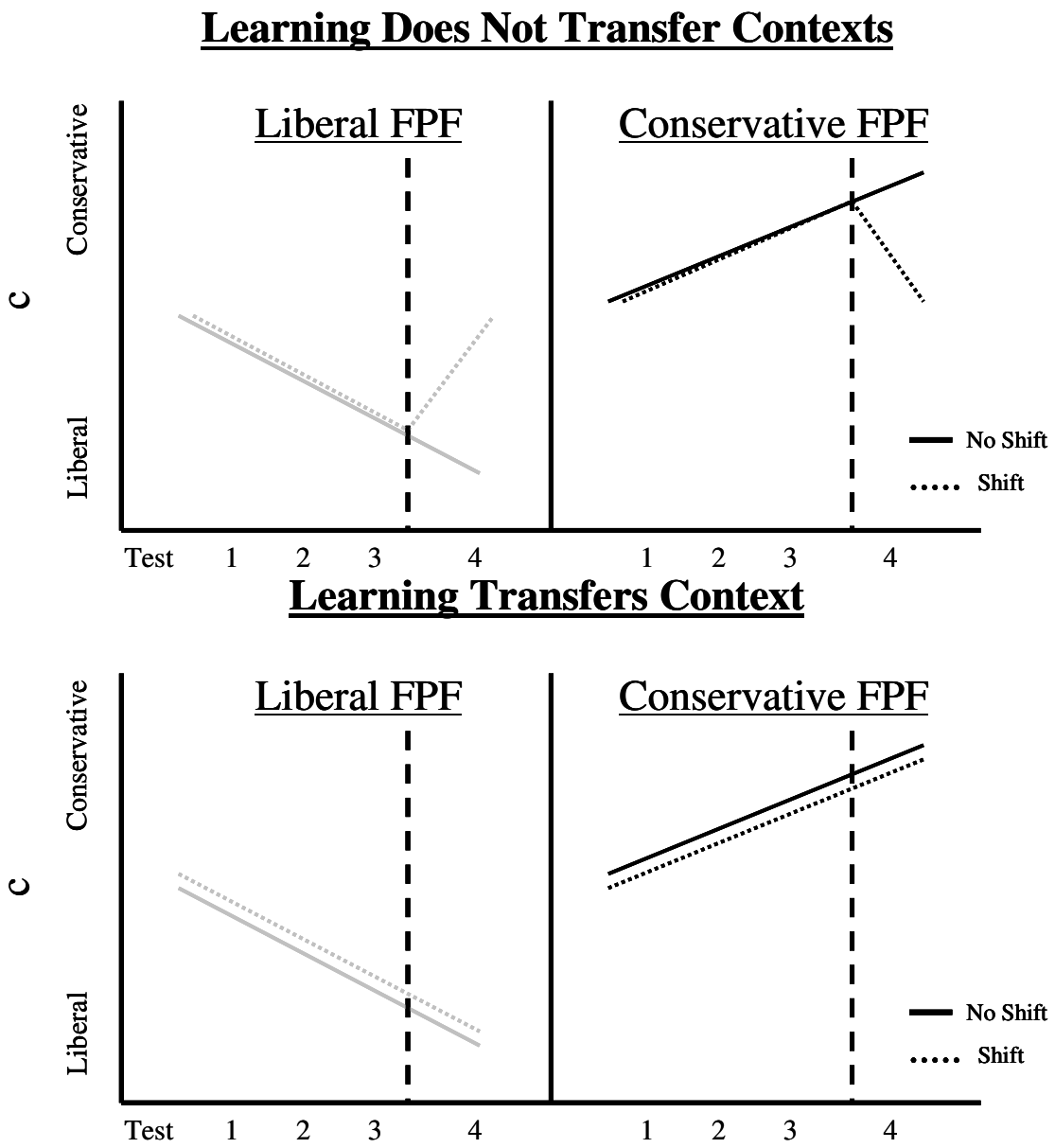


Figure 11: Hypothetical results. The top panel (predicted results) shows hypothetical results if the FPF effect is context-sensitive (i.e., does not transfer to a new context); criterion should return to baseline during the final test for the Shift Context groups. The bottom panel shows hypothetical results if the FPF effect is NOT context-sensitive; the FPF effect should remain in the new testing context for the Shift Context groups.

Experiment 3

Participants

114 individuals participated in Experiment 3 in return for either $15 or partial course credit.

Procedure

For this experiment, the BIS/BAS was presented before any study or test phases so participants could complete it before the experimental context manipulation. The experiment consisted of four study/test cycles, although the fourth test format differed for some participants (see below). The study and test phases were similar to Experiment 1. During study phases, 100 words were serially presented and participants rated the number of syllables in each word. Participants had two seconds to respond to each item; if this time was exceeded, the words 'TOO SLOW' appeared on screen and the computer moved onto the next trial. During test phases, 100 targets and 100 lures were randomly intermixed and presented serially; participants first indicated whether each item is old or new, followed by a confidence report ("low," "medium," or "high" confidence). Test phases were self-paced, and participants received FPF during the first three recognition tests. Each test was followed by questions 1-2 of the Subjective Awareness Questionnaire.

Following the fourth study cycle, participants were instructed to return to the experimenter. All participants were then taken to a separate conference room down the hall from the experimental space. This conference room was smaller than the original testing context and lacked any computers. The conference room also had one wall with a brick veneer, further

differentiating the contexts. All participants then completed paper copies of the GRAPES and RFQ in this separate room. Upon completing these two questionnaires, participants in the No Shift groups were taken back to the experimental space and completed the fourth recognition test at their original testing computer; this was designed to reinstate the original testing context and the original motor mappings (although no computer feedback was provided during this test). Participants in the Shift groups were given a paper copy of their fourth test (new testing context which also required a new motor response) and were read the following instructions before beginning:

> "You will now complete a pen and paper memory test. It will contain words you just studied intermixed with new words not previously shown in the experiment. Your task will be to determine whether each item is an old item from the most recent syllable-counting list or is a word that is new to the experiment by circling the appropriate answer in the "old/new" column. Following your old/new judgment, you'll indicate your confidence in that judgment by circling the corresponding confidence level in the "confidence column." For this test, it is very important that you answer each item in order and do not return to any previously answered items. Try to focus on each item individually when making your old/new and confidence assessments. If you have any questions, feel free to ask me."

Participants in the Shift groups did not receive feedback for their responses during the final test.

Participants in the Shift groups were taken back to their original testing computers following completion of the paper test. All participants completed the full Subjective Awareness

Questionnaire following the fourth recognition test. A final open-ended question was added to the Subjective Awareness Questionnaire for this experiment:

> "Briefly explain what you felt was the purpose of leaving the room during the experiment."

Results

Hits, false alarms, d', and c are presented in Table 6. Analyses in this section will initially examine performance during the first three tests as a function of Feedback (i.e., Liberal vs. Conservative FPF) in order to establish that FPF had an effect, then examine performance during the final test as a function of both Feedback and Context (i.e., No Shift vs. Shift) to investigate how context affected expression of the learned bias.

Liberal/No Shift					N = 29
Test	Hit Rate	FA Rate	d'	c	FPF trials
Test 1	.79 (.10)	.22 (.11)	1.67 (0.36)	-0.034 (0.31)	15.43 (8.04)
Test 2	.80 (.11)	.35 (.18)	1.34 (0.47)	-0.25 (0.40)	25.07 (13.61)
Test 3	.81 (.12)	.42 (.21)	1.18 (0.49)	-0.39 (0.49)	29.45 (14.84)
Test 4	.75 (.15)	.48 (.17)	0.81 (0.48)	-0.37 (0.46)	N/A

Conservative/No Shift					N = 29
Test	Hit Rate	FA Rate	d'	c	FPF trials
Test 1	.69 (.14)	.18 (.099)	1.52 (0.48)	0.23 (0.33)	20.69 (9.75)
Test 2	.63 (.19)	.19 (.10)	1.27 (0.55)	0.30 (0.42)	25.31 (13.17)
Test 3	.57 (.17)	.18 (.084)	1.12 (0.44)	0.39 (0.37)	30.07 (12.34)
Test 4	.50 (.20)	.24 (.12)	0.78 (0.56)	0.40 (0.45)	N/A

			Liberal/Shift		N = 27
Test	Hit Rate	FA Rate	d'	c	FPF trials
Test 1	.76 (.10)	.25 (.11)	1.48 (0.54)	-0.0088 (0.25)	17.89 (7.82)
Test 2	.81 (.096)	.37 (.20)	1.30 (0.57)	-0.29 (0.37)	27.81 (14.16)
Test 3	.83 (.11)	.50 (.24)	1.01 (0.62)	-0.52 (0.49)	35.59 (17.90)
Test 4	.70 (.11)	.44 (.13)	0.72 (0.39)	-0.20 (0.28)	N/A

			Conservative/Shift		N = 28
Test	Hit Rate	FA Rate	d'	c	FPF trials
Test 1	.71 (.13)	.15 (.096)	1.67 (0.43)	0.24 (0.31)	20.68 (10.32)
Test 2	.69 (.16)	.18 (.10)	1.51 (0.52)	0.22 (0.35)	22.36 (13.19)
Test 3	.60 (.19)	.16 (.092)	1.35 (0.57)	0.40 (0.38)	27.75 (14.06)
Test 4	.55 (.18)	.21 (.14)	0.98 (0.46)	0.37 (0.41)	N/A

Table 6: Hit rates, false alarm rates, accuracy, criterion, and number of FPF trials for Tests 1, 2, 3, and 4 (Experiment 3). Standard deviations in parentheses.

Accuracy

Accuracy during the first three tests was analyzed using a 2 × 3 mixed ANOVA examining a between-subjects factor of Feedback Group (Liberal vs. Conservative) and a within-subjects factor of Test (Test 1 vs. Test 2 vs. Test 3). The only significant effect on accuracy came from Test [$F(2,220) = 73.49$, $MSe = 0.069$, $p < .0001$, $\eta^2 = .40$], which indicated that accuracy declined across tests [$M_{T1} = 1.59$, $SE_{T1} = 0.043$; $M_{T2} = 1.34$, $SE_{T2} = 0.049$; $M_{T3} = 1.16$, $SE_{T3} = 0.051$]. Neither the effect of Feedback [$F(1,110) < 1$] nor the interaction [$F(2,220) = 1.55$, $MSe = 0.069$, $p = .21$, $\eta^2 = .014$] were significant.

Next, accuracy during the fourth test was analyzed using a 2 × 2 factorial ANOVA, examining factors of Prior Feedback (Liberal vs. Conservative) and Context (No Shift vs. Shift). There were no main effects or interactions (all p's > .11), indicating no accuracy differences among the four groups during the final test.

Criterion

Criterion during the first three tests was analyzed using a 2 × 3 mixed ANOVA examining a between-subjects factor of Feedback Group (Liberal vs. Conservative) and a within-subjects factor of Test (Tests 1-3). This analysis revealed a main effect of Feedback Group [$F(1,110) = 77.13$, $MSe = 0.33$, $p < .0001$, $\eta^2 = .41$], which indicated that participants in the Liberal group were more liberal than participants in the Conservative group [$M_L = -0.25$, $SE_L = 0.045$; $M_C = 0.30$, $SE_C = 0.044$]. There was also a main effect of Test [$F(2,220) = 12.67$, $MSe = 0.049$, $p < .0001$, $\eta^2 = .10$] which indicated a liberal trend in criterion across tests [$M_{T1} = 0.11$, $SE_{T1} = 0.028$; $M_{T2} = -0.009$, $SE_{T2} = 0.036$; $M_{T3} = -0.032$, $SE_{T3} = 0.041$]. These main effects were qualified by a robust Feedback Group × Test interaction [$F(2,220) = 49.81$, $MSe = 0.049$; $p < .0001$, $\eta^2 = .31$], which indicated that the two groups tended to increasingly diverge across tests (see Table 6 and Figure 12).

Before moving on to the next section, it was important to establish that there were no differences between the Context groups prior to the context manipulation. Future Context Group was added to the above ANOVA as a manipulation check. As would be expected, there was no main effect of Context Group and it did not interact with any other factors (all p's > .28).

The purpose of this experiment was to test the context-dependence of an implicitly learned criterion - that is, whether a criterion developed in one context would transfer to a vastly different testing context. To test this, criterion during Test 4 was examined using a 2 × 2 factorial ANOVA, examining factors of Prior Feedback Group (Liberal vs. Conservative) and Context Group (No Shift vs. Shift). There was a significant effect of Prior Feedback Group [$F(1,109) = 75.52$, $MSe = 0.17$, $p < .0001$, $\eta^2 = .41$], which indicated that participants who previously received Liberal feedback remained more liberal than participants who previously received

Conservative feedback [M_L = -0.29, SE_L = 0.055; M_C = 0.38, SE_C = 0.054], despite the absence of any feedback on the final test. The main effect of Context Group was not significant [$F(1,109) = 1.00$, MSe = 0.17, $p = .32$, $\eta^2 = .009$]. Surprisingly, the interaction among the factors was not significant [$F(1,109) = 1.64$, MSe = 0.17, $p = .20$, $\eta^2 = .015$], suggesting that context did not influence the effects of prior feedback learning (Figure 12).

The above analysis suggests that shifting context did not have any effect on criterion. In other words, differences in criterion as a result of FPF persisted despite a robust change in testing context, involving both a change in spatial location and response format. This notion can be tested more directly by contrasting criterion for the Liberal/Shift and Conservative/Shift groups. A significant difference between these two groups would provide yet further evidence that an implicitly learned criterion was insensitive to changes in context. Supporting this idea, criterion for the two groups was robustly different [$t(53) = -6.00$, $p < .0001$], which indicated that the Liberal/Shift group was more liberal than the Conservative/Shift group (see Table 6). In addition, criterion did not differ between the two Liberal groups [$t(54) = -1.70$, $p = .095$] nor the two Conservative groups [$t(55) < 1$]. These results all point to the conclusion that whatever is learned via FPF is not highly context-dependent. Further, the learning is not a simple motor preference as participants in the Shift groups completed a pen and paper version of the final test instead of a computerized version.

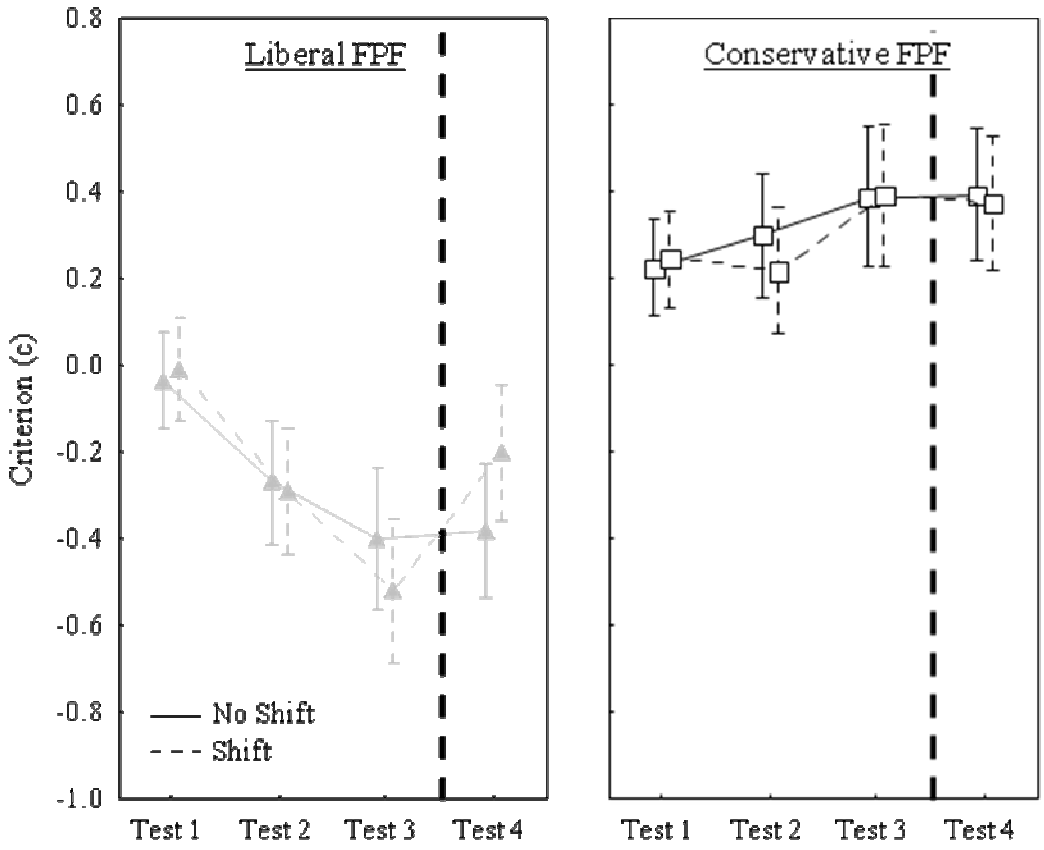


Figure 12: Effects of context change on criterion (Experiment 3). Dotted line indicates when all groups left the original testing room. No Shift participants returned to the original testing room to complete Test 4, while Shift participants remained in the new context and completed Test 4. Changing context did not appear to eliminate the effects of FPF (vertical bars represent 95% confidence intervals).

Confidence

Analysis of confidence was restricted to correct reports. Confidence across the first three tests was examined using a 2 × 2 × 3 mixed ANOVA examining a single between-subjects factor of Feedback Group (Liberal vs. Conservative) and two within-subjects factors, Response Type (Hit vs. CR) and Test (Test 1 vs. Test 2 vs. Test 3). There was no effect of Feedback Group on overall confidence [$F(1,110) < 1$]. There was a significant effect of Response Type on confidence [$F(1,110) = 64.25$, $MSe = 0.11$, $p < .0001$, $\eta^2 = .37$] which indicated that hits were

more confident than CRs [$M_H = 2.47$, $SE_H = 0.032$; $M_{CR} = 2.26$, $SE_{CR} = 0.042$]. The main effect of Test was not significant [$F(2,220) < 1$]. Turning to the interactions, the only significant two-way interaction was the Feedback Group × Response Type interaction [$F(1,110) = 14.36$, MSe = 0.11, $p < .001$, $\eta^2 = .12$]. This interaction indicated that the difference in confidence between hits and CRs was larger in the Liberal groups [$M_{LH} = 2.52$, $SE_{LH} = 0.045$; $M_{LCR} = 2.22$, $SE_{LCR} = 0.061$; $p < .05$, Tukey's] than in the Conservative groups [$M_{CH} = 2.41$, $SE_H = 0.044$; $M_{CCR} = 2.30$, $SE_{CCR} = 0.060$; $p < .05$, Tukey's]. Finally, the three-way interaction was significant [$F(2,220) = 8.89$, MSe = 0.021, $p < .001$, $\eta^2 = .075$]. This interaction indicated that hit and CR confidence tended to diverge in the Liberal feedback group, and tended to converge in the Conservative feedback group.

Confidence during the final test was examined using a 2 × 2 × 2 mixed ANOVA examining between-subjects factors of Prior Feedback Group (Liberal vs. Conservative) and Context (No Shift vs. Shift), and a single within-subjects factors of Response Type (Hit vs. CR). The only significant main effect was a main effect of Response Type [$F(1,109) = 22.74$, MSe = 0.074, $p < .0001$, $\eta^2 = .17$], which indicated that hits were more confident than CRs [$M_H = 2.24$, $SE_H = 0.034$; $M_{CR} = 2.07$, $SE_{CR} = 0.044$]. The only significant interaction was the Prior Feedback Group × Response Type interaction [$F(1,109) = 19.88$, MSe = 0.074, $p < .0001$, $\eta^2 = .15$]. This interaction indicated that while hits were more confident than CRs for the Prior Liberal feedback groups [$M_{LH} = 2.29$, $SE_{LH} = 0.048$; $M_{LCR} = 1.96$, $SE_{LCR} = 0.062$; $p < .001$, Tukey's HSD], there was no difference in response confidence for the Prior Conservative feedback groups [$M_{CH} = 2.20$, $SE_{CH} = 0.048$; $M_{CCR} = 2.18$, $SE_{CCR} = 0.061$; $p = .996$, Tukey's HSD].

Subjective Awareness Questionnaire

These data are shown in Table 7.

Believed Bias

Group	Test 1	Test 2	Test 3	Test 4
Liberal/No Shift	60.07 (10.86)	59.54 (12.39)	64.11 (13.80)	53.57 (13.53)
Conservative/No Shift	45.07 (14.62)	40.14 (12.24)	39.34 (11.62)	42.86 (15.30)
Liberal/Shift	62.48 (9.50)	67.22 (10.95)	64.85 (14.81)	58.27 (13.69)
Conservative/Shift	39.92 (15.89)	45.00 (14.08)	44.11 (16.12)	37.81 (16.28)

Believed Accuracy

Group	Test 1	Test 2	Test 3	Test 4
Liberal/No Shift	74.33 (9.68)	66.64 (13.59)	61.21 (14.97)	46.04 (15.76)
Conservative/No Shift	76.66 (8.20)	68.96 (10.60)	66.59 (11.80)	49.21 (14.92)
Liberal/Shift	75.63 (10.35)	71.48 (12.75)	67.07 (12.98)	53.81 (18.37)
Conservative/Shift	82.15 (7.40)	73.85 (12.08)	71.67 (13.69)	56.67 (20.70)

Believed Feedback Manipulation

Group	Hits	CRs	FAs	Misses	None
Liberal/No Shift	2	5	4	1	17
Conservative/No Shift	3	2	3	5	16
Liberal/Shift	3	0	5	1	18
Conservative/Shift	4	5	4	3	12

Feedback Influence

Group	Positive|"Old"	Positive|"New"	Negative|"Old"	Negative|"New"
Liberal/No Shift	4.59 (0.73)	4.59 (0.94)	4.45 (1.12)	4.62 (1.08)
Conservative/No Shift	4.38 (0.86)	4.03 (1.08)	4.72 (1.31)	4.52 (1.24)
Liberal/Shift	4.30 (0.95)	4.18 (1.08)	4.33 (1.21)	4.37 (1.18)
Conservative/Shift	3.67 (1.22)	3.68 (1.44)	4.46 (1.37)	4.14 (1.43)

Table 7: Subjective Awareness Questionnaire data, Experiment 3.

Believed Bias

Believed bias ratings were subjected to a 2 × 3 mixed ANOVA examining a between-subjects factor of Feedback Group (Liberal vs. Conservative) and a within-subjects factor of Test (Tests 1 - 3). This analysis revealed a main effect of Feedback Group [$F(1,102) = 117.76$, $MSe = 276.3$, $p < .0001$, $ç^2 = .54$] which mirrored the differences in criterion noted above [$M_L = 63.35$, $SE_L = 1.33$; $M_C = 42.92$, $SE_C = 1.33$]. Neither the effect of Test nor the two-way interaction were significant (both p's > .34).

Believed Bias during the final test was examined using a 2 × 2 factorial ANOVA, which examined factors of Prior Feedback Group (Liberal vs. Conservative) and Context Group (No Shift vs. Shift). Again, there was a main effect of Prior Feedback Group [$F(1,106) = 30.63$, $MSe = 217.7$, $p < .0001$, $\eta^2 = .22$] replicating the pattern noted above [$M_L = 55.92$, $SE_L = 2.01$; $M_C = 40.34$, $SE_C = 1.97$]. The main effect of Context Group was not significant [$F(1,106) < 1$]. The two-way interaction approached significance [$F(1,106) = 2.99$, $MSe = 217.7$, $p = .086$, $\eta^2 = .027$]. The nature of this interaction indicated that the Feedback Group difference between Believed Bias estimates was smaller for No Shift participants [$M_{LNS} = 53.67$, $SE_{LNS} = 2.79$; $M_{CNS} = 42.86$, $SE_{CNS} = 2.74$; $p < .05$, Tukey's] than for Shift participants [$M_{LS} = 58.27$, $SE_{LS} = 2.89$; $M_{CS} = 37.81$, $SE_{CS} = 2.84$; $p < .01$, Tukey's].

Believed Accuracy

Mean believed accuracy ratings for Tests 1, 2, and 3 were analyzed using a 2 × 3 mixed ANOVA examining a between-subjects factor of Feedback Group (Liberal vs. Conservative) and a within-subjects factor of Test (Test 1 vs. Test 2 vs. Test 3). There was a significant main effect of Feedback Group [$F(1,108) = 4.59$, $MSe = 291$, $p < .05$, $\eta^2 = .041$] which indicated that the

Liberal Feedback participants rated their performance lower than did Conservative Feedback participants [$M_L = 69.20$, $SE_L = 1.34$; $M_C = 73.22$, $SE_C = 1.32$]. The main effect of Test was highly significant [$F(2,216) = 50.15$, $MSe = 64$, $p < .0001$, $\eta^2 = .32$], which indicated that believed accuracy ratings declined across tests [$M_{T1} = 77.14$, $SE_{T1} = 0.87$; $M_{T2} = 69.98$, $SE_{T2} = 1.17$; $M_{T3} = 66.51$, $SE_{T3} = 1.30$]. Post-hoc tests indicated all three test means were significantly different from each other (all p's < .005). Finally, the two-way interaction was not significant [$F(2,216) < 1$]. The perceived accuracy difference between the Feedback Groups likely reflects the slight numerical accuracy advantage possessed by Conservative participants in this experiment (see Table 6).

Subjective performance during Test 4 was analyzed using a 2 x 2 factorial ANOVA, examining between-subjects factors of Prior Feedback (Liberal vs. Conservative) and Context (No Shift vs. Shift). The main effect of Prior Feedback was not significant [$F(1,107) < 1$]. The main effect of Context Group was significant [$F(1,107) = 5.24$, $MSe = 307.1$, $p < .05$, $\eta^2 = .047$], which indicated that No Shift participants rated their performance during Test 4 lower than did Shift participants [$M_{NS} = 47.62$, $SE_{NS} = 2.32$; $M_S = 55.24$, $SE_S = 2.38$], despite there being no significant differences in performance during Test 4 between these two groups. This may reflect the No Shift participants being used to receiving feedback in the original testing context; this surprising change from the original context may have altered how they perceived their own performance. The two-way interaction among the factors did not approach significance [$F(1,107) < 1$].

As in the previous two experiments, believed accuracy was contrasted with actual (percent correct out of 100) and the result was compared to 0 (i.e., perfect postdiction) for each test. This analysis collapsed across Feedback Group in order to increase power. Participants

postdicted their own performance relatively accurately during Test 1 [M = -0.32, SE = 0.89, t(109) = -0.36]. Participants underestimated their own performance during Test 2 [M = -2.69, SE = 1.11, t(110) = -2.42, $p < .05$] and Test 3 [M = -2.84, SE = 1.30, t(110) = -2.18, $p < .05$]. Believed minus actual accuracy during Test 4 was evaluated separately for the Context Groups. No Shift participants underestimated their own performance during Test 4 [M = -5.38, SE = 1.77, t(56) = -3.03, $p < .01$], whereas Shift participants postdicted their own performance relatively accurately [M = -0.16, SE = 1.86, t(53) = -0.084, p = .93]. The Context Group difference between these estimation differences approached, but was not significant [t(109) = 1.72, p = .09].

Feedback Influence

Participants rated their perceived influence of the various feedback outcomes (positive/negative following "old," and positive/negative following "new") on 6 point Likert scales. Feedback influence ratings were analyzed using two separate $2 \times 2 \times 2$ mixed ANOVAs examining a between-subjects factor of Feedback Group (Liberal vs. Conservative), and within-subjects factors of Feedback Valence (Positive vs. Negative) and Response Type ("old" vs. "new"); separate ANOVAs were run for the two Context Groups in order to simplify the resulting analyses. First examining the No Shift participants, the only significant effect was an interaction between Feedback Group and Response Type [F(1,56) = 4.38, MSe = 0.43, $p < .05$, $\eta^2 = .072$]. This interaction indicated that while the Liberal participants thought feedback following either response type equally influential [$M_O = 4.52$, $SE_O = 0.16$; $M_N = 4.60$, $SE_O = 0.15$] while Conservative participants thought feedback following "old" reports was more influential [$M_O = 4.55$, $SE_O = 0.16$; $M_N = 4.28$, $SE_N = 0.15$], although none of the post-hoc

comparisons among these values were significant (all p's > .12, Tukey's HSD). No other main effects or interactions were significant for the No Shift participants (all p's > .12).

Turning to the Shift participants, there was a significant main effect of Feedback Valence [$F(1,53) = 6.99$, $MSe = 1.07$, $p < .05$, $\eta^2 = .12$] which indicated that these participants rated negative feedback as more influential than positive feedback [$M_{Pos} = 3.96$, $SE_{Pos} = 0.14$; $M_{Neg} = 4.33$, $SE_{Neg} = 0.15$]. The Feedback Group × Feedback Valence interaction approached significance [$F(1,53) = 3.41$, $MSe = 1.07$, $p = .07$, $\eta^2 = .06$]. This interaction indicated that while Liberal participants rated positive and negative feedback as equally influential [$M_{Pos} = 4.24$, $SE_{Pos} = 0.21$; $M_{Neg} = 4.35$, $SE_{Neg} = 0.22$], Conservative participants felt that negative feedback was more influential than positive feedback [$M_{Pos} = 3.68$, $SE_{Pos} = 0.20$; $M_{Neg} = 4.30$, $SE_{Neg} = 0.21$]; this latter difference was significant according to Tukey's ($p < .05$).

Effects of Awareness

As with previous experiments, awareness was coarsely classified based on whether participants correctly guess how their feedback was manipulated. By this criterion, only 17 out of 114 participants qualified as aware (Liberal/No Shift = 4, Conservative/No Shift = 5, Liberal/Shift = 5, Conservative/Shift = 3). The number of participants selecting the correct feedback manipulation did not differ significantly from chance for either Feedback Group (Liberal: .140 ± .0944; Conservative: .158 ± .100). Additionally, the number of participants selecting "I don't agree with any of the [feedback] options" was greater than chance for both Feedback Groups (Liberal: .632 ± .131; Conservative: .491 ± .137). Fisher's exact test demonstrated no substantive differences in the distribution of aware vs. unaware participants

between the Shift and No Shift groups ($p > .99$, two-sided) or the Liberal and Conservative groups ($p > .99$, two-tailed).

As with previous experiments, the criterion analyses were re-run with the aware subjects excluded. None of the reported findings were affected.

Personality Measures

The BIS/BAS, RFQ, and GRAPES questionnaires did not correlate with any of the measures described above.

Discussion

This experiment was designed to test the hypothesis that FPF-induced criterion shifts are sensitive to the context in which they are learned. In other words, a response bias learned in one location should not show transfer to an entirely different testing context. Surprisingly, this was not the case; criterion learning was robust whether tested in the same context as the learning or a new context.

Further, FPF-based criterion learning does not reflect the reinforcement of a particular motor response tendency. This hypothesis was tested indirectly by the Shift Context groups. In many forms of implicit learning, what is learned is a binding of a given stimulus with a given distal response location (Ashby & Maddox, 2005; Willingham, Wells, Farrell, & Stemwedel, 2000). In these types of learning tasks, inconsistency in the response mappings (e.g., changing the locations of the responses) disrupts the expression of implicit knowledge (Ashby, Ell, & Waldron, 2003; Maddox, Bohil, & Ing, 2004). In the current experiment, participants in the Shift Context groups completed a pen and paper version of the final recognition test. If FPF induced a

more habitual or procedural-based style of learning, this switch alone should have disrupted the response bias that was learned. Since this experiment showed intact response bias despite changes in response format, it would seem that whatever is learned is not necessarily bound to a particular motor mapping or distal response location.

It is also interesting to note that participants in the Shift context group did not update their behavior despite being able to view their own response tendencies. Consider how a student might react when answering 'C' several times in a row on a multiple choice test – this would normally produce a feeling that something must be aberrant with either the test or the answers being reported. In this experiment, participants did not seem perturbed by runs of responses due to overly lax or strict criteria. Observers likely evaluate each item individually and do not aggregate across their response histories – even when those histories are sitting in front of them. Observers would likely have to be cued to pay attention to this sort of responding for it to be salient enough to advocate modifying their decision strategies (see Cox & Dobbins, 2011).

Whatever is learned via FPF appears insensitive to the context in which it was learned. However, caution should be expressed as this assertion relies on the null outcome. One could argue that the context manipulation wasn't powerful enough to disrupt learning; in this case, context could be more than just location and test format. However, prior work has shown that a new testing room can serve as an adequate-enough context shift to disrupt recall (Godden & Baddeley, 1975, 1980; Smith, Glenberg, & Bjork, 1980; Smith & Vela, 2001). Indeed, as described above, the shifted context was quite different from the original testing context, at least visually.

There were some interesting group differences in the Subjective Awareness Questionnaire, particularly in the Believed Accuracy question during Test 4. Participants who

returned to the original testing context believed themselves to have identified items more poorly than participants who tested in the new context. This is likely due to a variety of reasons. It could be that participants had begun to rely on the biased feedback during the prior tests, and the sudden removal of the feedback reduced their confidence in their own abilities. However, participants had generally underestimated their own performance on the previous tests as well, so underestimation may have been normal behavior. It may be more likely that the new testing context and format boosted confidence for the Shift Context participants. That is, participants may have felt more comfortable with the paper and pencil testing format, or they felt more comfortable with the new testing environment. Regardless, further research would have to be conducted to narrow down the list of possibilities.

Turning to the Believed Bias question, this experiment demonstrated that Feedback Group differences in Believed Bias persisted during the final test along with Feedback Group differences in criterion. This is in contrast with Experiment 1 where Feedback Group differences in Believed Bias disappeared despite persistent differences in criterion. This may be due to how participants were told the test list would be constructed. During the Subsequent Memory test in Experiment 1, participants were told that some of the items were coming from prior encounters, and some were new; participants in Experiment 3 just encountered a fourth version of the same procedure they'd been experiencing. The subsequent memory instructions may have served as a cue to more participants that the test was likely to be closer to 50/50 old and new items. For Experiment 3 participants, the similar test format likely reinforced prior beliefs about previous test makeup. This may be especially true for the Shift Context participants: because they took a written version of their final test, they could see their runs of responses. This is reflected in the Prior Feedback Group × Context Group interaction on Believed Bias (see Believed Bias section).

Participants who shifted to a new context made more extreme estimates of the test makeup than did participants who returned to the original testing context. Actually seeing a long run of "old" responses likely reinforced any prior beliefs about the test's perceived uneven construction rather than any beliefs about the feedback's nature.

Chapter 8

Evaluating a Temporal Difference Framework Prediction about Feedback Influence

5) Do unexpected positive outcomes drive learning more than do expected positive outcomes?

Under the temporal difference framework described in Chapter 3, learning is driven by errors of prediction. In other words, learning occurs when an organism mispredicts (either positively or negatively) the outcome of an action. This is germane to the FPF paradigm because the biased feedback is only presented during error trials. Presumably these are trials wherein the participant is expecting a low likelihood of success, thus a positive outcome should come as a surprise. These unexpected positive outcomes to a given type of low confidence error response (e.g., 'old'/false alarms) should drive learning – particularly, more so than a positive outcome to the analogous high confidence correct response (e.g., 'old', hits). In both of these cases the observer has responded old, and yet in the case of erroneous reinforced 'old' responses the outcome should be considerably more surprising that correct reinforced 'old' responses. There was some support for this notion in the Subsequent Memory test from Experiment 1 - errors that received FPF were subsequently remembered better than errors that received negative feedback. However, this is somewhat indirect since the core hypothesis of the FPF approach is that 'learning' in the paradigm reflects the acquisition of a judgment bias, not the memory of individual events that are linked to the development of the bias. Thus a more direct test of this would be a demonstration that trials in which false positive feedback occurs are more powerful in inducing subsequent biases than trials in which the same response is correctly reinforced. To answer this, trial-level data from Experiments 1 and 3 were analyzed using logistic regression.

Procedure

Trial-level data were collected from Experiments 1 and 3. These experiments were chosen due to the similarity in their design, as well as the fact that all trials received feedback during these experiments (i.e., not all trials received feedback during Experiment 2; thus it would require a different model). Trial-level data from Tests 2 and 3 were chosen from Experiment 1, and trial-level data from Tests 1 through 3 were chosen from Experiment 3; participants were separated by experiment for this analysis. Liberal and Conservative participants were analyzed separately using slightly different regression models (see below).

The analysis used what is sometimes referred to as a summary statistics approach and also referred to as random coefficients regression (Gumpertz & Pantula, 1989), in which each subject is modeled individually at level 1, and inferences about the reliability of any effects are determined at a second level (level 2) by considering the distributions of regression coefficients obtained at level 1 across the participants. As a simple example, one could assess whether subjects demonstrated above chance accuracy by modeling each separate individual's responses (1='old' and 0 = 'new') as a function of the presented stimulus type (1='old' and 0 = 'new'). The regression coefficient of each subject modeled thusly would indicate the degree of correspondence between his or her responses and the actual stimulus class, and hence accuracy. If these coefficients reliably diverged from a null value, then the group as a whole would be deemed more accurate than chance. This latter inference treats the subjects as a random effect, allowing a population inference. Of course, this two step approach is unnecessary when a summary accuracy statistic, such as d', can be calculated; however, in random coefficients regression one can model the separate contributions of multiple factors (other than stimulus type)

that may impinge on the subjects' response tendencies. Indeed, as shown below, one can even model the influence of prior outcomes on current response tendencies.

For the current paradigm, each participant's trial-wise data were dummy coded into several regressors. The outcome variable was the response on trial N, coded as 0 for "new" and 1 for "old." The first regressor (Item Type) represented the item type on trial N, coded 0 for a lure and 1 for a target. The second regressor was a set of four dummy coded variables representing the feedback outcome on trial N-1. The reference condition for the set of dummy variables was the correct, positive feedback outcome for the type of response for that group was exposed to FPF; these variables differed for Liberal and Conservative participants. For example, the reference condition for the Liberal Feedback group was hits, whereas the reference condition for the Conservative Feedback group was correct rejections. In this way, the remaining dummy coded feedback regressors reflect the influence of the remaining types of feedback outcomes on trial N-1 on the subjects responses on trial N, relative to a reference feedback condition of correctly reinforced hits or correctly reinforced correct rejections (again on trial N-1). Critically, these prior feedback influences are measured with subject accuracy statistically controlled, as this effect is modeled by the Item Type regressor for each subject.

The Liberal subjects' responses were modeled as:

Response(Trial N) = Item Type(Trial N) + PosFB CR(Trial N-1) +

NegFB M (Trial N-1) + **PosFB FA(Trial N-1)** + NegFB FA(Trial N-1)

The Conservative subjects' responses were modeled as:

Response(Trial N) = Item Type(Trial N) + PosFB H(Trial N-1) +

NegFB FA(Trial N-1) + **PosFB M(Trial N-1)** + NegFB M(Trial N-1)

For both groups, the first term in the model represents the contribution of the current item type to the response on the current trial. The remaining four terms are categorical dummy variables reflecting four potential feedback outcomes on the prior trial. The bolded variable represents FPF on the previous trial. The fifth potential feedback outcome (e.g., positive feedback for a hit for the Liberal subjects) serves as the reference condition for the dummy variables. Thus the theoretical question becomes "do prior responses that receive FPF drive subsequent responding over and above the same response that received true positive feedback?"

"Confidence on Trial N-1" was added as a regressor in a secondary analysis. The confidence parameter was not significant; thus the simplified model is presented below.

Results

A logistic regression model was fit to each individual using the glmfit function in Matlab. Parameter estimates obtained from this procedure were analyzed at the second level using a boostrapping procedure. Specifically, the logistic regression coefficients from a given group, for a particular variable, were repeatedly sampled with replacement and a mean was calculated for each sample. This procedure was repeated 10,000 times for the coefficient under consideration to construct a sampling distribution of the mean of that coefficient. The 95% confidence interval was constructed by determining where the middle 95% of the empirical sampling distribution fell. If this range excluded 0, the coefficient was deemed reliably different from the null.

Parameter estimates are shown in Figures 13 and 14. Of particular interest was whether the false feedback parameter (FA+ and M+) was significantly different from 0 for any group. Such a result would suggest that unexpected positive outcomes influenced subsequent

responding over and above expected positive outcomes (i.e., a correct response), that is, whether false positive feedback reliably biases individuals more than true positive feedback following the same response. Indeed, such was the case for three out of the four groups analyzed. This analysis was not significant for the Liberal FPF participants from Experiment 1 [M = 0.104, $CI_{95\%}$ = -0.014 : 0.23] although it approached significance as evidenced by the confidence interval. This parameter was significant for the Liberal participants from Experiment 3 [M = 0.239, $CI_{95\%}$ = 0.13 : 0.34], which indicated that FPF false alarms increased the odds of a subsequent "old" report over and above the influence of a hit. Turning to the Conservative participants, this parameter was significant for Conservative participants from Experiment 1 [M = -0.156, $CI_{95\%}$ = -0.296 : -0.017]. This parameter was also significant for the Conservative participants from Experiment 3 [M = -0.243, $CI_{95\%}$ = -0.35 : -0.14].

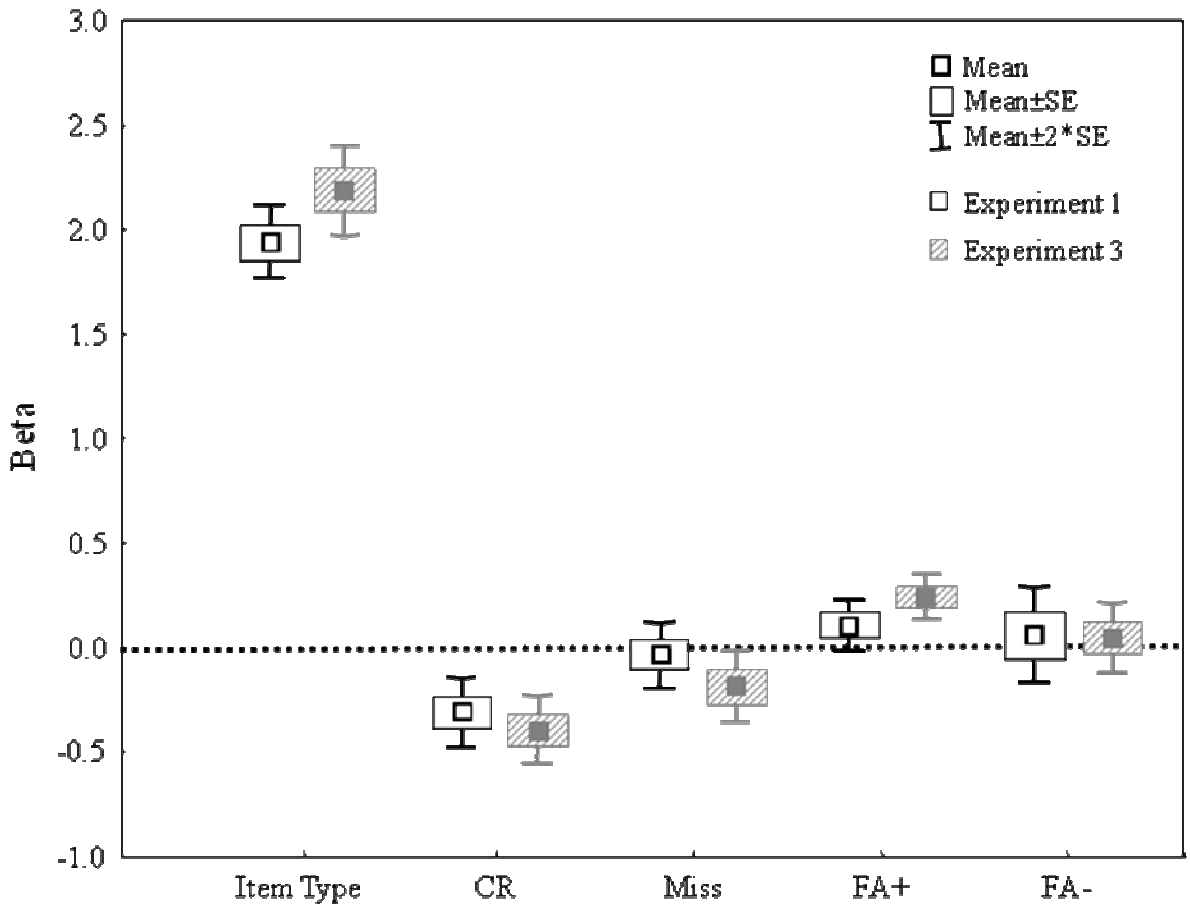


Figure 13: Logistic regression betas for Liberal participants. FA+ indicates false alarms that received FPF, while FA- indicates false alarms that received negative feedback.

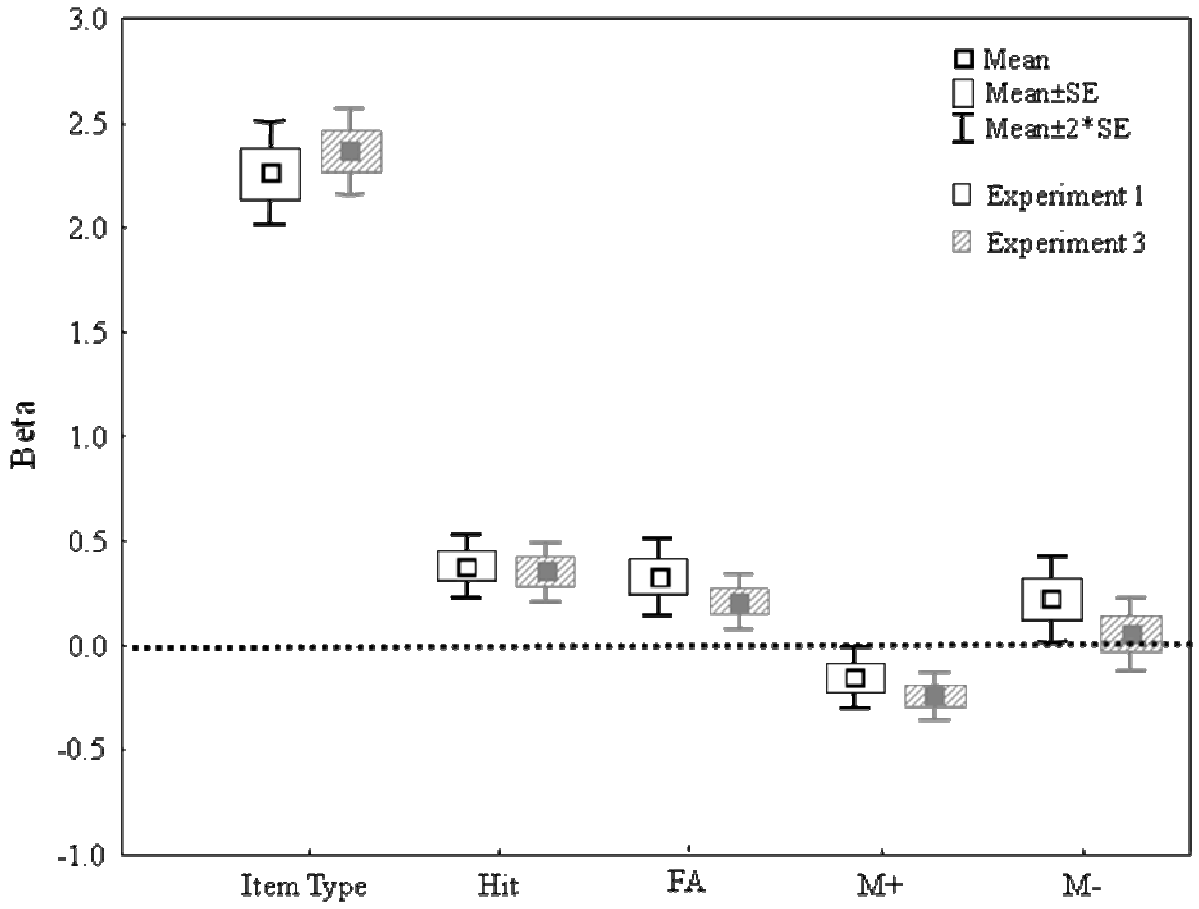


Figure 14: Logistic regression betas for Conservative participants. M+ indicates misses that received FPF, while M- indicates misses that received negative feedback.

Discussion

The modeling results all converge on the same idea: namely, that unexpected positive outcomes lead to larger changes in subsequent behavior than do expected positive outcomes. This is consistent with the reinforcement learning idea that learning increases with the size of prediction error (Dayan & Daw, 2008; Sutton & Barto, 1981), in this case, positive prediction error. The idea is that misses and false alarms are made with low confidence under the assumption that they may likely be incorrect. The unexpected (false) positive feedback elicits a

large prediction error, shifting behavior to respond in accordance with the feedback leading to the inducement of a subsequent bias more so than the receipt of positive feedback for the same response when correct. This yields a robust change in responding, even when FPF is only delivered on a very small minority of trials (see experimental tables for values).

As mentioned in the Procedure section, a secondary analysis added response confidence to the model. From a prediction error standpoint, low confidence errors that receive FPF should drive responding more so than high confidence errors. In other words, one would expect to find an interaction between "Confidence on Trial N-1" and "FA+/M+ on Trial N-1." This was not the case. Upon further consideration, confidence is in fact already represented in the simplified model in the reference condition for both Liberal and Conservative participants (i.e., hits and correct rejections, respectively). Hits are more confident than false alarms, and correct rejections are more confident than misses. Because the observer has no knowledge of whether he is committing an error, these responses simply differ in their overall feeling of confidence. In this way, the reference condition represents "high confidence old" and "high confidence new" reports to Liberal and Conservative participants. The corresponding FPF errors represent "low confidence old" and "low confidence new" reports to the Liberal and Conservative participants. This explains why there was no interaction between confidence and prior feedback in the secondary version of the model – the effect of confidence was essentially already captured in the various response types.

Chapter 9

General Discussion and Conclusions

Three experiments were conducted examining the potential influence of an implicit learning mechanism in governing the placement of the episodic recognition criterion. The topic is important because with very few exceptions (Han & Dobbins, 2008, 2009) demonstrations that implicit learning processes may directly influence the way explicit memories are reported are rare and could have implications in applied areas such as eyewitness identification. The current thesis examined this topic by addressing three basic questions applied to a FPF manipulation, namely, a) can observers ignore FPF, b) does the FPF effect transfer across vastly different perceptual domains, and c) is the FPF effect sensitive to shifts in testing context and motor response requirements. As discussed below, the answers to these questions buttress the idea that the FPF effect relies on implicit learning and further, they refine the operating characteristics of this learning.

Experiment 1 demonstrated that, even when explicitly asked to do so, observers could not block the biasing influence of FPF by attempting to ignore the feedback information. When coupled with the awareness questionnaire data demonstrating chance levels of identifying the actual manipulation in place (and the similar rates of detection across ignore and use groups) the data are consistent with an implicit phenomenon. Critically, this appears to be the first study that examines the ability of observers to volitionally ignore feedback during learning and it appears that observers are largely unable to do so in this case. Aside from the fact that they appear unaware of the biasing nature of the feedback there are other aspects of the FPF manipulation that may potentiate its influence even when observers are attempting to ignore it. First and foremost, the vast majority of feedback in the paradigm is correct and hence frequently will

conform to the participants' trial-wise beliefs about their own performance. More specifically, hits and correct rejections always received valid feedback indicating a correct response, and these are also the response types that tend to yield the greatest subjective confidence. Because the computer provides positive feedback on the trials in which the subjects most strongly believe they are correct, faith in the accuracy of feedback is likely quite strong; this point is further buttressed by the Subjective Awareness data indicating a strong preference for saying nothing was aberrant with the feedback. Additionally for each subject, one class of errors was always correctly signaled as incorrect, which would also conform to the low confidence subjects tend to exhibit during errors. Finally, the small subset of trials receiving the FPF manipulation was selected probabilistically, and hence one class of error receives probabilistic positive reinforcement; yet this appears sufficient to generate prominent shifts of decision criteria.

Speculatively, the ability of a small number of FPF trials (see Table 1) to evoke gross criterion shifts during IGNORE instruction probably reflects the fact that surprisingly positive outcomes may be particularly hard for subjects to ignore. This idea is supported by the modeling results discussed in Chapter 8, in that unexpected positive outcomes drove learning over expected positive outcomes. Additionally, Experiment 1 indicated the stimuli triggering FPF events were subsequently recognized better than those triggering correct error feedback. Although these findings point to an implicit learning phenomenon it is important to more fully outline what is meant by ‘implicit’ in the context of the FPF manipulation and the results of Experiment 1. The concepts of "expectation" and "surprise" are really explicit phenomena despite the learning in this paradigm presumably being implicit. What is "implicit" in this paradigm is the effect of FPF on criterion as it satisfies the key criteria of implicit learning in that: it occurs automatically, seemingly without conscious attempt to do so; the representations

that are learned appear to be outside of conscious access; and the learning is built up gradually across time (Frensch & Rünger, 2003). Further, to the extent that participants in the Ignore condition are actually attempting to suppress the feedback, the fact that it still demonstrates an effect further supports the notion that the learning is implicit (otherwise the induced criterion shift would be corrected by the observer).

The fact that observers could not ignore the feedback is in keeping with several findings from the implicit learning literature showing the limited utility of explicit control during an implicit learning process. This experiment provided an interesting contrast to the implicit learning literature. When observers are cued to try and seek meaning in an implicit learning task, the goal is to improve their overall performance. In general, this is not useful as the rules and structure that govern an implicit learning task are typically too complex to overtly discern (Ashby & Maddox, 2005; Reber, 1976). When rules are simpler, or observers are given more useful information, explicit strategies can actually aid in implicit learning performance (Reber, Kassin, Lewis, & Cantor, 1980). Although this is apparently the first study to examine how observers can ignore feedback designed to influence their behavior, the results are consistent with the idea that explicit control has little power over implicit learning.

Having established that observers cannot reliably ignore FPF, Experiment 2 examined the degree to which this learned phenomenon transferred across recognition test stimuli with grossly different features (words versus faces). This experiment confirmed the hypothesis that an implicitly learned criterion would transfer to a set of perceptually distinct stimuli to which recognition judgments were never reinforced. This suggests that, at some level, recognition stimuli, regardless of perceptual domain, are translated into an abstracted memory evidence signal. Despite the fact that the bias induced in word recognition reliably transferred to face

recognition judgment, the transfer was not perfect as the induced biases in the latter were less extreme. It was argued that this was due to an incomplete overlap in the type of evidence being evaluated – although both words and faces share some level of underlying familiarity, presumably faces also carry idiosyncratic details independent from words. This begs the question as to whether one can fashion stimulus categories that are so perceptually disparate that transfer would fail to occur. However, in thinking about this possibility it is important to note a limitation of Experiment 2 in that there was no control for potential differences in the relative levels of familiarity and recollection processes across the two classes. This may be critical because reinforcement-induced decision biases may not transfer or express on trials in which the memoranda evoke vivid recollections, as these are presumably impervious to altered mappings of familiarity-based judgments. A strong test of this notion of shared abstract memory evidence would be to make the two stimuli even more distinct (e.g., use real-world scenes instead of faces) until the learning did not transfer.

Because Experiment 2 demonstrated considerable transfer learning and bolstered the notion that subjects are using an abstracted strength of evidence variable, it begs the question of whether subjects can adopt fundamentally different biases for intermixed stimulus classes. For example, could FPF be used to induce a liberal criterion for words and a conservative criterion for faces within the same subject during an intermixed test list? This represents an interesting avenue for further generalization research; based on the current characterization of this learning, such a manipulation should fail, with the resultant net criterion being relatively unbiased. This follows based on the notion that what is reinforced through the procedure is a particular mapping between an abstract evidence variable and decision process. If correct, then liberal and conservative FPF trials would tend to cancel or offset one another and this would represent an

important limitation of this form of learning. In contrast, it would presumably be easy to provide explicit instructions to the observers asking them to use different strategies for responding on the basis of stimulus class. For example, one could emphasize that observers "should be particularly cautious in endorsing faces as recognized since even new faces will often strike one as familiar," while noting that, "it is important that you identify as many studied words as possible such that even if a word only strikes you as mildly familiar you should endorse it." Alternatively, two different payoff matrices could be used that differentially penalized the type of error depending upon stimulus class. Regardless, this would serve as a useful contrast if, as anticipated, FPF was ineffective in inducing qualitatively different biases across stimulus classes within a recognition test list.

Finally, the third experiment demonstrated a somewhat surprising result - an implicitly learned criterion will easily transfer across vastly different spatial and testing contexts. The experiment was designed anticipating a potentially large disruption by manipulating both the spatial context (testing room) and testing context (test type). If disruption were observed then follow-up experiments would separately manipulate these two variables to isolate whichever lead to the disruption of the criterion learning. However, neither variable appeared to adversely affect the learning that occurred in the original testing context, and indeed, the learned bias perfectly transferred in the context shift condition. This lies in contrast with the procedural and habit learning literature whereby learning is strongly tied to the contextual cues associated with it (Graybiel, 2008; LaBar & Phelps, 2005) and suggests that the learning during FPF is even more abstracted than implied by Experiments 1 and 2. That said, there are several caveats regarding Experiment 3 to consider. First, it may be that the context shift of the rooms was too subtle to reliably interfere with the expression of the learning. Although prior reports have suggested that

room alterations can be effective (Godden & Baddeley, 1975, 1980; Smith, Glenberg, & Bjork, 1980; Smith & Vela, 2001), we did not go to great lengths to alter features such as lighting, color, temperature or ambient noise in an attempt to maximize the differences in the two rooms, although the rooms were of a different size and also differed with respect the presence of computers and the texture of one wall. Regardless, Experiment 3 clearly demonstrated that the FPF phenomenon does not rest on a learned motor preference and it is not hyperspecific with respect to the context of acquisition, although further work is required on this point.

In the current paradigm, we only manipulated positive outcomes, using these during presumably low confidence errors to induce biases. Another potentially interesting extension of the current work concerns the valence of the manipulated feedback. FPF increases the average proportion of positive outcomes for one decision. How would the results of this procedure change, if at all, if instead biased *negative* feedback were used? In other words, instead of errors being selectively reinforced, what if correct responses (namely low confidence correct responses) were selectively punished? From a temporal difference reinforcement perspective, there is no difference between a positive and negative prediction error in what they forecast to the subject (Dayan & Daw, 2008); from this perspective, one would not expect any differences in learning. However, an abundant amount of research shows negative feedback is processed differently from positive feedback, both behaviorally and neurally (Atallah, Frank, & O'Reilly, 2004; Shen, Flajolet, Greengard, & Surmeri, 2008). For instance, patients with unmedicated Parkinson's disease learn to avoid negative outcomes more effectively than they learn to pursue positive outcomes; medication reverses this behavior (Frank, Seeberger, & O'Reilly, 2004). Patients with severe depression tend to show a hypersensitivity to negative outcomes (e.g., fixation on negative feedback which disrupts further task performance; Eshel & Roiser, 2010). These

findings all suggest that a false negative feedback paradigm may yield different results from the FPF paradigm used here. Critically however, this would require constraining any false negative feedback to low confidence trials so as to not raise the subjects' suspicions. Nonetheless, a direct comparison of false negative and false positive feedback may yield interesting differences.

One issue that bears examination is the relationship between discriminability and the opportunity for learning. It seems reasonable that participants with lower discriminability should be more susceptible to the influence of FPF. In other words, lower discriminability should lead to larger shifts in criterion. This can be tested by examining the relationship between absolute amount of criterion learning (absolute amount of criterion change due to exposure to FPF) and accuracy on the initial FPF test. Collapsed across experiments, there was a strong negative relationship between initial FPF test accuracy (Test 2 for Experiment 1, Test 1 for Experiments 2 and 3) and absolute criterion change (absolute value of Test 3 minus Test 1 c; $r = -.32$, $p < .0001$). Thus, lower initial accuracy was associated with larger criterion shifts across tests. This is likely due to lower levels of discriminability increasing the likelihood of erring, yielding more opportunities for FPF to drive behavior.

In general, the confidence findings are consistent with the criterion findings in each experiment. However, hits were universally more confident than correct rejections, even in the Conservative groups when “new” reports were being disproportionally reinforced. This likely reflects the fact that when recollections occur, they are assigned the highest confidence level (see Jaeger, Cox, & Dobbins, 2012 for a discussion and simulation). Thus hits will generally always be assigned greater confidence than correct rejections.

The strong claim laid out here is that implicit criterion learning involves a fundamental remapping of a strength of memory evidence continuum onto a recognition decision process. As

a collection, Experiments 1 through 3 suggest not only an abstract evidence representation but also a fair degree of context flexibility. Nonetheless, the judgments were always ones of recognition, and so even when reinforcement was removed, the transfer was tested for a criterion induced *during* simple yes-no recognition to a transfer test (without reinforcement) *examining* simple yes-no recognition. Although Experiment 3 ruled out the notion that peripheral learned motor preferences contribute to the FPF effect, and Experiment 2 demonstrated flexible transfer of the learning across test stimulus categories, it could nonetheless be the case that what is learned is specific to judgments of recognizing. If however, the FPF does not induce an explicit response strategy and actually leads observers to subjectively feel that probes are less or more familiar then it could even generalize to memory domains outside of recognition judgment. For example, what would be the consequences of inducing a liberal bias using FPF, then having the observer engage in a source memory test where studied materials have been encoded under two prior sources ("Source A", "Source B," or "New")? Here the observer's retrieval orientation is presumably heavily focused on recovering prior contextual specifics, not on making simple judgments of recognition. Thus, a key question for future research would be whether the tendency to make source endorsements would be influenced by the FPF procedure administered during prior recognition tests.

The Basal Ganglia and Memory Decisions

An obvious future extension of the current work lies in the realm of functional neuroimaging. Prior to any actual study, it is possible to speculate on potential candidate brain systems involved in this task. The basal ganglia are a set of midbrain structures implicated in many different implicit learning tasks. For instance, more implicit styles of learning (as opposed

to explicit memorization strategies) in the weather prediction task are associated with activity in the striatum, a region of the basal ganglia targeted by dopamine neurons thought to be associated with reinforcement learning and reward prediction (Poldrack et. al., 1999). Although typically thought to be primarily motor structures, it is becoming clear that the basal ganglia play a role in higher order cognition as well (Koziol, Budding, & Suth, 2009). Presumably, the basal ganglia play at least a partial role in building associations between specific classifications and specific sets of cards. Indeed, several studies have demonstrated that patients with striatal damage have difficulty with the weather prediction task (Shohamy et. al., 2004; Shohamy, Kalanithi, & Gluck, 2008), whereas patients with hippocampal damage generally perform no differently than controls during the early half of testing (Knowlton, Squire, & Gluck, 1994).

Interestingly, the basal ganglia are often implicated in studies of recognition memory, particularly those that also involve feedback (Han, Heuttel, Raposo, Adcock, & Dobbins, 2010; McDermott, Szpunar, & Christ, 2009; Scimeca & Badre, 2012; Spaniol et. al., 2009). Computational models of basal ganglia function typically ascribe it a role in modulating representations of cognitive and motor actions in cortex (Atallah, Frank, & O'Reilly, 2004). In other words, the role of the basal ganglia in implicit learning tasks is to aggregate across feedback outcomes (i.e., prediction errors) and select the potential actions in cortex via distinct corticostriatal loops (Alexander & Crutcher, 1990; Atallah, Frank, & O'Reilly, 2004; Di Martino et. al., 2008; Wimmer & Shohamy, 2011). A similar model could easily be in place here; future work could be designed to model this effect using a neuroanatomically computational model.

Generalizing outside of the laboratory, it is possible that the basal ganglia mold everyday recognition decision-making. Wixted and Gaitain (2002) proposed a model whereby pigeons learn via feedback to approximate statistically optimal (i.e., likelihood ratio) decision making in

a delayed match to sample task, a simpler variant of human recognition memory tasks. They speculated that humans may rely on a similar mechanism to train their own recognition memory systems via a lifetime of reinforcement such that they develop response tendencies given particular levels of recognition evidence analogous to probability matching behavior. The Wixted and Gaitain (2002) learning account may explain why observers enter the lab with an established or habitual memory evidence-to-judgment mapping. However, it is clearly speculative since the proposed long term learning has never been documented. Furthermore, it faces hurdles in that it doesn't explain why there is considerable variance in the biases observers bring to the lab, nor does it explain how observers select the appropriate mapping from among the infinite range of situations they might encounter in the lab. For example, Benjamin (2003) demonstrated that observers are typically unaware that low frequency words are easier to recognize than high frequency words and will in fact often explicitly claim the reverse; yet they nonetheless show better recognition of, and more conservative responding to, low frequency versus high frequency words. Of course frequency is just one of a myriad of factors that influence recognition discrimination, and so this begs the question as to how the observer can appropriately alter mappings across these different manipulations without often explicitly understanding their implications for retrieval accuracy or having been previously been trained on comparable tasks outside the lab.

If implicit criterion learning is dependent on the basal ganglia, it stands to reason that patients with basal ganglia dysfunction should show little, if any criterion learning. For example, patients with Parkinson's disease demonstrate difficulty with a variety of implicit learning tasks (Frank, Seeberger, & O'Reilly, 2004; Knowlton, Mangels, & Squire, 1996; Packard & Knowlton, 2002; Shohamy, Myers, Onlaor, & Gluck, 2004). For example, Shohamy and

colleagues (2004) showed that Parkinson's patients performed worse on the weather prediction than matched controls. Further, they showed that patients tended to use sub-optimal explicit strategies (such as memorizing the most common outcome associated with a single card). This would suggest that Parkinson's patients might show resistance to the effects of FPF. An interesting future experiment would be to compare Parkinson's patients with age-matched controls in the FPF paradigm.

Conclusions

Three experiments were conducted examining how biased feedback influences judgments of recognition. It is clear from this work that FPF leads to a remapping of memory evidence to decision outcomes. The learning that is developed appears difficult to ignore, transfers between stimuli with vastly different perceptual features, and transfers between different spatiotemporal contexts. This work presented a brief first step in what is hoped to be an exciting new leg of research, and the great deal of overlap that exists between two seemingly disparate literatures. Several questions were answered, but like most research, several more were opened up; from basic questions about paradigm specificities to grand questions about the fundamental workings of recognition memory. Several potential areas of new research include further behavioral work, functional neuroimaging, and potentially neuropsychological work. Regardless, the current work solidifies the notion that implicit learning mechanisms influence the way we evaluate and make decisions about explicit memories.

References

Adcock, R. A., Thangavel, A. Whitfield-Gabrieli, S., Knutson, B., & Gabrieli, J. D. E. (2006). Reward-motivated learning: Mesolimbic activation precedes memory formation. *Neuron, 50*, 507-517.

Alexander, G. E., & Crutcher, M. D. (1990). Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends in Neurosciences*, *13*(7), 266-271.

Ashby, F. G., Ell, S. W., & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition, 31(7),* 1114-25. doi:10.3758/BF03196132

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56, 149*-78. doi:10.1146/annurev.psych.56.091103.070217

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review, 6(3)*, 363-378.

Atallah, H. E., Frank, M. J., & O'Reilly, R. C. (2004). Hippocampus, cortex, and basal ganglia: Insights from computational models of complementary learning systems. *Neurobiology of Learning and Memory, 82*, 253-267.

Ball, S. A., & Zuckerman, M. (1990). Sensation seeking, Eysenck's personality dimensions and reinforcement sensitivity in concept formation. *Personality and Individual Differences*, *11*(4), 343-353.

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

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition*, *31*(2), 297-305.

Benjamin, A. S. & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language, 51*, 159-172. doi:10.1016/j.jml.2004.04.001

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: applications to recognition memory. *Psychological review*, *116*(1), 84.

Bouton, M. E. (2002). Context, ambiguity, and unlearning: Sources of relapse after behavioral extinction. *Biological Psychiatry, 52(10)*, 976-986.

Bouton, M. E., & Moody, E. W. (2004). Memory processes in classical conditioning. *Neuroscience and Biobehavioral Reviews, 28*, 663-674

Brainard, D. H. (1997) The Psychophysics Toolbox, Spatial Vision 10:433-436. doi:10.1163/156856897X00357

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition Learning, 1*, 69-84.

Carterette, E. C., Friedman, M. P., & Wyman, M. J. (1966). Feedback and psychophysical variables in signal detection. *The Journal of the Acoustical Society of America, 39(6)*, 1051-1055.

Carver, C.S., White, T.L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the BIS/BAS Scales. *Journal of Personality and Social Psychology, 67*, 319-333.

Chun, M. M., & Jiang, Y. (2003). Implicit, long-term spatial contextual memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(2), 224.

Chun, M. M., & Phelps, E. A. (1999). Memory deficits for implicit contextual information in amnesic subjects with hippocampal damage. *Nature neuroscience*, *2*(9), 844-847.

Cohen, M. X., Schoene-Bake, J. C., Elger, C.E., & Weber, B. (2009). Connectivity-based segregation of the human striatum predicts personality characteristics. *Nature Neuroscience, 12*, 32-34.

Cox, J. C., & Dobbins, I. G. (2011). The striking similarities between standard, distractor-free, and target-free recognition. *Memory & Cognition, 39,* 925-940. doi: 10.3758/s13421-011-0090-3

Criss, A. H., & McClelland, J. L. (2006). Differentiating the differentiation models : A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM). *Journal of Memory and Language, 55*, 447-460. doi:10.1016/j.jml.2006.06.003

da Silva, L., & Sunderland, A. (2010). Effects of immediate feedback and errorless learning on recognition memory processing in young and older adults. *Neuropsychological Rehabilitation, 20(1)*, 42-58.

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective & Behavioral Neuroscience, 8(4)*, 429-453. doi:10.3758/CABN.8.4.429

Dayan, P., & Watkins, C. J. C. H. (2001). Reinforcement learning, Encyclopedia of Cognitive Science. *ed: MacMillan Press, UK*.

Devan, B. D., Goad, E. H., & Petri, H. L. (1996). Dissociation of hippocampal and striatal contributions to spatial navigation in the water maze. *Neurobiology of Learning and Memory*, *66*(3), 305-323.

Di Martino, A., Scheres, A., Margulies, D. S., Kelly, A. M. C., Uddin, L. Q., Shehzad, Z., Biswal, B., Walters, J. R., Castellanos, F. X., & Milham, M. P. (2008). Functional connectivity of human striatum: a resting state FMRI study. *Cerebral cortex*, *18*(12), 2735-2747.

Dobbins, I. G., & Kroll, N. E. (2005). Distinctiveness and the recognition mirror effect: evidence for an item-based criterion placement heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1186.

Eichenbaum, H. (2000). Hippocampus: Mapping or memory? *Current Biology, 10*, 785-787.

Eshel, N., & Roiser, J. P. (2010). Reward and punishment processing in depression. *Biological Psychiatry, 68*(2), 118-124.

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, 31(5)*, 1075–1095. doi:10.1037/0278-7393.21.5.1075

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C (2004). By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. *Science, 306*(5703), 1940-1943.

Frensch, P. A., & Rünger, D. (2003). Implicit learning. *Current Directions in Psychological Science*, *12*(1), 13-18.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16(1)*, 5-16.

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

Gluck, M. A., Shohamy, D., & Myers, C. (2002). How do people solve the "weather prediction" task?: Individual variability in strategies for probabilistic category learning. *Learning & Memory, 9*, 408-418.

Godden, D. R., & Baddeley, A, D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology, 66(3)*, 325-331.

Graybiel, A. M. (2008). Habits, rituals, and the evaluative brain. *The Annual Review of Neuroscience, 31*, 359-387.doi:10.1146/annurev.neuro.29.051605.112851

Gumpertz, M., & Pantula, S. G. (1989). A simple approach to inference in random coefficient models. *The American Statistician*, *43*(4), 203-210.

Han, S. (2009). Trait individual difference of reinforcement-based decision criterial learning during episodic recognition judgments. *Korean Journal of Cognitive Science, 20(3)*, 357-381.

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. doi: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. doi:10.3758/PBR.16.3.469

Han, S., Huettel, S. A., Raposo, A., Adcock, R. A., & Dobbins, I. G. (2010). Functional significance of striatal responses during episodic decisions: Recovery or goal attainment?. *The Journal of Neuroscience*, *30*(13), 4767-4775.

Healy, A. F., & Jones, C. (1975). Can subjects maintain a constant criterion in a memory task? *Memory & Cognition, 3(3)*, 233-238.

Healy, A. F., & Kubovy, M. (1978). The effects of payoffs and prior probabilities on indices of performance and cutoff location in recognition memory. *Memory & Cognition, 6(5)*, 544-553.

Heit, E., Brockdorff, N., & Lamberts, K. (2003). Adaptive changes of response criterion in recognition memory. *Psychonomic Bulletin & Review, 10(3)*, 718-723.

Henriques, J. B., Glowacki, J. M., & Davidson, R. J. (1994). Reward fails to alter response bias in depression. *Journal of Abnormal Psychology, 103(3),* 460-466.

Higgins, E. T., Friedman, R. S., Harlow, R. E., Idson, L. C., Ayduk, O. N., & Taylor, A. (2001). Achievement orientations from subjective histories of success: Promotion pride versus prevention pride. *European Journal of Social Psychology*, *31*(1), 3-23.

Hirsh, J. (1974). The hippocampus and contextual retrieval of information from memory: A theory. *Behavioral Biology, 12*(4), 421-444.

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21(2)*, 302-313.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*, 513-541.

Jaeger, A., Cox, J. C., & Dobbins, I. G. (2012). Recognition confidence under violated and confirmed memory expectations. *Journal of Experimental Psychology: General*, *141*(2), 282-301. doi: 10.1037/a0025687

Jaeger, A., Lauris, P., Selmeczy, D., & Dobbins, I. G. (2012). The costs and benefits of memory conformity. *Memory & cognition*, *40*(1), 101-112.

Kantner, J., & Lindsay, D. S. (2010). Can corrective feedback improve recognition memory? *Memory & Cognition, 38(4),* 389-406.

Koziol, L. F., Budding, D. E., & Suth, A. (2009). *Subcortical structures and cognition*. New York: Springer.

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science*, *273*(5280), 1399-1402.

Knowlton, B. J., Ramus, S. J., & Squire, L. R. (1992). Intact artificial grammar learning in amnesia: Dissociation of classification learning and explicit memory for instances. *Psychological Science, 3*(3), 172-179.

Knowlton, B. J., & Squire, L. R. (1994). The information acquired during artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(1), 79-91.

Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning & Memory, 1*, 106-120.

LaBar, K. S., & Phelps, E. A. (2005). Reinstatement of conditioned fear in humans is context dependent and impaired in amnesia. *Behavioral neuroscience*, *119*(3), 677.

Lampinen, J. M., Scott, J., Pratt, D., Leding, J. K., & Arnal, J. D. (2007). 'Good, you identified the suspect… but please ignore this feedback': Can warnings eliminate the effects of post-identification feedback? *Applied Cognitive Psychology*, *21*(8), 1037-1056.

Lau, B. & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior, 84(3)*, 555-579.

Ley, R., & Long, K. (1987). A distractor-free test of recognition and false recognition. *Bulletin of the Psychonomic Society, 25(6)*, 411-414.

Ley, R., & Long, K. (1988). Distractor similarity effects in tests of discrimination recognition and distractor-free recognition. *Bulletin of the Psychonomic Society, 26(5)*, 407-409.

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

Maddox, W. T., & Bohil, C. J. (1998). Base-rate and payoff effects in multidimensional perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24(6),* 1459-1482.

Maddox, W. T., & Bohil, C. J. (2005). Optimal classifier feedback improves cost-benefit but not base-rate decision criterion learning in perceptual categorization. *Memory & Cognition, 33(2)*, 303-319.

Maddox, W. T., Bohil, C. J., & Ing, A. D. (2004). Evidence for a procedural-learning-based system in perceptual category learning. *Psychonomic Bulletin & Review, 11(5)*, 945-952.

Mathews, R. C., Buss, R. R., Stanley, W. B., Blanchard-Fields, F., Cho, J. R., & Druhan, B. (1989). Role of implicit and explicit processes in learning from examples: A synergistic effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(6), 1083-1100.

McCabe, D. P., & Balota, D. A. (2007). Context effects on remembering and knowing: The expectancy heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 536-549. doi: 10.1037/0278-7393.33.3.536

McDermott, K. B., & Roediger, H. L. (1994). Effects of imagery on perceptual implicit memory tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1379-1390.

McDermott, K. B., Szpunar, K. K., & Christ, S. E. (2009). Laboratory-based and autobiographical retrieval tasks differ substantially in their neural substrates. *Neuropsychologia, 47(11),* 2290-8. doi:10.1016/j.neuropsychologia.2008.12.025

McDonald, R. J., King, A. L., & Hong, N. S. (2001). Context-specific interference on reversal learning of a stimulus-response habit. *Behavioural Brain Research, 121*, 149-165.

McKelvie, S. J. (1993). Effects of spectacles on recognition memory for faces: Evidence from a distractor-free test. *Bulletin of the Psychonomic Society, 31 (5),* 475-477.

Miller, M. B., Handy, T. C., Cutler, J., Inati, S., & Wolford, G. L. (2001). Brain activations associated with shifts in response criterion on a recognition test. *Canadian Journal of Experimental Psychology, 55(2)*, 162-173.

Morrell, H. E. R., Gaitan, S., & Wixted, J. T. (2002). On the nature of the decision axis in Signal-Detection-based models of recognition memory. *Cognition*, *28*(6), 1095-1110. doi:10.1037//0278-7393.28.6.1095

MRC Psycholinguistic Database: Machine Usable Dictionary. Version 2.00. Informatics Division Science and Engineering Research Council Rutherford Appleton Laboratory Chilton, Didcot, Oxon, OX11 0QX Michael Wilson 1 April 1987.

Nadel, L. (1991). The hippocampus and space revisited. *Hippocampus*, *1*(3), 221-229.

Nadel, L., Hoscheidt, S., & Ryan, L. R. (2013). Spatial cognition and the hippocampus: The anterior–posterior axis. *Journal of cognitive neuroscience*,*25*(1), 22-28.

Neal, D. T., Wood, W., Labrecque, J. S., & Lally, P. (2012). How do habits guide behavior? Perceived and actual triggers of habits in daily life. *Journal of Experimental Social Psychology, 48*, 492-498.

Neyman, J., & Pearson, E. S. (1992). *On the problem of the most efficient tests of statistical hypotheses* (pp. 73-108). Springer New York.

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.

O'Connor, A. R., Han, S., & Dobbins, I. G. (2010). The inferior parietal lobule and recognition memory: Expectancy violation or successful retrieval? *The Journal of Neuroscience, 30(8)*, 2924-2934.

Packard, M. G., & Knowlton, B. J. (2002). Learning and memory functions of the basal ganglia. *Annual Review of Neuroscience*, *25*(1), 563-593.

Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonparametric" A' and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review, 10(3)*, 556-569.

Poldrack, R. A., & Packard, M. G. (2003). Competition among multiple memory systems: Converging evidence from animal and human brain studies. *Neuropsychologia, 41*, 245-251.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*(3), 353-363.

Pothos, E. M. (2007). Theories of artificial grammar learning. *Psychological Bulletin, 133*(2), 227-244.

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99(3)*, 518-535. doi:10.1037//0033-295X.99.3.518

Reber, A. S. (1969). Transfer of syntactic structure in synthetic languages. *Journal of Experimental Psychology, 81*(1), 115-119.

Reber, A. S. (1976). Implicit learning of synthetic languages: The role of instructional set. *Journal of Experimental Psychology: Human Learning and Memory, 2*(1), 88-94.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General, 118*(3), 219-235.

Reber, A. S., Kassin, S. M., Lewis, S., & Cantor, G. (1980). On the relationship between implicit and explicit modes in the learning of a complex rule structure. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(5), 492.

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. doi:10.1037/0278-7393.33.2.305

Rotello, C. M., & Macmillan, N. A. (2008). Response bias in recognition memory. In Benjamin, A. S., & Ross, B. H. (Eds.), *Skill and Strategy in Memory Use: A Volume in The Psychology of Learning and Motivation* (pp. 61-94). London: Elsevier. doi: 10.1016/S0079-7421(07)48002-1

Rugg, M. D., & Curran, T. (2007). Event-related potentials and recognition memory. *Trends in cognitive sciences*, *11*(6), 251-257.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84*, 1-66.

Schnyer, D. M., Dobbins, I. G., Nicholls, L., Schacter, D. L., & Verfaellie, M. (2006). Rapid response learning in amnesia: Delineating associative learning components in repetition priming. *Neuropsychologia*, *44*(1), 140-149.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599.

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience*, *23*(1), 473-500.

Scimeca, J. M., & Badre, D. (2012). Striatal contributions to declarative memory retrieval. *Neuron*, *75*(3), 380-392.

Seger, C. A. (2008). How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. *Neuroscience and Biobehavioral Reviews, 32,* 265-278.

Selmeczy, D., & Dobbins, I. G. (2013). Metacognitive awareness and adaptive recognition biases. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 39*(3), 678-690. doi: 10.1037/a0029469

Selmeczy, D., & Dobbins, I. G. (2014). Relating the content and confidence of recognition judgments. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 40*(1), 66-85. doi: 10.1037/a0034059

Sheeran, P., Aarts, H., Custers, R., Rivis, A., Webb, T. L., & Cooke, R. (2005). The goal-dependent automaticity of drinking habits. *British Journal of Social Psychology, 44*, 47-63.

Shen, W., Flajolet, M., Greengard, P., & Surmeier, D. J. (2008). Dichotomous dopaminergic control of striatal synaptic plasticity. *Science*, *321*(5890), 848-851.

Shohamy, D., Myers, C. E., Grossman, S., Sage, J., Gluck, M. A., & Poldrack, R. A. (2004). Cortico-striatal contributions to feedback-based learning: Converging data from neuroimaging and neuropsychology. *Brain, 127*, 851-859.

Shohamy, D., Myers, C. E., Kalanithi, J., Gluck, M. A. (2008). Basal ganglia and dopamine contributions to probabilistic category learning. *Neuroscience and Biobehavioral Reviews, 32*, 219-236.

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition, 6(4)*, 342-353.

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review, 8(2)*, 203-220.

Spaniol, J., Davidson, P. S., Kim, A. S., Han, H., Moscovitch, M., & Grady, C. L. (2009). Event-related fMRI studies of episodic encoding and retrieval: Meta-analyses using activation likelihood estimation. *Neuropsychologia*, *47*(8), 1765-1779.

Squire, L.R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans". *Psychological Review, 99(2)*, 195–231. doi:10.1037/0033-295X.99.2.195

Strong, M. H., & Strong, E. K., Jr. (1916). The nature of recognition memory and of the localization of recognitions. *American Journal of Psychology, 27*, 341-362. doi:10.2307/1413103.

Stella, F., Cerasti, E., Si., B., Jezek, K., & Treves, A. (2012). Self-organization of multiple spatial and context memories in the hippocampus. *Neuroscience and Biobehavioral Reviews, 36(7)*, 1609-1625.

Sutton, R. S. & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88(2)*, 135-170.

Tom, S. B., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science, 315,* 515-518.

Tricomi, E., Balleine, B. W., & O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience, 29*(11), 2225-2232. doi:10.1111/j.1460-9568.2009.06796.x.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne*, *26*(1), 1.

Turner, B., Van Zandt, T., & Brown, S. (2011). A dynamic stimulus-driven model of signal detection. *Psychological Review, 118(4)*, 583-613.

Underwood, B. J. (1972). Word recognition memory and frequency information. *Journal of Experimental Psychology, 94*, 276-283. doi:10.1037/h0032785

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

Van Zandt, T., & Maldonado-Molina, M. M. (2004). Response reversals in recognition memory. *Journal of Experiment Psychology: Learning, Memory, and Cognition, 30(6)*, 1147-1166.

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

Wallace, W. P. (1978). Recognition failure of recallable words and recognizable words. *Journal of Experimental Psychology: Human Learning and Memory, 4(5)*, 441-452. doi:10.1037/0278-7393.4.5.441

Wallace, W. P., Sawyer, T. J., & Robertson, L. C. (1978). Distractors in recall, distractor-free recognition, and the word-frequency effect. *American Journal of Psychology, 91(2)*, 295-304. doi:10.2307/1421539

Wallace, W. P. (1980). On the use of distractors for testing recognition memory. *Psychological Bulletin, 88(3)*, 696-704. doi:10.1037/0033-2909.88.3.696

Wallace, W. P. (1982). Distractor-free recognition tests of memory. *American Journal of Psychology, 95(3)*, 421-440. doi:10.2307/1422134

Willingham, D. B., Wells, L. A., Farrell, J. M., & Stemwedel, M. E. (2000). Implicit motor sequence learning is represented in response locations. *Memory & Cognition, 28*(3), 366-375.

Wimmer, G. E., & Shohamy, D. (2011). The striatum and beyond: Contributions of the hippocampus to decision making. In M. R. Delgado, E. A. Phelps, & T. W. Robbins

(Eds.), *Decision Making, Affect, and Learning: Attention and Performance XXIII* (pp. 281–309). Oxford, UK: Oxford University Press.

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.

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review, 117(4)*, 1025-1054.

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review*, *11*(4), 616-641.

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews: Neuroscience, 7(6),* 464-476. doi:10.1038/nrn1919

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*(3), 441-517. doi:10.1006/jmla.2002.2864

Appendix A – Supplemental Material

Experiment 1

Reaction Time

As with confidence, analysis of mean reaction times (RTs) was restricted to correct reports. RTs were first converted to z-scores for each individual, then sorted by response type (e.g., hits and correct rejections). As with prior sections, RT during Test 1 was analyzed separately using a 4×2 mixed model ANOVA examining a between-subjects factor of Group, and a within-subjects factor of Response Type (Hit vs. Correct Rejection). This analysis revealed a main effect of Response Type [$F(1,92) = 14.90$, $MSe = 0.024$, $p < .001$, $\eta^2 = .14$] which indicated that hits were faster than correct rejections [$M_H = 0.15$, $SE_H = 0.023$; $M_{CR} = 0.24$, $SE_{CR} = 0.027$]. No other effects were significant (all p's > .30).

Next, RTs during Tests 2 and 3 were analyzed using a separate 2×2×2×2 mixed model ANOVA examining between-subjects factors of Feedback Group (Liberal vs. Conservative) and Volition (Use vs. Ignore), and within-subjects factors of Response Type (Hit vs. Correct Rejection) and Test (Test 2 vs. Test 3). This analysis revealed a main effect of Response Type [$F(1,92) = 26.24$, $MSe = 0.036$, $p < .0001$, $\eta^2 = .22$] which indicated that hits were faster than correct rejections [$M_H = -0.19$, $SE_H = 0.014$; $M_{CR} = -0.092$, $SE_{CR} = 0.016$]. There was also a significant main effect of Test [$F(1,92) = 48.93$, $MSe = 0.056$, $p < .0001$, $\eta^2 = .35$] which indicated that RTs decreased across tests [$M_{T2} = -0.058$, $SE_{T2} = 0.015$; $M_{T3} = -0.23$, $SE_{T3} = 0.018$]. No other main effects or interactions were significant (all p's > .11).

Finally, RT during the Subsequent Memory Test was analyzed using a 2×2×2 mixed model ANOVA examining between-subjects factors of Prior Feedback Group (Liberal vs. Conservative) and Prior Volition (Use vs. Ignore), and a single within-subjects factor of

Response Type (Hit vs. Correct Rejection). This analysis revealed a main effect of Prior Feedback Group [$F(1,92) = 7.03$, $MSe = 0.060$, $p < .01$, $\eta^2 = 0.071$] which indicated that the Prior Liberal feedback groups were significantly slower than the Prior Conservative feedback groups [$M_{PL} = -0.026$, $SE_{PL} = 0.026$; $M_{PC} = -0.12$, $SE_{PC} = 0.024$]; this likely reflects differences in the test construction between the groups (i.e., the Prior Liberal groups were rating prior lures from Test 2, whereas the Prior Conservative groups were rating prior targets from Test 2). There was also a main effect of Response Type [$F(1,92) = 8.82$, $MSe = 0.044$, $p < .01$, $\eta^2 = .087$] which indicated that hits were faster than correct rejections [$M_H = -0.12$, $SE_H = 0.023$; $M_{CR} = -0.028$, $SE_{CR} = 0.023$]. No other main effects or interactions were significant (all p's > .12).

Experiment 2

Reaction Time

As with Experiment 1, RTs were first converted to z-scores for each individual, then sorted by test and response type. Reaction time for correct reports was examined using a similar $2 \times 2 \times 2 \times 4$ mixed ANOVA, again examining between-subjects factors of Feedback Group (Liberal vs. Conservative) and within-subjects factors of Stimulus (Word vs. Face), Response (Hit vs. CR), and Test (Test 1 vs. Test 2 vs. Test 3 vs. Test 4). This analysis revealed a robust main effect of Stimulus [$F(1,50) = 192.54$, $MSe = 0.52$, $p < .0001$, $\eta^2 = .79$] which indicated that responses to words were faster than responses to faces [$M_W = -0.14$, $SE_W = 0.012$; $M_F = 0.56$, $SE_F = 0.043$], despite similar levels of accuracy. There was also a robust main effect of Test [$F(3,150) = 67.91$, $MSe = 0.23$, $p < .0001$, $\eta^2 = .58$] which indicated that RT decreased across tests [$M_{T1} = 0.57$, $SE_{T1} = 0.044$; $M_{T2} = 0.24$, $SE_{T2} = 0.027$; $M_{T3} = 0.086$, $SE_{T3} = 0.034$; $M_{T4} = -0.060$, $SE_{T4} = 0.031$]. The effect of Response Type trended toward significance [$F(1,50) = 3.62$,

MSe = 0.21, p = .063, η^2 = .067] which indicated that hits were generally faster than correct rejections [M_H = 0.24, SE_H = 0.026; M_{CR} = 0.18, SE_{CR} = 0.024]. Turning to the interactions, there was a significant Feedback Group × Response Type interaction [F(1,50) = 13.07, MSe = 0.21, p < .001, η^2 = .21]. This interaction indicated that while hits and correct rejections were equally as fast for the Liberal group [M_H = 0.18, SE_H = 0.038; M_{CR} = 0.23, SE_{CR} = 0.034; p = .62, Tukey's HSD], correct rejections were actually faster than hits for the Conservative group [M_H = 0.31, SE_H = 0.038; M_{CR} = 0.13, SE_{CR} = 0.034; p < .01, Tukey's HSD]. Finally, there was a significant Stimulus × Response Type interaction [F(1,50) = 7.68, MSe = 0.094, p < .01, η^2 = .13]. This interaction indicated that while hits and correct rejections were equally as fast for word stimuli [M_H = -0.14, SE_H = 0.018; M_{CR} = -0.14, SE_{CR} = 0.018], hits were slower than correct rejections for face stimuli [M_H = 0.62, SE_H = 0.051; M_{CR} = 0.50, SE_{CR} = 0.047], suggesting that these face stimuli were easier to correctly reject as new than to correctly endorse as familiar. No other main effects or interactions were significant (all p's > .09).

Experiment 3

Reaction Time

As with confidence, analyses of reaction times were restricted to correct reports. RTs were first converted to z-scores for each individual, then sorted by test and response type (e.g., hits and correct rejections). RT across the first three tests was examined using a 2 × 2 × 3 mixed ANOVA examining a single between-subjects factor of Feedback Group (Liberal vs. Conservative) and two within-subjects factors, Response Type (Hit vs. CR) and Test (Test 1 vs. Test 2 vs. Test 3). This analysis revealed a significant main effect of Response Type [F(1,111) = 13.32, MSe = 0.076, p < .001, η^2 = .11] which indicated that hits were generally faster than correct rejections [M_H = -0.067, SE_H = 0.013; M_{CR} = 0.010, SE_{CR} = 0.012]. There was also a

significant main effect of Test [$F(2,222) = 103.37$, $MSe = 0.088$, $p < .0001$, $\eta^2 = .48$] which indicated that RT generally decreased across tests [$M_{T1} = 0.19$, $SE_{T1} = 0.021$; $M_{T2} = -0.088$, $SE_{T2} = 0.012$; $M_{T3} = -0.19$, $SE_{T3} = 0.018$]. There was a robust Feedback Group × Response Type interaction [$F(1,111) = 32.86$, $MSe = 0.076$, $p < .0001$, $\eta^2 = .23$] This crossover interaction indicated that hits were significantly faster than correct rejections for the Liberal group [$M_H = -0.12$, $SE_H = 0.019$; $M_{CR} = 0.079$, $SE_{CR} = 0.018$; $p < .001$, Tukey's HSD], while hits were numerically slower than correct rejections for the Conservative group [$M_H = -0.014$, $SE_H = 0.018$; $M_{CR} = -0.058$, $SE_{CR} = 0.017$; $p = .45$, Tukey's HSD]. There was a significant Response Type × Test interaction [$F(2,222) = 6.54$, $MSe = 0.020$, $p < .01$, $\eta^2 = .056$] This interaction indicated that while both hits and correct rejections tended to get faster across tests, the RT advantage for hits increased across tests [$M_{HT1} = 0.18$, $SE_{HT1} = 0.026$; $M_{CRT1} = 0.21$, $SE_{CRT1} = 0.022$; $M_{HT2} = -0.14$, $SE_{HT2} = 0.018$; $M_{CRT2} = -0.035$, $SE_{CRT2} = 0.018$; $M_{HT3} = -0.25$, $SE_{HT3} = 0.022$; $M_{CRT3} = -0.14$, $SE_{CRT3} = 0.025$]. Finally, the three-way interaction amongst the factors was significant [$F(2,222) = 10.74$, $MSe = 0.020$, $p < .0001$, $\eta^2 = .088$]. This three-way interaction indicated that while hits were always faster than correct rejections for the Liberal group, correct rejections were equally as fast as hits across all three tests for the Conservative group.

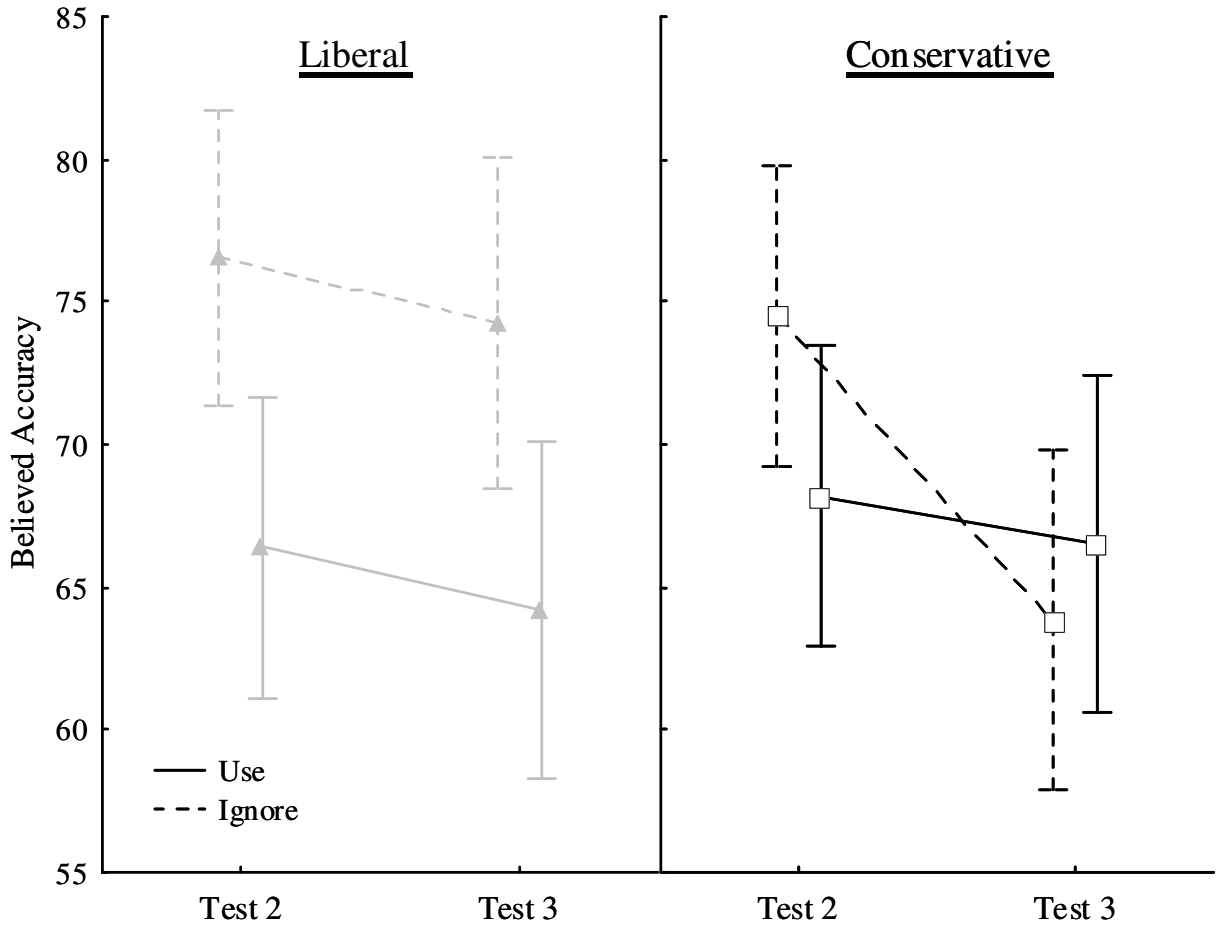
Only the reaction times for the No Shift groups from Test 4 were analyzed as RTs were not collected from the participants completing a paper version of the final test. RTs from the No Shift groups were analyzed using a 2 × 2 mixed ANOVA examining a between-subjects factor of Prior Feedback (Liberal vs. Conservative) and a within-subjects factor of Response Type (Hit vs. CR). The effect of Prior Feedback approached significance [$F(1,55) = 3.36$, $MSe = 0.033$, $p = .072$, $\eta^2 = .057$] which indicated that the Prior Liberal group was slightly faster than the Prior Conservative group [$M_{PL} = -0.0098$, $SE_{PL} = 0.024$; $M_{PC} = 0.052$, $SE_{PC} = 0.024$]. The two-way

interaction was significant [$F(1,55) = 8.76$, $MSe = 0.073$, $p < .01$, $\eta^2 = .14$]. This interaction indicated that while hits were numerically faster than correct rejections for the Prior Liberal group [$M_H = -0.055$, $SE_H = 0.053$; $M_{CR} = 0.035$, $SE_{CR} = 0.031$; $p = .60$, Tukey's HSD], correct rejections were significantly faster than hits for the Prior Conservative group [$M_H = 0.16$, $SE_H = 0.052$; $M_{CR} = -0.052$, $SE_{CR} = 0.030$; $p < .05$, Tukey's HSD].

Discussion of RT Effects

In general, the RT findings are consistent with the criterion findings in each experiment. That is, participants respond more quickly to the response that is consistent with their bias. This led to large differences in hit and correct rejection RT in the Liberal groups, and small differences in the Conservative groups since hits were typically faster than correct rejections initially. Only in experiment 2 were correct rejections faster than hits, and this effect may have been driven largely by reaction times to face stimuli. Given enough opportunities for reinforcement, it appears that correct rejections eventually overtake the initial response time advantage afforded to hits.

Interestingly, confidence and reaction time results for Experiment 2 diverged for faces. Specifically, confidence for faces was higher than confidence for words, but reaction times to faces were *slower* than reaction times to words; these effects occurred despite equivalent accuracy between the two stimulus types. This may reflect that faces led to more deliberative retrieval attempts than did words. The additional processing time afforded to faces likely produced a more confident report when issued.



Supplementary Figure 1: Three-way interaction between Test, Feedback Group, and Volition on Believed Accuracy (vertical bars denote 95% confidence intervals).

BIS/BAS	
BAS Drive	6.72 (2.17)
BAS Fun-Seeking	7.88 (2.42)
BAS Reward Responsiveness	12.24 (1.88)
BIS	16.88 (3.63)

RFQ	
Promotion Focus	21.99 (3.38)
Prevention Focus	17.37 (3.70)

GRAPES	
Reward Expectancy	9.08 (2.84)
Punishment Expectancy	7.88 (2.84)

Supplementary Table 1: Individual difference measures. Standard deviations in parentheses.

Supplementary Table 2

BIS/BAS (Han, 2009)	
BAS Drive	10.50 (2.03)
BAS Fun-Seeking	9.13 (1.67)
BAS Reward Responsiveness	17.50 (2.03)
BIS	21.13 (4.27)

Supplementary Table 2: BIS/BAS scores from Han, 2009 (standard deviations in parentheses). Presented for comparison.