Response Bias in Recognition Memory as a Stable Cognitive Trait

by

Justin David Kantner

B.A., Purdue University, 2000
M.A., Indiana University, 2005

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Psychology

Response Bias in Recognition Memory as a Stable Cognitive Trait

by

Justin David Kantner

B.A., Purdue University, 2000
M.A., Indiana University, 2005

Supervisory Committee

Dr. D. Stephen Lindsay, Supervisor
(Department of Psychology)

Dr. Michael E. J. Masson, Departmental Member
(Department of Psychology)

Dr. Catherine A. Mateer, Departmental Member
(Department of Psychology)

Dr. Neena L. Chappell, Outside Member
(Department of Sociology)

Supervisory Committee

Dr. D. Stephen Lindsay, Supervisor
(Department of Psychology)

Dr. Michael E. J. Masson, Departmental Member
(Department of Psychology)

Dr. Catherine A. Mateer, Departmental Member
(Department of Psychology)

Dr. Neena L. Chappell, Outside Member
(Department of Sociology)

Abstract

Recognition is the cognitive process by which we judge whether a given object, person, place, or event has occurred in our previous experience or is new to us. According to signal detection theory, old/new recognition decisions are based on how much evidence one finds in memory that an item has appeared previously (e.g., its familiarity) but can be affected substantially by response bias, a general proclivity to respond “old” or “new.” When experimental conditions evoke a “conservative” response bias, participants will require a relatively high amount of memory evidence before calling an item “old” and will give a high proportion of “new” responses to both old and new items; when conditions promote a “liberal” bias, participants will relax their required level of memory evidence and will call a high proportion of both old and new items “old.”

Response bias is usually analyzed at a group level, but substantial individual differences in bias can underlie group means. These differences suggest that, independent

of any experimental manipulation, some people require more memory evidence than others before they are willing to call an item "old." The central motivation for the present work is the possibility that these individual differences are meaningful and reflect bias levels that inhere within individuals. Seven experiments were designed to test the hypothesis that response bias can be characterized as an intra-individually stable cognitive "trait" with an influence extending beyond recognition memory.

The present experiments are based on the expectation that if response bias is a cognitive trait, it should a) be consistent within an individual across time, to-be-recognized materials, and situations; b) generalize beyond recognition memory to other tasks involving binary decisions based on accumulated evidence; c) be associated with personality traits that represent one's willingness to take action based on limited information; and d) carry consequences for recognition in applied settings. The results indicated substantial within-individual bias consistency in two recognition tests separated by 10 minutes (Experiment 1) and a similar level of consistency when the two tests were separated by one week (Experiment 2). Bias was strongly correlated across the stimulus domains of words and paintings (Experiment 3) and words and faces (Experiment 7). Correlations remained significant across two ostensibly independent experiments differing markedly in context and materials and separated by an average of 2.5 weeks (Experiments 6 and 7). Recognition bias predicted frequency of false recall in the Deese-Roediger-McDermott (DRM) paradigm (Experiment 4) and false alarms in an eyewitness identification task (Experiment 7). No relationship was detected between bias and grain size in estimation from general knowledge (Experiment 2), risk avoidance through the use of report option on a trivia task (Experiments 4 and 5), or speed and accuracy on a

go-no go task (Experiment 6). Personality measures suggested relationships between response bias and need for cognition, maximizing versus satisficing tendencies, and regret proneness. Collectively, these findings support the idea that response bias as measured in recognition memory tasks is a partial function of stable individual differences that have broad significance for cognition.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First, and unquestionably foremost, I express my profound appreciation for the generosity and the guidance of Steve Lindsay, my supervisor, whose advice I sought at virtually every turn in carrying out the work of this dissertation. I simply could not have asked for a more caring, conscientious, and enthusiastic mentor during six fantastic years at UVic. O Captain, My Captain!

I thank my the members of my candidacy exam and doctoral committees, Steve, Mike Masson, Katy Mateer, and Neena Chappell, with whom it was a pleasure and an honor to work over the course of five years. I also greatly valued the participation and the insight of Andy Yonelinas, External Examiner on my doctoral committee.

I had the immeasurable benefit of working with a brilliant group of research assistants who made it possible to collect all of the data reported in this dissertation in roughly a year's time: Mayumi Okamoto, Priya Rosenberg, Sarah Kraeutner, Caitlin Malli, Jordy Freeman, Emily Cameron, and Graeme Austin.

Dave Hamilton showed me the fruits of a long and highly successful career as my officemate for several summers, suggested two of the personality measures used in my dissertation research, and made the introduction that led to my postdoctoral fellowship.

My parents, James and Patricia, and my brother, Joe, cheered me on through every stage of this endeavor, adding meaning to everything I did.

Finally, I thank my wife, Sarah, for her love and all that it allows.

Dedication

To Claude E. Kantner, who walked the road before me.

To Sarah E. Kantner, who walked the road with me.

To Sebastian P. Kantner, who has the road ahead of him.

Introduction

Our everyday experience of the world is filled with encounters of places, objects, people, and events that we have come across in the past as well as those that are new to us. Recognition is the cognitive process by which we judge whether a given encounter belongs to the former category or to the latter. Although judgments of recognition are often made quickly and with high accuracy (e.g. Brady, Konkle, Alvarez, & Oliva, 2008), the simplicity implied by the binary nature of the decision is deceptive. Laboratory experiments have revealed a range of systematic errors made by the recognition system in rendering simple "old" or "new" judgments to recently presented items (e.g., Roediger & McDermott, 1995), and the component processes underlying the recognition decision have been a matter of theoretical debate since the 1960s (for a review see Yonelinas & Parks, 2007).

In addition, there is broad consensus that the mnemonic elements of a recognition decision are supplemented by "response bias," a proclivity to respond "old" (or "new") that may be independent of memory per se. A liberal response bias is associated with a high proportion of "old" judgments and reflects an apparent reluctance to call items "new," while a conservative response bias is associated with a high proportion of "new" judgments and a reluctance to call items "old." Historically, interest in response bias has been peripheral to interest in recognition accuracy. That is, investigators have generally been more concerned with the conditions mediating the accuracy with which old items are discriminated from new items, and less concerned with the proportion of trials on which participants give "old" versus "new" responses per se. In the last several years, however, response bias has attracted increasing attention as an informative measure in its

own right, at least in terms of understanding recognition memory. The present research was motivated by the hypothesis that response bias indexes a more central component of cognitive processing than previously thought. The overarching purpose of the experiments described is to characterize response bias as a cognitive "trait" whose influence extends beyond recognition memory.

I begin with a discussion of recognition memory as it is studied in the laboratory, and then define response bias from the perspective of signal detection theory, the dominant framework used to describe performance in recognition memory tasks. I then summarize the properties of response bias revealed by research and identify a rarely examined aspect of response bias data: substantial individual differences. I argue that these individual differences may result from the fact that response bias is a cognitive trait, varying between individuals from more conservative to more liberal but intra-individually stable. I then review a collection of published findings consistent with the characterization of bias as a cognitive trait.

I next report seven experiments designed as multifaceted tests of the hypothesis of trait response bias. These experiments are organized by the following four themes. First, if response bias is a trait, it should be consistent within individuals across time, to-be-recognized materials, and situations. Second, from the perspective of signal detection theory, individual differences in response bias suggest that some people require more evidence of previous encounter than others before they will declare an item to be old. If required level of evidence is a cognitive trait, it should generalize beyond recognition memory to other tasks involving a binary decision based on accumulated evidence. Third, trait response bias might be associated with personality traits that represent one's

willingness to act versus withhold action (e.g., impulsivity). Fourth, if response bias is a general cognitive trait, it should carry important consequences for behavior in applied tasks.

Recognition Memory in the Laboratory

Studies of recognition typically employ a study-delay-test design; this design formed the basis for the recognition component of the experiments reported here. Participants begin by studying a list of materials (e.g., words, pictures, faces, shapes) presented one-at-a-time on a computer screen. A delay interval follows the presentation of the study list and separates the study and test phases. At test, the studied ("old") items are randomly intermixed with some number of non-studied ("new") items and are presented one at a time with instructions to respond "old" to any items that were on the study list and "new" to any items that were not. The materials are usually familiar to participants from extra-experimental sources (e.g., words), so the recognition judgment is not one of whether the test probe has ever been encountered, but whether it was encountered during the study phase. This old/new judgment is sometimes coupled with a measure of confidence in each decision, as will be the case in each of the experiments reported below.

Signal Detection Theory and Recognition Memory

As applied to recognition memory (Parks, 1966), signal detection theory holds that every item we encounter falls at some point along a continuum of "evidence" of prior encounter in the designated study context. To take an example from a natural setting, one may see a rare and exotic species of flower that one has never seen before and that bears

little resemblance to any flower one has seen in the past; in this instance, there is very little evidence in memory to suggest that the species of flower is recognized. By contrast, there may be substantial memory evidence to suggest that a more commonplace flower such as a tulip has been seen before. In the context of recognition memory, the term "familiarity" is often used as a proxy for memory evidence; thus, the exotic flower engenders a minimal sense of familiarity and is unlikely to be recognized while the tulip is quite familiar and relatively likely to be recognized.

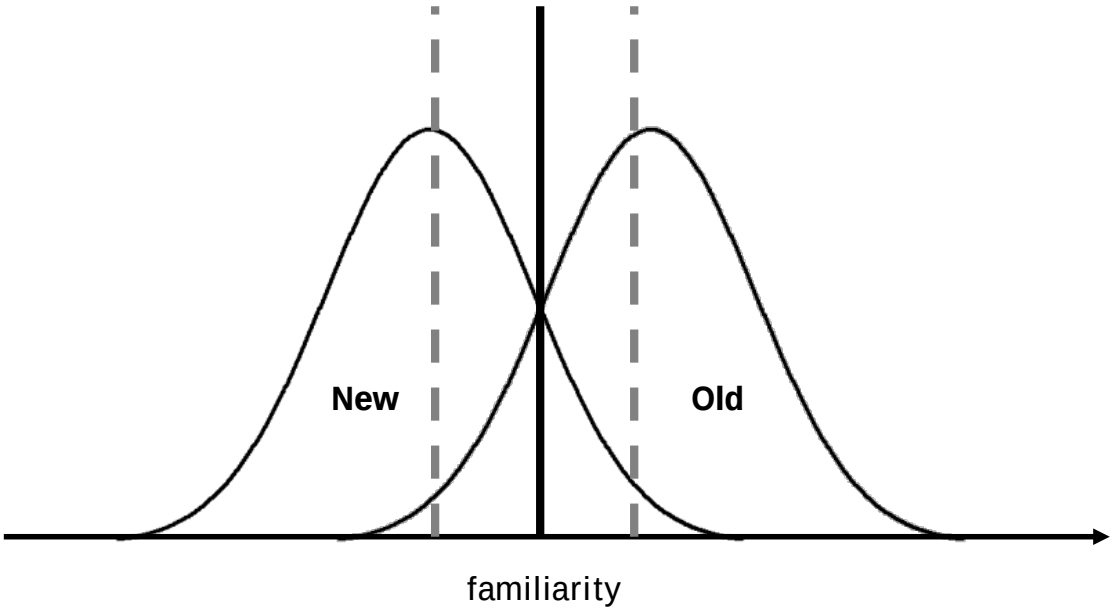


Figure 1. Illustration of signal detection model of recognition.

Signal detection theory assumes that old items will generally be more familiar than new items, and that the familiarity of each is normally distributed. The simplest version of the theory, in which the variance of the old and new distributions is equal, is illustrated in Figure 1. Critically, the distributions typically overlap to some degree along a central region of the familiarity continuum. This overlap reflects the fact that new items may be moderately familiar despite their newness (e.g., because they are similar to some

old items) and that some old items may *only* be moderately familiar despite their oldness (e.g., because they were not well encoded during the study phase). Thus, some new items may be as familiar as or more familiar than some old items. Familiarity, then, is not alone sufficient to make an accurate old/new decision for all items, especially those of neither high nor low familiarity. Signal detection theory assumes that individuals make judgments according to a decision criterion: a point along the familiarity continuum below which an item will be judged "new" and above which an item will be judged "old." The use of a criterion does not improve the accuracy of recognition judgments (as will be discussed below), but it serves as a heuristic that allows a decision to be made on test trials on which memory evidence of oldness versus newness is ambiguous.

The bold vertical line in Figure 1 depicts a neutral decision criterion, one that lies at the midpoint of the old and new distributions. A participant using this criterion will give "old" and "new" responses equally often across the course of a recognition test. The end of the old-item distribution falling below the criterion represents items that were presented on the study list but are not sufficiently familiar at test to yield an "old" judgment. Such items will incorrectly be called "new," a "miss" in signal detection terminology. Similarly, a portion of the new-item distribution containing the most familiar non-studied items will surpass the decision criterion and will incorrectly be called "old" (a "false alarm"). Because most of the old items surpass the criterion while most of the new items fail to surpass it, the majority of old items will correctly be called "old" (a "hit") and the majority of new items will correctly be called "new" (a "correct rejection"). By convention, and in present work, recognition accuracy is described and calculated in terms of the hit and false alarm rates. The correction rejection and miss rates

are simply (1 - false alarm rate) and (1 - hit rate), respectively, and thus are not needed to evaluate performance.

Criterion placement need not be neutral: one may set a low ("liberal") criterion, such that very little familiarity is required of an item before it will be called old (see the dashed gray line to the left of the neutral criterion). Because more of the old-item distribution lies above a liberal criterion than a neutral criterion, and the hit rate of a subject using a liberal criterion will be higher than that of a subject using a neutral criterion. However, because more of the new-item distribution also surpasses the criterion, false alarm rates will be higher with a liberal criterion. Alternatively, one may set a high, or "conservative," criterion, such that items will not be called "old" unless they elicit a very strong feeling of familiarity (see the dashed gray line to the right of the neutral criterion), yielding fewer false alarms and fewer hits than a neutral criterion. Importantly, differences in criterion placement are not associated with differences in recognition accuracy. Rather, they represent alternative approaches to the task: a liberal criterion sacrifices a low false alarm rate for a high hit rate and a conservative criterion sacrifices a high hit rate for a low false alarm rate. In the recognition literature, these approaches are captured by the terms *liberal* and *conservative response bias*, respectively.

Properties of Response Bias

Early work on the application of signal detection theory to recognition memory found evidence for the proposed role of a response criterion by demonstrating that its location follows in predictable ways from various task manipulations. Perhaps the simplest of these is instructional motivation. If the test instructions encourage participants

to be highly confident before they endorse an item as old or to respond new whenever uncertain, response bias is conservative (e.g., Egan, 1958). Revealing to participants the proportion of old items in the test also readily influences bias: if participants know a priori that 80% of test items will be old, they will err on the side of an "old" response when unsure, adopting a liberal bias (e.g., Parks, 1966; Van Zandt, 2000). An imbalance in the proportion of old and new items may also be learned during the course of a test if corrective trial-by-trial feedback is administered, resulting in base-rate-appropriate criterion setting (Kantner & Lindsay, 2010; Rhodes & Jacoby, 2007; Titus, 1973). Payoff schedules that encourage old or new responses (e.g., a gain of one dollar for every hit coupled with a loss of only ten cents for every false alarm) yield similar results (Healy & Kubovy, 1978, Van Zandt, 2000).

Subsequent studies have shown that criterion placement is also influenced by the materials used. For example, when the stimuli to be recognized are perceived to be particularly memorable, either through manipulations designed to increase memory strength such as repetition (e.g., Hirshman, 1995) or by virtue of their distinctiveness or infrequency (e.g., Brown, Lewis, & Monk, 1977; Wixted, 1992), participants adopt a more conservative response bias, apparently because they expect to remember studied items well, such that a moderate level of familiarity at test does not suffice to elicit an "old" judgment (e.g, Brown et al., 1977). In addition, a substantial literature indicates that emotionally arousing words (e.g., "horror") elicit a more liberal response bias than emotionally neutral words (e.g., Budson, Todman, Chong, Adams, Kensinger, Krangel, & Wright, 2006; Dougal & Rotello, 2007). The same liberal bias effect holds for emotionally arousing faces (Johannson, Mecklinger, & Treese, 2004) and bizarre actions

(e.g., Worthen & Wood, 2001). One explanation for this effect is that the arousal induced by these items at test may be misattributed to familiarity.

Despite the centrality of the response criterion to the signal detection account of recognition and accumulating evidence concerning experimental manipulations that affect its location, fundamental questions such as how a criterion is established and under what circumstances it changes have been slower to receive attention (Dobbins & Han, 2008; Estes & Maddox, 1995; Whittlesea, 2002). Thus, recent research has assessed the ability of subjects to change criterion over the course of an experiment in response to task manipulations (Hockley, 2011). One such manipulation is a change of difficulty during the test. Benjamin and Bawa (2004), for example, found that when the similarity of new items to old items was increased partway through the test, participants adjusted to a more conservative criterion (in order to avoid an increase in false alarms). Brown, Steyvers, and Hemmer (2007) changed target-lure similarity from low to high several times during test and determined that subjects can toggle between more liberal and more conservative levels of bias in an adaptive manner, though the timing of the shifts was estimated to be an average of three trials behind the point of change.

Other studies tested two classes of items, one with high memory strength (e.g., presented 5 times during the study phase) and one with low memory strength (e.g., presented just once), in order to determine whether subjects can apply a more conservative criterion when judging items of the strong class and a more liberal criterion when judging items of the weak class. Although results have been equivocal (e.g., Morrell, Gaitan, & Wixted, 2002; Stretch & Wixted, 1998; Verde & Rotello, 2007), evidence suggests that subjects can make such criterion shifts, at least under some

conditions, when the two item classes are tested in different lists (e.g., Hirshman, 1995) and even when they are mixed within a single test list (e.g., Lindsay & Kantner, 2010; Singer, 2009).

Is Response Bias an Intra-individually Stable Cognitive Trait?

The preceding review highlights some of the major themes of research on response bias and exemplifies an accelerating interest in understanding and modeling patterns of bias under various experimental conditions (Hockley, 2011; Rotello & Macmillan, 2008). The present experiments were designed to examine response bias from a different perspective. The principle objective is to ask whether bias is strictly a function of prevailing experimental conditions or inheres to a degree in an individual recognizer as a cognitive trait.

The departure represented by the present work is captured by the point that all of the above lines of research involve the analysis of bias at a group level. For example, one can conclude that informing participants of a high base rate of old items results in a liberal response criterion because the mean response bias of the informed group is liberal while the mean bias of an uninformed group is neutral. Based on numerous recognition experiments conducted in our laboratory, however, substantial individual differences in bias often underlie group means. Figure 2 illustrates an example of this phenomenon from Experiment 1 of Kantner and Lindsay (2010), which involved a standard recognition task. In the control condition of that experiment (N = 23), the mean response bias was statistically neutral, represented by the bold line near the meeting point of the old and new distributions. A plot of the criteria used by each of the 23 individual participants (represented by the gray lines) reveals that considerable variability

characterized the bias scores that compose the mean. While an unbiased criterion was used across subjects, a number of individual subjects were either liberally or conservatively biased.

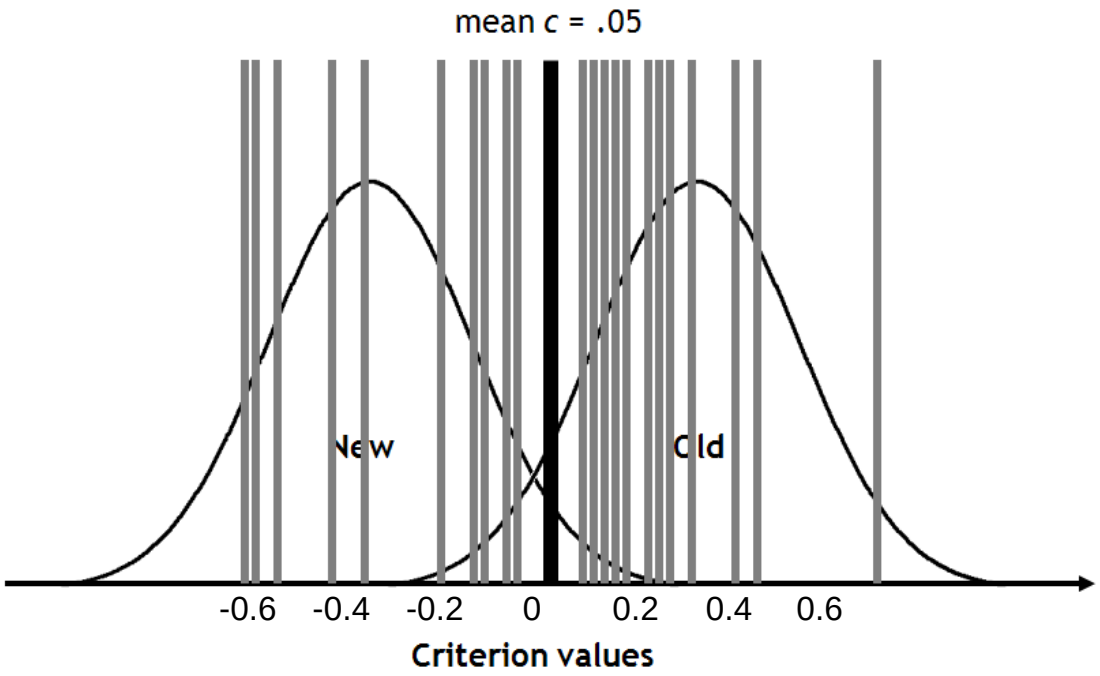


Figure 2. The spread of individual criterion values in Kantner and Lindsay (2010), Experiment 1, control condition.

A central motivation for the current experiments is the possibility that this variability is meaningful (i.e., not merely the result of measurement error) and reflects bias proclivities within individuals that are independent of the parameters of the recognition task. From a signal detection theory perspective, the spread of criterion values represented in Figure 2 suggests that some participants require more evidence of oldness than others before they will make an "old" judgment. To take the most extreme example from Figure 2, consider the leftmost and the rightmost individual criteria. These two participants achieved similar levels of accuracy on the test, but through two very different means: one was highly liberal, accepting items of even modest familiarity as old

and thus maximizing hits, while the other was highly conservative, requiring a high degree of familiarity before calling items “old” and minimizing false alarms. An intriguing possibility is that these two participants did not merely happen to respond in this manner on this particular test but that they are *generally* liberally and conservatively biased recognizers, respectively; that is, the level of memory evidence they require before committing to an "old” decision is a stable trait, and the bias they display on a recognition test a manifestation of that trait. Evidence of trait-like stability would suggest an entirely different component to response bias than that studied by examining its reaction to task variables and would raise questions as to the cognitive and behavioral consequences of such a trait. The experiments reported here constitute an initial investigation into these issues.

Before describing evidence from the recognition literature suggesting trait-like attributes of response bias, an alternative characterization of apparent criterion variability is worth noting. It might be the case that all participants have an equivalent criterion but that some participants experience more familiarity in response to both old and new items than do others. A participant given to experiencing a small amount of familiarity at test would be expected to give a larger proportion of “new” responses than one that experiences a great deal of familiarity with both old and new items, even if their response criteria are equal. Although this account cannot be ruled out by signal detection theory, it is unappealing because it is not clear why participants would vary dramatically in familiarity with new items (discussed further in the General Discussion). In contrast, the notion of individual differences in response bias has substantial theoretical appeal.

Previous Evidence Suggestive of Trait Bias

Although response bias is not generally characterized as representing a trait in the recognition literature, some past research appears to have been motivated implicitly by the possibility. A substantial number of studies have examined the relationship of response bias to a range of neural and behavioral pathologies. The consistency in the central result of such studies is compelling. Response bias is found to be more liberal for elderly individuals (Harkins, Chapman, & Eisdorfer, 1979; Huh, Kramer, Gazzaley, & Delis, 2006; Trahan, Larrabee, & Levin, 1986; though see Gordon & Clark, 1974), patients with Alzheimer's disease (Beth, Budson, Waring, & Ally, 2009; Gold, Marchant, Koutstall, Schacter, & Budson, 2007; Snodgrass & Corwin, 1988), patients with dementia (Woodard, Axelrod, Mordecai, & Shannon, 2004; Snodgrass & Corwin, 1988), individuals with mental retardation (Carlin, Toglia, Wakeford, Jakway, Sullivan, & Hasel, 2008), patients with schizophrenia (Moritz, Woodward, Jelinek, & Klinge, 2008), and individuals with panic disorder (Windmann & Kruger, 1998) compared to appropriate controls. A frequent explanation for such effects is that they arise from damage to or deterioration of prefrontal cortex (PFC; Gold et al., 2007; Huh et al., 2006; Windmann, Urbach, & Kutas, 2002), which is associated with planning and control of responses and is assumed to be involved in criterion setting. This connection of the PFC and response bias suggests that variability in PFC function may help explain variability in bias across individuals (e.g., Kramer, Rosen, Du, Schuff, Hollnagel, Weiner, Miller, & Delis, 2005). More generally, the association of liberal response bias and the above conditions is consistent with the idea that groups of individuals may be differentiated from one another on the basis of response bias without a specific experimental

intervention. This idea is highly consistent with the notion of response bias as a stable cognitive trait.

More closely related to the goals of the proposed experiments are a small number of studies examining the correlation of response bias and cognitive or personality traits within an individual; significant relationships between bias and established traits suggest that bias also possesses trait-like qualities. Following the theory of frontal region involvement in criterion setting, Huh et al. (2006) correlated response bias on a recognition test with performance on four measures of executive function from the Delis-Kaplan Executive Function System (Delis, Kramer, & Kaplan, 2001) associated with the frontal lobe: inhibition (via a Stroop task), concept formation (via a card sorting task), set shifting (via the trail making test) and verbal fluency (via a word generation task). Inhibition was the only significant predictor of response bias (r = .31), and some of the executive function measures were not statistically related to each other, leading Huh et al. to declare the analysis inconclusive.

In a 30-year-old study that might constitute the published work most relevant to the current experiments (cited just twice according to Web of Science), Gillespie and Eysenck (1980) investigated response bias in introverts and extraverts using a continuous recognition task. Introverts were found to use a more conservative response criterion than extraverts and were described as exercising greater “response cautiousness.” This result and characterization of the conservative recognizers are wholly consistent with the notion of response bias as the manifestation of a cognitive trait: introverts are expected to exercise greater caution than extraverts (Patterson & Newman, 1993), leading them to require more evidence before committing to an “old” response in a recognition task.

Response bias, then, may arise from a stable trait corresponding to a required level of evidence before action is taken, a trait that, like introversion/extraversion, is stable within an individual and generalizes to tasks and situations beyond recognition memory.

While few published studies have approached response bias as a potential trait, a greater number have investigated individual differences in false memory proneness via the Deese-Roediger-McDermott (DRM) paradigm (Deese, 1959; Roediger & McDermott, 1995). In the DRM paradigm, participants read lists of words that are semantically united (e.g., nurse, patient, hospital, surgeon, medicine) but do not include a critical, highly related associate (e.g., doctor). These critical lures are readily brought to mind by true list members, creating a compelling sense at test that they, too, were part of the list. Participants falsely recall or recognize the critical lure at rates sometimes meeting or exceeding accurate recall/recognition of presented list items. Given that a liberal recognition bias is associated with increased endorsement of test probes that were not studied, evidence that DRM false recognition has trait-like qualities could suggest the same for response bias (e.g., Miller & Wolford, 1999; Miller, Guerin, & Wolford, in press).

DRM performance has been correlated with a number of individual difference measures (e.g., age, working memory, frequency of dissociative experiences; for a review see Gallo, 2010). Some experiments have identified populations with particularly high rates of DRM errors: individuals reporting recovered memories of childhood abuse (Clancy, Schacter, McNally, & Pittman, 2000), individuals reporting having recovered such memories through therapy (as opposed to spontaneously; Geraerts, Lindsay, Merckelbach, Jelicic, Raymaekers, Arnold, & Schooler, 2009), and individuals claiming

memories of alien abduction (Clancy, McNally, Schacter, Lenzenweger, & Pitman, 2002) or past lives (Meyersburg, Bogdan, Gallo, & McNally, 2009). Each of these findings suggests that some individuals are inherently more prone than others to accept memories as true even when memory evidence is weak, making them especially vulnerable to false memories.

Two studies have assessed the within-individual stability of DRM false recognition. Salthouse and Siedlecki (2007) found reliable stability within a single test but not across separate tests differing in stimulus type, and false recognition of critical lures was uncorrelated with a host of cognitive and personality measures in two experiments. However, Blair, Lenton, and Hastie (2002) found high levels of reliability in tests of the same DRM lists given two weeks apart, indicating that false recognition does not vary unpredictably within an individual.

Two further findings from the DRM literature are suggestive with respect to trait response bias. Although Blair et al. (2002) were interested in the stability of false DRM recognition independent of response bias, they reported a significant correlation of critical and non-critical false alarms during the first test (but a non-significant correlation during the second), a result that hints at a relationship between general recognition bias and DRM false memories. Relatedly, Qin, Ogle, and Goodman (2008) did not find evidence for a hypothesized relationship between DRM errors and susceptibility to adopting fictitious childhood events as autobiographical, but response bias calculated from the non-critical DRM trials was significantly (if weakly) predictive of such susceptibility. These results are consistent with the possibility that response bias might

generalize to tasks outside of recognition memory, a facet of trait-like stability tested in several of the current experiments.

Current Experiments: Measurement of Response Bias

The measurement of response bias raises complex theoretical and statistical issues relevant to any recognition memory experiment. This complexity arises from the fact that response bias must be estimated from patterns of recognition test responses, and the optimal method of estimation has been a matter of extensive debate (see Rotello & Macmillan, 2008). There are many methods for calculating bias, and each is tied to certain theoretical assumptions that may or may not hold true for a given dataset.

The estimate used in the current work is *c* (Macmillan, 1993), a simple and widely-used measure given as

$$-(z[H] + z[FA])/2 \qquad (1)$$

where H is a participant's hit rate and FA is the false alarm rate. The conversion of both values to a z-score reflects the classical signal detection model assumption of standard normal old- and new-item distributions along the evidence continuum (depicted in Figure 1). Despite its popularity in the recognition literature, the *c* measure is not without shortcomings. Two primary concerns and the means taken to address them in the analyses of the current experiments are discussed below.

Equal variance of the old- and new-item distributions. In addition to the assumption of normal distributions noted above, *c* assumes that the two distributions have equal variance. Evidence from Receiver Operating Characteristic (ROC) curves, functions relating hit rates to false alarms rates across several potential response criteria inferred from confidence ratings, has shed light on both of these assumptions (Yonelinas

& Parks, 2007). While the shape of the distributions does appear to be approximately Gaussian in item recognition tasks such as those used in the current experiments (Yonelinas & Parks, 2007), there is broad consensus that the variances of the distributions are unequal. A rule of thumb is that the variance of the new-item distribution is about 80% of that of the old-item distribution (Ratcliff, Sheu, & Gronlund, 1992), a regularity often explained in terms of encoding variability: while new test items possess only background variance in familiarity (i.e., through exposure in everyday life), old test items possess background variance in addition to variability in strength of encoding when presented at study (Wixted, 2007). When the equal variance assumption is violated, *c* will misrepresent the proportion of the distributions falling to the right of the criterion, leading to an inaccurate estimate of bias.

A more accurate but less wieldy alternative to the *c* measure is c_a, which produces an estimate of response criterion at each level of confidence that takes into account the relative variances of the old-item and new-item distributions (Macmillan & Creelman, 2005). The accuracy of *c* can be robust with respect to violations of the equal variance assumption, however (e.g., Curran, DeBuse, & Leynes, 2007). When the two variances truly are equal, *c* will be equivalent to c_a at the middle (neutral) confidence level; to the extent that the variances differ, *c* will deviate from middle c_a. To gain a sense of the accuracy of *c* in the current experiments, both *c* and c_a were calculated for all participants in Experiment 7, an experiment correlating bias across two highly distinct stimulus domains (faces and words) and possessing a large sample size ($N = 74$). The correlation of *c* and middle c_a in this dataset was extremely high ($r = 0.97$), and the observed bias

correlation only differed by 0.02 across the two measures. Therefore, *c* was retained as the bias measure of choice in all of the present experiments.

Independence of response bias and sensitivity. The two components held by signal detection theory to underlie recognition judgments, response bias and sensitivity in the ability to discriminate old items and new (a proxy for accuracy in a recognition task), are, in theory, representatives of independent psychological processes. The measures used to index bias and sensitivity, however, are usually not fully independent (e.g., Wixted, 2007). The most common measure of sensitivity, and the one used in the present analyses (*d'*), is calculated as

$$z(\mathrm{H}) - z(\mathrm{FA}) \qquad (2)$$

or the distance between the centers of the old-item and new-item distributions (Green & Swets, 1966). That both *d'* and *c* are calculated from hit and false alarm rates can blur their separability at the interpretive stage. Imagine, for example, that a participant completes two recognition tests with the following hit and false alarm rates: H = .74, FA = .28 (Test 1); H = .84, FA = .28 (Test 2). The increase in the hit rate, coupled with an unchanging false alarm rate, produces changes in both *d'* (from 1.23 to 1.58) and *c* (from -0.03 to -0.21), leading to the conclusion that sensitivity has increased across tests while bias has become more liberal.

Unfortunately, signal detection theory can be used to model an increased hit rate and consistent false alarm rate in any number of ways, and does not require that both sensitivity and bias have changed. For example, a shift of the old-item distribution farther to the right along the familiarity axis in Test 2 than in Test 1 (perhaps because the participant devised a more effective strategy for studying items on Test 2 than was used

on Test 1) coupled with an unmoved new-item distribution (i.e., because the background familiarity of items not on the study list is unchanged across tests) would predict the observed pattern of hit and false alarm rates in the absence of any change in bias across the tests. Under these circumstances, the liberal shift in the *c* parameter would be misleading.

Because sensitivity and bias cannot be completely decoupled in certain patterns of hit and false alarm data, the most straightforward method for checking their co-dependence is to determine their statistical independence. Therefore, in each of the experiments including two recognition tests, a correlation was calculated between each participant's *d'* and *c* scores. Where no relationship was apparent, it was assumed that *d'* and *c* were measuring essentially independent components of the recognition decision. When small correlations were present, partial correlations of *c* on two different tests were used to control for the influence of *d'*. In many such cases, correlations of *d'* and *c* were driven by the tendency for both measures to taken on extreme values as hit rates approach 1 or false alarm rates approach 0. Partial bias correlations generally approximated the corresponding full correlations.

Experiment 1

If response bias represents a cognitive trait, it should remain consistent within an individual across time. Therefore, an important first step in establishing response bias as trait-like is to determine whether a given subject will show the same level of bias on two different recognition tests. Experiment 1 was designed to test this possibility in a straightforward manner: each subject took two recognition tests (each preceded by its

own study list) that were separated by a filled 10-minute interval. The measure of interest was the correlation between bias on Test 1 and bias on Test 2.

Method

Participants. In each of the present experiments, University of Victoria students participated for optional bonus credit in an undergraduate psychology course. The vast majority of participants were 18-24 years old and approximately 70% were female. English was a second language for some, but such participants' data were only withheld from analysis on the rare occasion that a lack of fluency in English precluded full comprehension of instructions or verbal stimulus materials.

There were 41 participants in Experiment 1.

Materials. The stimuli were 192 4- to 8-letter medium- to high-frequency English nouns drawn from the MRC psycholinguistic database (http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm; Coltheart, 1981). Study and test lists were created via random selection from the 192-word pool for each participant. Sample words appear in Appendix A. Forty-eight randomly selected words composed Study List 1. Test List 1 consisted of the 48 words from Study List 1 and 48 non-studied words. Study List 2 contained 48 words not included in Study List 1 or Test List 1, and Test List 2 consisted of the 48 words from Study List 2 plus 48 words presented at no earlier point in the experiment. Three primacy and 3 recency buffers (not included in the pool of 192 words but adhering to the same specifications) were included in each study list. Thus, each study list contained 54 words and each test list contained 96 words. Study and test lists were presented in a randomized order. Stimuli appeared in the center of a computer screen and were black against a white background. All of the current

experiments were conducted with E-Prime experimental software (Psychology Software Tools, http://www.pstnet.com).

Procedure. Unless stated otherwise, all participants in the present experiments were tested individually with an experimenter present throughout the session. Participants were informed that they would first view a list of words one at a time and that the task was to try to memorize each word as well as possible for a subsequent memory test. Study items were presented for 1 s each and a blank 1-s interstimulus interval (ISI) separated the items.

Upon completion of the study list, participants received memory test instructions informing them that they would see another list of words, that some of these words had appeared in the preceding study list and some had not, and that their task was to indicate whether or not each item had been studied. Recognition judgments were made on a six-point, confidence-graded scale (1 = Definitely Not Studied, 2 = Probably Not Studied, 3 = Maybe Not Studied, 4 = Maybe Studied, 5 = Probably Studied, 6 = Definitely Studied). Each test word appeared in the center of the screen with the response scale centered beneath it. Responses were non-speeded. Both the word and the scale remained on the screen until a response was made via key press. Entry of the response triggered a 1-s intertrial interval (ITI) during which only the response scale remained on the screen.

At the end of the test, participants were given a sheet of paper and a pen and were asked to spend 8 minutes writing down the names of as many countries as they could. Participants occasionally commented ahead of the 8-minute deadline that their productivity had stalled; these participants were encouraged to continue working on the task for the rest of the allotted time in the event that a new country might spring to mind.

The 8-minute duration of the task was intended to combine with the brief instructional period to follow in forming an approximately ten-minute interval between the first and second recognition study/test cycles.

The procedure for the second study/test cycle was identical to that of the first, with the exception of some instructional modifications. Study instructions emphasized that although the task was the same as it had been during the first half of the experiment, all of the words to be presented would be new (i.e., none would be repeated from earlier phases of the experiment). Test instructions similarly emphasized that no words from the first half of the experiment would appear in the second test, and, consequently, that one only needed consider the immediately preceding study list in determining whether a given item had been studied.

Results and Discussion

In this and each subsequent experiment, recognition rating data were converted to hits and false alarms by scoring responses of 4, 5, or 6 as hits for old items and as false alarms for new items. Occasional participant false alarm rates of 0 were replaced with $0.5/n$, where n is the number of new test items; hit rates of 1 were replaced with $(1 - [0.5/n])$, where n is the number of old test items (Macmillan & Kaplan, 1985). The bias measure c is positive when response bias is conservative, negative when it is liberal, and close to zero when it is neutral. In general, the report of results for each experiment will begin with a summary of the group or condition means for the dependent measures of interest, followed by the critical measures of bias correlation across tests.

Table 1. Recognition means in Experiment 1.

	H		FA		*c*		*d'*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Test 1	.70	.13	.26	.13	.07	.38	1.25	.45
Test 2	.74	.17	.27	.17	-.02	.47	1.50	.78

Note. H = hit rate, FA = false alarm rate, M = mean, SD = standard deviation.

The mean hit (H) and false alarm (FA) rates and their corresponding sensitivity (*d'*) and bias (*c*) values for Tests 1 and 2 are displayed in Table 1. Mean recognition sensitivity increased significantly from Test 1 to Test 2, $t(40) = -2.48$, $p < .05$, driven by a moderate but significant rise in hit rates, $t(40) = -2.26$, $p < .05$. Mean false alarm rates were nearly identical across tests, $t = -.05$. Bias was roughly neutral and did not differ significantly between Test 1 and Test 2, $t(40) = 1.59$, $p = .12$.

The results of primary interest in the present experiments concern individual differences in response bias. As observed in numerous prior experiments from our laboratory (see, e.g., Figure 2), bias varied greatly at the level of the individual, ranging from extremely conservative to extremely liberal. The highest value of *c* in a single test was 1.10 (H = .44, FA = .02); the lowest value was -1.01 (H = .92, FA = .73). The question was whether these values were predictive of bias across the two recognition tests.

Unless otherwise stated, all correlations reported in this manuscript were calculated with Pearson's *r* statistic. Test 1 bias is plotted against Test 2 bias for each participant in Figure 3. As is clear from inspection of the figure, participants with a

liberal bias on Test 1 tended to be liberal on Test 2, those who were conservative on Test 1 tended to remain conservative on Test 2, and those who were essentially neutral on Test 1 remained essentially neutral on Test 2. Overall, there was a strong positive correlation between bias on the first and second tests, $r(39) = 0.69$, $p < .001$. While sensitivity was also strongly correlated across tests, $r(39) = 0.58$, $p < .001$, mean bias and mean sensitivity (averaged across the two tests) did not correlate with one another, $r(39) = 0.003$.

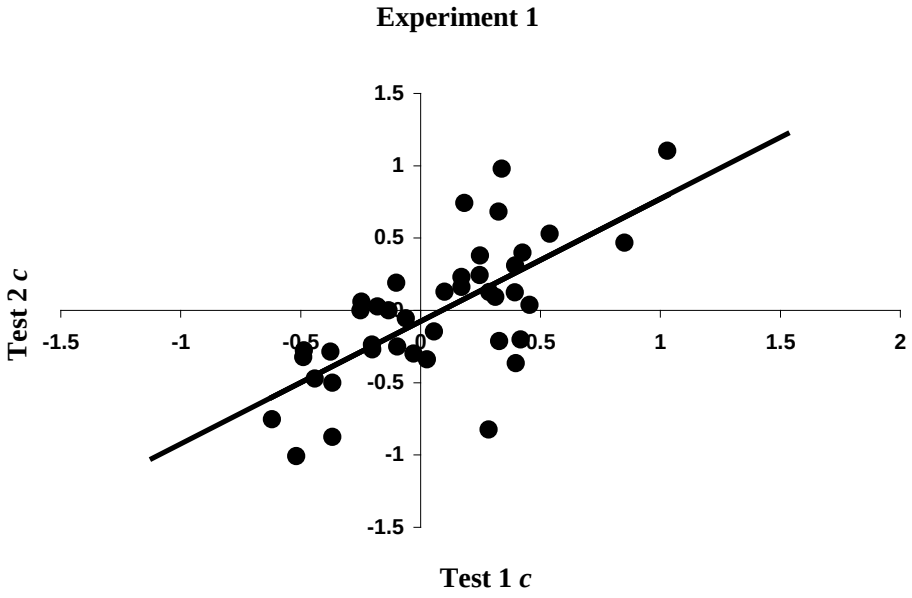


Figure 3. Correlation of recognition bias at Test 1 and Test 2 in Experiment 1.

To establish a benchmark against which to compare inter-test bias correlations, the split-half reliability of bias within a single test was measured. In theory (i.e., error variance notwithstanding), within-test reliability should index the strongest measurable bias correlation, and, therefore, represent the ceiling for bias correlations across tests. For each participant, test responses were divided randomly into halves, bias on each half was computed, and the correlation of bias across the two halves was calculated. Because this

analysis derives bias estimates from only half the trials of an inter-test correlational analysis, the procedure was performed on both Test 1 and Test 2. The within-test correlations were 0.69 and 0.78 on Tests 1 and 2, respectively, for a mean within-test correlation of 0.73. Thus, the level of stability in bias across tests in Experiment 1 was similar to that observed within a single test, an indication that a delay of 10 minutes and a separate study/test cycle had virtually no effect on participants' response bias.

The results of Experiment 1 demonstrate that compelling levels of inter-individual variability can characterize response bias in a recognition test despite the neutrality suggested by the group mean. Moreover, they provide support for the hypothesis that these individual differences are consistent across two recognition tests.

Experiment 2

The finding of bias consistency when 10 minutes separate two recognition tests provides important evidence that estimates of individuals' bias on a given recognition test, and the resulting variability in bias scores across participants, are not solely the result of measurement error. Experiment 2 was designed to provide a stronger test of lasting consistency in bias. As in Experiment 1, bias was correlated across two recognition tests using words as stimuli. In Experiment 2, however, the two tests were separated by one week.

The second goal of Experiment 2 was to investigate a second dimension of trait-like stability in response bias: its transfer to non-recognition memory tasks. The idea motivating such an investigation is that if response bias is the manifestation of an "evidence requirement" trait (as described above), it should correlate with performance on other tasks in which an evidence requirement might guide judgments.

This possibility was tested with two such tasks in Experiment 2. The first was a DRM list recall task (see p. 14). Given the decreased caution exercised by liberal recognizers in accepting words as having been encountered previously, the prediction was that such participants would be more likely to commit false recall of critical DRM lures than participants exhibiting a conservative recognition bias. That is, while liberal recognizers might recall DRM lures solely by virtue of the fact that they fit well with the list being recalled and feel familiar, conservative recognizers might be disposed to question whether the familiarity evoked by the critical lure is diagnostic of study list presence, and, in the absence of explicitly recollecting such presence, might be more likely to resist reporting that it was on the list.

The second non-recognition measure correlated with recognition bias in Experiment 2 was grain size in estimating answers to general knowledge questions (Goldsmith, Koriat, & Weinberg-Eliezer, 2002). Participants were asked questions to which they did not usually know the exact answers (e.g., “What year did CBC make its first television broadcast?”) and responded with numerical ranges that they believed were likely to contain the exact answer. Subjects could choose relatively fine-grained answers (e.g., “1950-1955”) or relatively coarse-grained answers (e.g., “1900-1970”). Fine-grained answers are less likely to be accurate but are more informative than coarse-grained answers. The grain size with which one answers a question is understood to reflect preference for accuracy or informativeness in responding (Ackerman & Goldsmith, 2008); most people over-emphasize informativeness and are highly inaccurate (Yaniv & Foster, 1995). We predicted that participants exhibiting a more conservative recognition bias would tend to use wider ranges than liberal recognizers, again on the

basis that recognition response bias is a reflection of a "required evidence" trait: conservative recognizers were hypothesized to require more evidence of their knowledge of a topic before committing to a narrow range answer.

Method

Participants. There were 46 participants in Experiment 2.

Materials. The stimuli used in the recognition portions of the experiment were identical to those used in Experiment 1. The stimuli used in the DRM task were the doctor, window, rough, bread, anger, sweet, couch, and smell lists from Stadler, Roediger, and McDermott (1999). These eight lists were chosen based on the following criteria: first, they did not include the sleep list, which was suspected to be well-known to participants pre-experimentally through classroom demonstrations of the DRM false memory effect; second, they did not include words that also appeared in the recognition portions of the experiment; and third, they were among the lists reported to elicit the highest rates of critical lure recall by Stadler et al. (1999). Each list contained 15 words in decreasing order of semantic relatedness to the category prototype, a structure thought to increase subsequent false recall of the critical lure (see Roediger & McDermott, 1995).

The general knowledge task included 50 trivia-style questions, each with an exact numerical answer, drawn from a set written by the author and a research assistant. The questions were designed such that current university undergraduates would be unlikely to know the exact answers but would possess enough relevant knowledge to provide, for each question, a numerical range they were confident contained the true answer. A typical undergraduate might, for example, be aware that Elvis Presley was famous in the latter half of the 20th century and died roughly a decade or two before they were born, but

not be able to recall that he died in 1977. A typical response to the question “What year did Elvis Presley die?” might then be “1965-1985.” Pilot testing indicated that the question set was effective in eliciting range estimates and that the variability in the size of the ranges used was suitable for testing hypotheses about individual differences in general knowledge-based estimation.

All questions selected for use in Experiment 3 called for answers in the form of specific years (as in the examples above). Each question began with the words “In what year” and referred to a historical, political, scientific, or pop cultural event from the last 200 years. The restricted historical range of the events queried was intended to reduce the occurrence of extremely large range sizes in the dataset; such outliers had occasionally exerted an undesirable level of influence on participant means in pilot testing.

Procedure. Participants took part in two sessions scheduled at the same time of day exactly one week apart. Session 1 consisted of a recognition study/test cycle and either the general knowledge or DRM task. Session 2 consisted of a second recognition study/test cycle (using different words from Session 1) and whichever of the general knowledge and DRM tasks was not included in Session 1. The assignment of the non-recognition tasks to Sessions 1 and 2 was random for each participant, as was the order of the two tasks within each session. There were no intervals between tasks beyond those needed to convey task instructions.

The procedure for the recognition phases was identical to that of Experiment 1. The procedure for the DRM and general knowledge tasks was as follows.

DRM task. Participants were informed that on each of a number of trials they would see a list of words presented one-at-a-time on the computer screen, that they were

to read each word aloud, and that they would subsequently be asked to write down as many words from the list as they could recall within a 2-minute time limit. The experimenter provided the participant with a pen and a stack of eight slips of paper for use in recalling the lists.

Each list was preceded by a screen encouraging participants to focus attention and hit a key when ready to proceed. Words were presented for 2 s each with a 1-s ISI separating the words. The final word on each list was followed by a screen containing the words “Recall List Now” that remained up throughout the recall period. A high tone sounded after 2 minutes to signify the end of the recall period, whereupon the participant placed the completed slip of paper on the bottom of the stack and proceeded to the next study list. The ordering of the eight lists was random for each participant.

General knowledge task. Participants were informed that they would be answering a series of questions related to history, government, science, and culture, and that each question required an answer in the form of a range of years. Instructions stated that participants were not expected to know the precise answers to many of the questions, and that in such cases they were to respond with a range of years within which they were “reasonably certain the event in question occurred, such that you would be comfortable giving this information to a friend if asked.” When participants did feel certain of a precise answer, they were to express that value as both the beginning the end of the range (e.g., “1958-1958”).

Each question was displayed near the top of the screen. Two boxes were positioned to the left and right of the center of the screen into which participants entered lower and upper range bounds, respectively. To enter either bound, the participant made a

mouse click inside of the corresponding box and typed the year in a response window. Participants were given as long as needed to enter responses, could enter the lower and upper bound in any order, and could edit responses once entered simply by clicking on a response box and entering a new value. Once both response boxes had been filled, a new box appeared at the bottom of the screen called “Enter Range.” Clicking this box initiated a 1-s ITI, followed by the next trial. A single practice trial preceded the 50 test trials.

Results and Discussion

The general knowledge task data from five participants were removed from the analyses reported below. In three of these cases, the participant was new to North America and lacked the knowledge base necessary to formulate ranges on a substantial proportion of questions. In one case, the participant was far older than the remainder of participants and had lived through an inordinate number of the events in question. In the final case, the participant gave several ranges beginning much earlier than 200 years ago despite instructions to the contrary. These participants’ DRM and recognition data were included in subsequent analyses.

Recognition test means are displayed in Table 2. Performance on the two tests was very similar: hit rates, false alarm rates, sensitivity, and bias were all statistically equivalent (all ts < 0.7). Mean bias across all participants was again approximately neutral.

Table 2. Recognition means in Experiment 2.

	H		FA		*C*		*d'*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Test 1	.71	.16	.28	.14	.01	.41	1.29	.59
Test 2	.70	.14	.28	.17	.06	.41	1.26	.66

Test 1 bias is plotted against Test 2 bias in Figure 4. The scatterplot reveals a pattern similar to that seen in Experiment 1. The correlation of bias across the two tests was again highly significant, $r(44) = 0.67$, $p < .001$. Bias and recognition sensitivity were again uncorrelated, $r(44) = -0.03$.

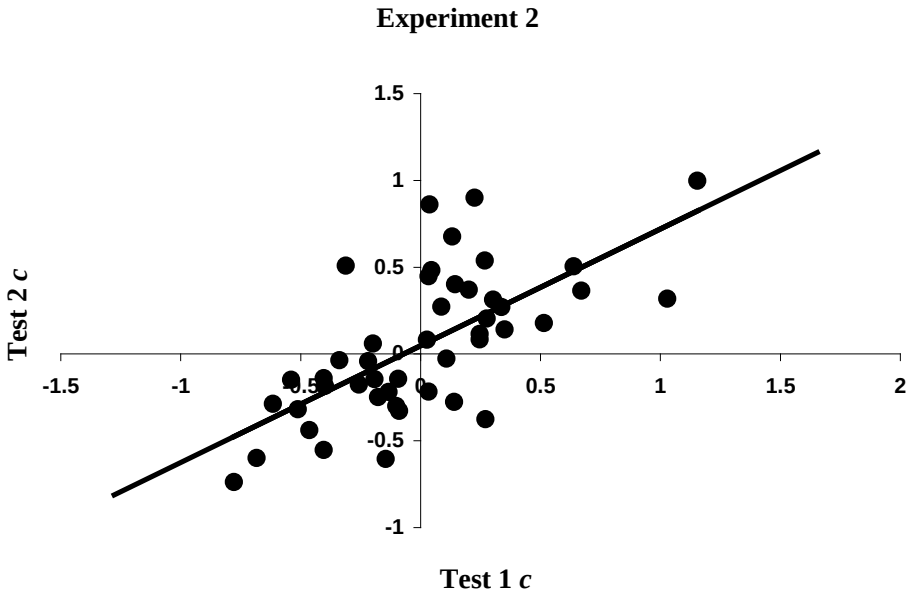


Figure 4. Correlation of recognition bias at Test 1 and Test 2 in Experiment 2.

Performance on the DRM and general knowledge tasks varied widely. Participants falsely recalled an average of 2.86 critical lures out of 8 possible (SD = 1.73,

range = 0 to 7). The critical measure was the correlation of the number of critical lures recalled and the average of Test 1 *c* and Test 2 *c* for each participant. Contrary to expectations, the correlation between these two quantities was close to zero (r = -0.08).

The critical measure in the general knowledge task was the mean number of years contained within each participant's range estimates (i.e., the mean range width). Across participants, the average range was 25.4 years (SD = 17.9). As is typically observed in studies of interval estimation (e.g., Yaniv & Foster, 1995), the majority of ranges were too narrow to capture the correct answer; the mean proportion of accurate ranges was 0.412 (SD = 0.150). Neither mean range width nor accuracy were significantly correlated with response bias (rs = 0.11 and -0.08, respectively; both ps > .48). Because the weak range width/bias correlation fell in the predicted direction, range width was analyzed according to a median split of bias scores. This analysis revealed that the average range of the most conservative recognizers was 7.3 years longer than that of the most liberal recognizers. However, this difference did not approach significance (p = .20).

To address the extent to which the range width/bias relationship was constrained by the reliability of the general knowledge task, the split-half reliability of the range width measure was calculated. The procedure was analogous to that used with *c* in Experiment 1 (see p. 24). Given the reliability estimate of 0.73 obtained for *c* in Experiment 1 and an estimate of reliability for the range width measure, a "correction for attenuation" can be applied in which the correlation between the two measures is adjusted to account for the noise in each individual measure (Murchinsky, 1996). While the resulting disattenuated correlation coefficient cannot be tested for significance, it provides an indication of the extent to which the relationship between the two variables is

undervalued by the traditional coefficient. This analysis was applied to non-recognition tasks in Experiments 4-6 and to the correlations involving personality measures, each time using the 0.73 as the estimate of the reliability of *c*.

The split-half reliability of the range width measure was 0.82, indicating that reliability was not concealing a relationship between recognition bias and conservatism in range estimates. Accordingly, the correction for attenuation raised the correlation between the two factors only slightly, to 0.14.

Experiment 2 was designed to test the within-individual stability of bias across time and the generality of recognition bias to two particular non-recognition tasks. The results were straightforward in both respects. Bias on a test given during the first session of the experiment was highly predictive of bias one week later; indeed, the correlation was nearly as strong as in Experiment 1, when the two tests were only ten minutes apart. While this comparison spans separate experiments and groups of participants, it is nonetheless worth emphasizing that the differences between the ten-minute and one-week intervals transcend duration. With a 10-minute interval, participants remain within the context of the experiment between tests, changing only the task with which they are engaged; when one week separates the tests, participants return to the laboratory for Test 2 having accumulated a week of life experiences since Test 1. The fact that these two intervals were associated with similar correlations of bias is strongly suggestive of trait-like stability.

Evidence of extension beyond recognition memory was not obtained, however. Bias was uncorrelated with false recall in the DRM paradigm and range size in estimation from general knowledge, despite the fact that the means and variability associated with

all three measures were well-suited to measuring individual differences. Potential explanations for these null results are given in the rationale for Experiment 4, which returned to the use of these tasks.

Experiment 3

While Experiment 2 provided evidence that response bias is consistent across time, Experiment 3 tested a second facet of trait-like stability: consistency across stimulus materials. In Experiments 1 and 2, the correlated bias measures were derived from two tests of word recognition, leaving open the possibility that bias is consistent for words (or, more generally, that it is consistent within the same stimulus domain), but differs unpredictably when the to-be-recognized stimuli change. To address this possibility, Experiment 3 included conditions in which two recognition study/test cycles varied in the class of materials used.

The stimulus domains chosen for the experiment were words and digital images of masterwork paintings. These materials are well suited to an examination of bias consistency across stimuli in two respects. First, words and paintings share few features beyond their visual presentation modality and contrast sharply along several dimensions: paintings are richly detailed, complex in subject matter, and thematically (and sometimes emotionally) evocative, while the common word stimuli used in the present experiments possess none of these attributes. The use of such qualitatively distinct stimulus sets provides a strong test of the within-individual consistency of bias across materials.

A second advantage of words and paintings in providing a rigorous test of bias consistency is their tendency to elicit very different magnitudes of bias on recognition tests: while words tend to produce roughly neutral responding, paintings are associated

with dramatic conservatism (Lindsay & Kantner, 2011). Note that bias consistency does not require that the obtained measure of bias is the same or even similar for a given participant across tests if the two tests use different stimuli. Rather, it requires that bias on Test 1 *predicts* bias on Test 2, such that a participant with a more liberal than average word recognition bias should to show a more liberal than average painting recognition bias (even though the bias may generally be more conservative for the latter than for the former). In Experiment 3, participants should show very different magnitudes of bias on the two tests. A finding that bias levels of individuals remain correlated across words and paintings (e.g., that most participants' bias becomes more conservative from words to paintings, but to a similar degree) would provide substantial evidence of trait-like stability in bias across materials.

The issue of differential biases across stimuli is also important because those differences may reflect variation in the subjective memorability of the stimuli (Brown et al., 1977; see Properties of Response Bias). As mentioned above, if two experimental conditions are distinguished by the degree to which participants expect stimuli to be memorable, participants will tend to apply a more conservative criterion in the condition with greater subjective memorability. Recent work in our laboratory (Lindsay & Kantner, 2010) found evidence that paintings may yield a more conservative bias than words due at least in part to the fact that they are perceived to be more memorable than words. Not all participants shared this belief, however: some predicted that they would have better memory for words than paintings and others predicted approximately equivalent memory for the two.

To the extent that the relative subjective memorability of the two item classes varies across individuals, bias in word recognition will become more difficult to predict as a function of bias in painting recognition. Thus, the impact of variability in subjective memorability on response bias has the potential to mask intra-individual stability. A result indicating within-person consistency in bias despite this potential source of metamnemonic variability would represent a significant extension of Experiments 1 and 2.

In Experiment 3, the materials difference across the two tests was embedded within a 2 (words or paintings at Test 1) x 2 (words or paintings at Test 2) between-subjects design. Thus, some participants received words on both Test 1 and Test 2, others received paintings on both Test 1 and Test 2, and others received one stimulus type on Test 1 and the other on Test 2. This design allowed an assessment of the consistency of bias within and between two stimulus domains.

Experiment 3 also initiated the collection of personality data in an effort to explore potential relationships between recognition bias and inherent personality characteristics. As these data were collected across several experiments, the personality measures and associated results will be discussed collectively following Experiment 7.

Method

Participants. 143 undergraduates participated in Experiment 3. Participants were randomly assigned to one of four conditions: the Word-Word (WW) condition (words in the first study/test cycle, words in the second study/test cycle), the Painting-Painting (PP) condition, the Word-Painting (WP) condition, and the Painting-Word (PW) condition.

The WW, PP, WP, and PW conditions included 40, 37, 35, and 31 participants, respectively.

Materials. Word stimuli were identical to those used in Experiments 1 and 2. Several hundred images of masterwork paintings were obtained from a computer-based memory training game called Art Dealer by permission of its creator (Jeffrey P. Toth of the University of North Carolina-Wilmington). This set contains large, full color, high definition images of works by well-known artists from the 17th to early 20th centuries (e.g., Rembrandt, Matisse, Cole, Modigliani, Caillebotte). Sample paintings appear in Appendix B. Two hundred and four of these images, representing a wide array of artists, styles, and subject matter, were selected for use in Experiment 3. Very famous works (e.g., Van Gogh's self-portraits) were avoided. All paintings were rectangular but dimensions varied widely; most covered approximately half of the computer screen area. Paintings and words were assigned to study and test phase(s) by the same method as words in Experiments 1 and 2.

Procedure. The procedure was identical to that of Experiment 1, with the following three exceptions. First, participants in the WP and PW conditions were not informed during the second study/test cycle that none of the items would be repeated from the first cycle; the use of different materials across the two cycles rendered the point self-evident. Second, the filler task was amended to include cities and geographical landmarks in addition to countries beginning with participant number 40. Third, following the second study/test cycle, participants completed two personality questionnaires (see p. 77).

Results and Discussion

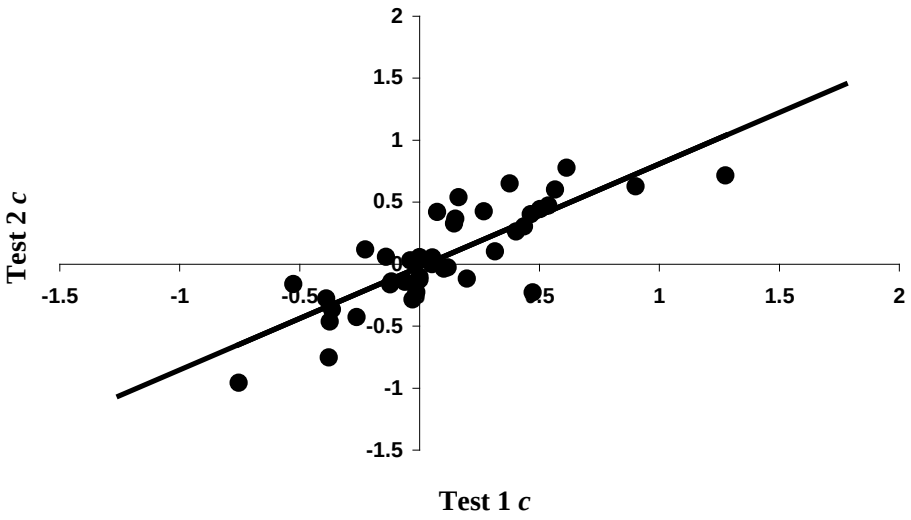
Recognition means are displayed as a function of condition in Table 3. Test 1/Test 2 comparisons and bias correlations are discussed separately by condition. Small but non-negligible positive correlations of *d'* and *c* emerged in two of the four conditions (*r*s = .02, .26, .05, and .26 in the PP, PW, WP, and WW conditions respectively; all *p*s > .09). Therefore, bias consistency was assessed with partial correlations controlling for Test 1 and Test 2 d' in each of the four conditions. Test 1 bias is plotted against Test 2 bias for each of the four conditions in Figure 5.

Table 3. Recognition means in Experiment 3.

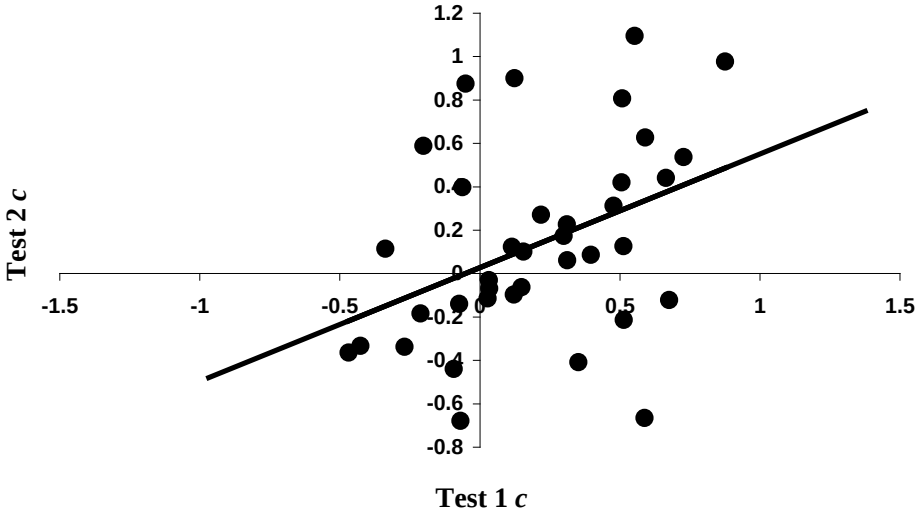
	H		FA		*c*		*d'*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
WW Condition								
Test 1 (Words)	.68	.15	.26	.15	.11	.39	1.25	.61
Test 2 (Words)	.71	.13	.26	.17	.07	.40	1.36	.69
PP Condition								
Test 1 (Paintings)	.69	.15	.19	.12	.21	.35	1.51	.60
Test 2 (Paintings)	.77	.14	.16	.12	.14	.45	1.97	.60
PW Condition								
Test 1 (Paintings)	.68	.13	.17	.12	.28	.41	1.58	.51
Test 2 (Words)	.75	.14	.26	.17	.00	.40	1.46	.65
WP Condition								
Test 1 (Words)	.70	.12	.24	.15	.09	.34	1.34	.55
Test 2 (Paintings)	.76	.13	.16	.11	.13	.32	1.93	.86

Note. WW = word-word, PP = painting-painting, PW = painting-word, WP = word-painting. Type of stimulus used in each test is in parentheses.

Experiment 3: WW condition



PP condition



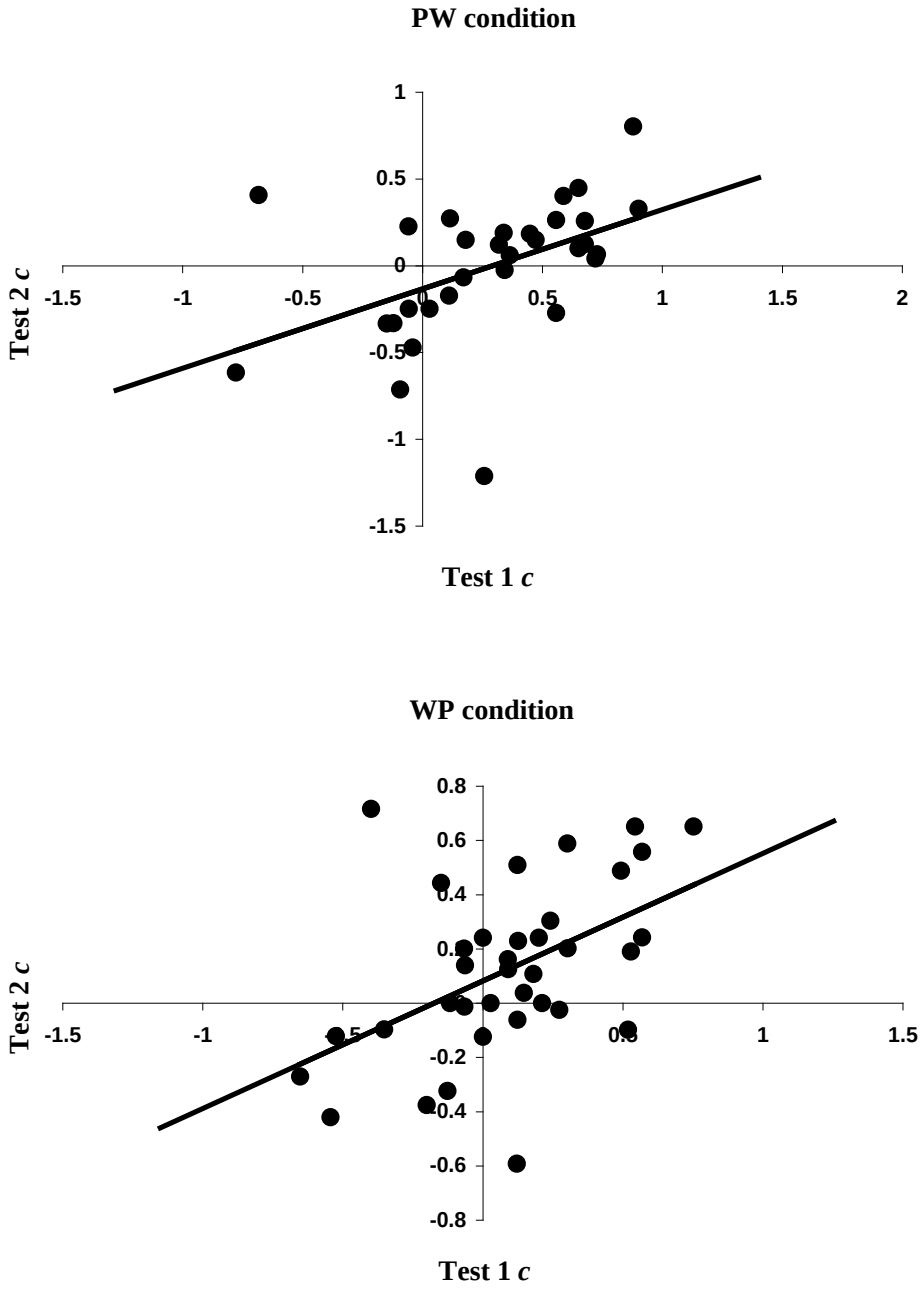


Figure 5. Correlation of recognition bias at Test 1 and Test 2 in Experiment 3.

WW condition. Neither *d'* nor *c* differed significantly across tests. Bias was significantly correlated across the tests, $r(36) = 0.81$, $p < .001$, replicating the findings of Experiments 1 and 2.

PP condition. Sensitivity rose significantly from Test 1 to Test 2, $t(36) = -5.18$, $p < .001$, while bias was unchanged ($t < 1$). Bias was significantly correlated across tests, $r(33) = 0.39$, $p < .05$.

PW condition. There was no significant difference in sensitivity on the painting and word tests ($t = 1.121$, $p = .27$). As expected, painting bias was much more conservative than word bias, $t(30) = 3.837$, $p < .001$. The bias correlation across tests was again significant, $r(27) = 0.45$, $p < .05$.

WP condition. Group differences in sensitivity and bias followed the opposite pattern of the PW condition: sensitivity differed significantly, $t(34) = 3.966$, $p < .001$, while bias was statistically equivalent ($t < 1$). The correlation of bias was significant and attained a magnitude similar to that seen in the PW condition, $r(31) = 0.49$, $p < .01$.

Thus, Test 1 bias remained strongly predictive of Test 2 bias when different materials were used in the two tests. The similarity of the correlations observed in the PW and WP conditions is sensible given the identical content of the two conditions (i.e., one study/test cycle with paintings and another with words). It is informative, however, in light of the divergent trends distinguishing the two conditions. The PW condition showed a sizable shift in bias from paintings to words but no change in sensitivity; the WP condition showed no shift in bias across stimuli (contrary to expectation) but significantly greater sensitivity to paintings than words. Bias stability, then, is apparently not reliant on a match in general discrimination or response bias across tests. The former finding is consistent with the signal detection theory assumption that discrimination and bias are independent properties, while the latter confirms the important hypothesis that bias need not be similar across tests to have a predictive relationship across tests (see p. 35).

Stability was not equivalent across all four conditions. Fisher's tests confirmed that the magnitude of the WW correlation was significantly greater than that of the other three conditions (all *z*s > 2.4, all *p*s < .05), which did not differ significantly from each other (all *z*s < 0.5). The decreased stability in the PW and WP conditions relative to the WW condition is not surprising and indicates that consistency in stimuli contributes to consistency in bias across tests. The relatively weak bias relationship observed in the PP condition, however, was an unexpected result, given that materials did not differ in this condition. One possible explanation for this finding was investigated in Experiment 6.

In sum, despite the differences in materials and inter-test bias and sensitivity trends across the four conditions, all showed reliable bias correlations across tests. Thus, the results of Experiment 3 followed those of Experiments 1 and 2 in providing evidence of trait-like stability in response bias.

Experiment 4

While Experiments 1-3 began to establish a foundation of evidence for the characterization of recognition bias as a manifestation of a stable trait, evidence in the form of generalization beyond recognition memory has been absent. The results of Experiment 2 suggest that bias is highly consistent within an individual over a period of at least one week, and Experiment 3 demonstrated that bias remains predictable with dramatic changes in to-be-recognized materials. In Experiment 2, however, recognition bias was uncorrelated with performance on a DRM free recall test and a general knowledge task tapping strategic adjustments of grain size, two tasks hypothesized to involve the same evidence criterion at work in producing trait recognition bias.

Experiment 4 returned to the DRM and grain size paradigms under conditions expected to increase the likelihood of detecting a relationship with recognition bias if one exists.

The DRM task in Experiment 4 was unchanged from Experiment 2, but the timing of the experiment within the course of the academic term was believed to better suit the DRM paradigm. Unfortunately, Experiment 2 took place midway through the spring term at the University of Victoria and overlapped with lectures on the DRM paradigm in various psychology courses. Interviews conducted during the debriefing revealed that more than half of the participants came to the experiment with fresh insight that they should avoid recall of critical lures. In experimental settings, warnings about the critical lure decrease DRM false alarm rates (see Starns, Lane, Alonzo, and Roussel [2007] for a review). Variability in this foreknowledge across Experiment 2 participants may have driven differences in false recall of critical lures, undermining the detection of other mediators (e.g., inherent response bias). Experiment 4 was conducted in the first half of the fall semester, at which time very few of the introductory psychology students that constitute the majority of research participants have been familiarized with the DRM paradigm. As in Experiment 2, liberal recognizers were predicted to recall a higher proportion of critical lures than conservative recognizers.

The general knowledge task was revised for Experiment 4 on the suspicion that the null result in Experiment 2 arose from the use of range size as the dependent measure. Despite the tendency for participants to overestimate their own knowledge levels, prior knowledge may have driven variability in range sizes to a far greater degree than response bias (which may, in addition, simply be unrelated to range estimation). Therefore, a new version of the task was created in which each question was

accompanied by two response options, one of which was correct (e.g., “What did year did CBC make its first television broadcast? a. 1953 b. 1963”). Participants were informed that they would gain 10 cents for every correct answer and lose 10 cents for every incorrect answer. They were also given the right to “pass” on any question to which they did not feel confident giving an answer (called *report option* by Koriat & Goldsmith [1994]). In this scenario, any question to which the subject does not have prior knowledge of the answer is a small gamble in which giving a response incurs risk that can be avoided with the exercise of report option.

The dependent measure of interest is the proportion of trials on which liberal versus conservative recognizers use report option. Conservative recognizers, assumed to require more memory evidence than liberal recognizers before committing to an “old” judgment, were hypothesized to require more confidence in their knowledge of the answer to a given question before committing to the gamble. Thus, conservative recognizers should exercise report option significantly more often. This task is particularly appealing, given the goals of this experiment, in that risk-taking behaviors have been associated with extraversion (Patterson & Newman, 1993), and extraversion, in turn, has been associated with a liberal recognition bias (Gillespie & Eysenck, 1980).

Method

Participants. There were 50 participants in Experiment 4.

Materials. The same words used in Experiments 1-3 served as recognition task materials. DRM task materials were identical to those of Experiment 2. The general knowledge task included 50 questions, each with two response alternatives. Approximately half of the questions were retained from Experiment 2; in order to

increase variety within the task, the remainder of the set comprised questions requiring numerical responses other than names of years. These questions were drawn from the original pool of 208 (see Experiment 2 Method).

Two response alternatives were prepared for each question. One alternative was always the correct answer. The second option was chosen by the author and was designed to pose as a plausible alternative that would generate uncertainty without making the task overly difficult. Generally, the incorrect alternative was a value of moderate distance from the correct answer.

Procedure. Experiment 4 consisted of four stages: a recognition study/test cycle, the DRM task, the general knowledge task with report option, and the two personality questionnaires administered in Experiment 3. Participants were tested in groups of 1 to 3, a measure taken to increase the efficiency of data collection given the unusual length of the experiment. In sessions of more than one participant, a second experimenter was present and aided in the transition between phases. The order of the recognition, DRM, and general knowledge tasks was counterbalanced across groups; within groups, each participant completed the tasks in the same order. All participants completed the personality questionnaires at the end of the experiment.

The recognition task followed a procedure identical to that of Experiments 1-3 (note, however, that only one study/test cycle was included in Experiment 4). The procedure for the DRM task was identical to that of Experiment 2, with two exceptions required by the group testing format. First, participants were asked to read list words silently. Second, a black and white flashing screen (rather than an auditory tone) alerted participants to the end of each two-minute recall period.

The general knowledge task was similar to the one used in Experiment 2, with differences reflecting the change to a two-alternative forced choice (2AFC) response format with report option. Task instructions were analogous to those in Experiment 2, with an additional component informing participants that they would gain ten cents for each correct response, lose ten cents for each incorrect response, and gain or lose nothing by choosing to "pass" on answering a given question. Instructions emphasized that a negative balance at the end of the task would not result in any loss of money.

Questions were again presented near the top of the screen with two boxes positioned underneath; these boxes contained response options A and B. Near the bottom of the screen appeared the words, "Press spacebar to pass." Participants chose an answer by entering 'a' or 'b' or passed by hitting the spacebar. Passing initiated the next trial. Selection of one of the response alternatives prompted the appearance of a confidence scale ranging from 50% to 100% near the top of the screen. Participants indicated their confidence in the selected answer via key press, initiating the next trial.

Upon finishing the general knowledge task, a "bonus code" appeared that revealed the participant's monetary balance to the experimenter but not to the group of participants. Winnings were distributed at the end of the session; participants were given the amount of their final balance or 50 cents, whichever was greater.

Results and Discussion

The data of two participants near chance in recognition accuracy were removed from subsequent analyses. The general knowledge test data of one additional participant were deleted due to a failure to follow task instructions. Therefore, the following analyses

included 48 participants in the DRM and recognition tasks and 47 participants in the general knowledge task.

Table 4. Recognition means in Experiment 4.

	H		FA		*c*		*d'*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Test 1	.66	.15	.28	.15	.11	.41	1.10	.54

Group recognition measures are displayed in Table 4. In the DRM task, participants falsely recalled an average of 2.98 critical lures out of 8 possible (SD = 1.92, range = 0 to 7). Unlike in Experiment 2, values of *c* were significantly correlated with frequency of false recall, $r(46) = -0.35$, $p < .05$, with a negative relationship indicating that increasing liberality of recognition bias was associated with increasing frequency of false recall (see Figure 6). Values of *d'* were also significantly correlated with false recall, with higher recognition sensitivity predicting fewer critical false alarms, $r(46) = -0.39$, $p < .01$. Because both bias and sensitivity predicted DRM performance, a partial correlation of *c* and frequency of false recall controlling for *d'* was calculated. The relationship remained significant, $r(45) = -0.29$, $p < .05$. Similarly, the *d'*-DRM relationship remained reliable when controlling for *c*, $r(45) = -0.35$, $p < .05$. These analyses suggest that both bias and sensitivity are related to false recall of critical lures in the DRM paradigm.

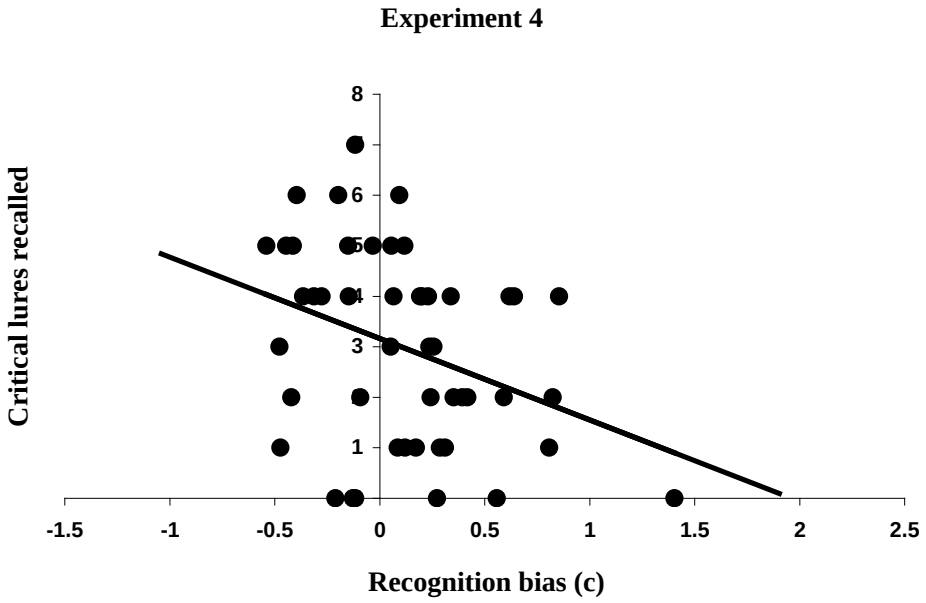


Figure 6. Correlation of recognition bias and frequency of DRM false recall in Experiment 4.

Participants chose to pass on an average of 12.53 out of 50 (25.1%) general knowledge questions (SD = 8.69). Individual participants' use of the pass option ranged from 0 to 40 times. When questions were answered with one of the two response alternatives, mean accuracy was 68.9% (SD = 10.5%) and mean confidence was 58.2% (SD = 12.5%). No relationship was detected between recognition bias and frequency of passing (r = 0.06), accuracy (r < .001), or confidence (r =-0.08). Split-half reliability of the pass measure was high (0.81), and correction for attenuation resulted in little adjustment of these correlations (rs = 0.08, 0.001, and -0.10, respectively). Recognition sensitivity was also unrelated to these measures (strongest r = -0.06).

One further analysis concerned an individual's frequencies of passing versus giving responses at a 50% confidence level (I thank Jordy Freeman for suggesting this analysis). Since both types of responses signify an expectation of chance-level ability in

answering a question, it was expected that this comparison would discriminate liberal and conservative responders: the former should be more likely to risk an incorrect response while the latter should be more likely to pass. However, preference for the pass option (the number of passes minus the number of 50% confidence responses) was uncorrelated with recognition bias ($r = 0.09$).

Experiment 4 provided the first indication of a relationship between recognition response bias and performance on a non-recognition task. Individuals using a more lax criterion for calling items old in a recognition test also used a more lax standard for recalling related but non-presented list items. Though further replication of this relationship is warranted, its presence in Experiment 4 supports the suspicion that the lack of relationship in Experiment 2 was due to the noise added by widespread foreknowledge of the task.

No relationship was observed between bias and performance in the general knowledge task. Experiment 5 addressed this null result.

Experiment 5

Experiment 5 was intended to address a shortcoming in the design of the Experiment 4 general knowledge task that may have undermined the detection of a relationship with bias: the offsetting values of reward and penalty for correct and incorrect answers, respectively. While the availability of the “pass” option was meant to give individuals lacking sufficient evidence of their ability to answer a question a means of avoiding the risk of penalty, the expected value of the pass option and of purely guessing one of the response alternatives was, in fact, the same. That is, even when two response alternatives to a question appeared equally plausible (i.e., the probability of a

correct response was at chance), there was no payoff-based advantage in passing: across trials, guessing would results in gains of ten cents 50% of the time and losses of ten cents 50% of the time (expected value = 0), while passing would bring a neutral result 100% of the time (expected value = 0). The fact that several participants never exercised the pass option in Experiment 4 supported the concern that it was not considered a useful alternative to guessing.

Use of the pass option, then, might have depended in part on whether a given participant apprehended its functional equivalence to guessing and incorporated that knowledge into his or her approach to the task. The goal of Experiment 5 was to remove this source of variation in the hopes of better detecting a report option-recognition bias relationship. The penalty for an incorrect response was raised to 15 cents, rendering the pass option objectively advantageous under high uncertainty.

Method

Participants. There were 81 participants in Experiment 5.

Materials. The materials were the same as those in the recognition and general knowledge tasks of Experiment 4.

Procedure. Participants completed the experiment individually. The experiment consisted of a recognition study/test cycle and a general knowledge task with report option. The order of the two tasks was random for each participant. The procedure differed in only two respects from that of Experiment 4: first, participants were correctly informed that the penalty for an incorrect response would be 15 cents. Second, the phrase “Press spacebar to pass” appearing on each trial was surrounded by a thick black border to increase its salience.

Results and Discussion

The data of three participants were dropped prior to analysis. Two of these participants reporting during debriefing that they had forgotten about the pass option; the third knew the answers to nearly all of the general knowledge questions and did not require the pass option. The below analyses include data from the remaining 78 subjects.

Table 5. Recognition means in Experiment 5.

	H		FA		*c*		*d'*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Test 1	.72	.15	.28	.14	-.01	.40	1.31	.62

Recognition measures are presented in Table 5. The general knowledge task results were very similar to those of Experiment 4. Participants chose to pass on an average of 11.32 out of 50 (22.6%) general knowledge questions (SD = 10.81). Individual participants' use of the pass option again ranged from 0 to 40 times. Mean accuracy on questions not passed was 70.2% (SD = 8.0%); mean confidence was 58.5% (SD = 11.4%). As in Experiment 4, no significant relationship was detected between recognition bias and frequency of passing (r = 0.11), accuracy (r = -0.11), or confidence (r = 0.10). Correction for attenuation based on the reliability estimates obtained in Experiments 1 (for *c*) and 4 (for pass frequency) augmented these values only slightly (rs = 0.14, -0.14, and 0.13, respectively). Preference for the pass option versus a 50% confidence response was again uncorrelated with recognition bias (r = 0.07).

Experiment 5 replicated the null relationship of recognition response bias and risk avoidance in a general knowledge task observed in Experiment 4. Unfortunately, the reasons for these null results are not evident from the data. It might be the case that conservatism in a recognition task and conservatism in a gambling-oriented general knowledge test have independent cognitive substrates, and that trait-like stability in recognition bias is not relevant to the class of decisions exemplified in the general knowledge task. Alternatively, the task itself might not have been an adequate test of risk tolerance. The risk and reward associated with each question was relatively small, and the extra 5 cents associated with an incorrect response (incorporated in Experiment 5) might have been insufficient incentive for conservative responders to pass more often when uncertain. Indeed, the mean number of passes was directionally (but not significantly) lower in Experiment 5 than in Experiment 4, suggesting that the reward/penalty imbalance had little influence on participants. Future approaches to studying the response bias-risk relationship are mentioned in the General Discussion.

Experiments 6 and 7: Cross-situational Consistency in Response Bias

The final two experiments reported here tested an array of hypotheses related to the trait-like status of recognition memory response bias. Those addressed by the two individual experiments are described separately below. Jointly, Experiments 6 and 7 held the overarching purpose of assessing the cross-situational consistency of response bias. The motivation and method for establishing cross-situational consistency are introduced here.

Thus far the evidence held to support the idea of response bias as a cognitive trait has taken the form of significant correlations of bias across time (especially in

Experiment 2), materials (Experiment 3), and tasks (the bias-DRM relationship in Experiment 4). If recognition bias is a manifestation of a trait, trait-like stability should also be in evidence across distinct, unrelated situations. The results of Experiments 1-5 do not speak to this facet of bias consistency because all of the correlated measures were taken within the same experimental context: the "situation" did not change. The experimenter, testing environment, day of the session (except in Experiment 2), and time of day were all constants. Even in Experiment 2, composed of two sessions that took place one week apart, the identity of the experiment remained the same, and participants were fully aware that the two sessions were connected. Perhaps the observed consistency of response bias within individuals was upheld by the consistency of the measurement context rather than an inherent characteristic of the individuals.

The overarching goal of Experiments 6 and 7 was to create two ostensibly independent testing situations and to measure response bias in the same individuals in both situations. Central to the endeavor was having as many participants as possible complete both experiments, but without any awareness of a direct connection between the two prior to debriefing. The recruitment method was as follows. Experiments 6 and 7 were posted on the University of Victoria's online subject pool management system (Sona; http://www.sona-systems.com) under two separate names, Cognitive Judgments and Many Faces. Initially, participation in the two studies was mutually exclusive: signing up for one disqualified an individual from signing up for the other. After a sufficient *N* had accrued in both experiments (approximately three weeks later), they were changed to become mutual pre-requisites: one could not sign up for Cognitive Judgments *unless* one had already completed Many Faces, and vice versa. From this

point forward, any new participants in either study were completing the two-experiment sequence, allowing for the calculation of bias correlations across what participants believed to be independent situations.

Steps were taken to ensure that the sign-up process did not betray a connection between the two studies. Their names (Cognitive Judgments and Many Faces) suggested no obvious link, and one was posted with an optional description while the other was not. Some minor commonalities in the postings of the two studies – the presence of the author's name and a contact phone number for the laboratory – could not be avoided. A more serious concern was that the information screen for each study listed the name of the other as a prerequisite. This cross-listing could not be eliminated, and presented perhaps the most salient clue that the two experiments were connected. In order to dilute the influence of the cross-listing, 15 other studies were added as potential prerequisites for each study. For example, the list of potential prerequisites for Cognitive Judgments included Many Faces and 15 additional experiment names. By default setting of the Sona system, having completed any one these studies made one eligible for Cognitive Judgments. The "filler" prerequisites were, however, defunct studies from several years prior in which very few (if any) current students were likely to have taken part. Importantly, then, Many Faces participants still constituted virtually the sole source of participants for Cognitive Judgments (and vice versa). Embedded within a 16-item list of experiments, however, it was hoped that the critical prerequisite would not be obvious to participants.

To further disguise the relationship between the two studies, the experiments themselves were designed to be as distinct from one another as possible. Experiment 6

included two recognition study/test cycles using paintings as stimuli, a preceding practice phase, and a go-no go reaction time task; debriefing centered on the impact of the early practice phase on improvements in accuracy across the two recognition tests. Experiment 7 included recognition cycles using words and faces as stimuli, the viewing of video clips depicting crimes, and a subsequent suspect identification task; here the debriefing emphasized the exploration of a link between recognition performance in a standard laboratory paradigm and in a more realistic eyewitness scenario. Experiment 6 took place in a small room on the first floor of the building while Experiment 7 took place in a larger room with a different general appearance in the basement of the building. The experimenters differed between the studies. The fonts of the instructions and the appearance of the consent forms used in each study varied. Although Experiments 6 and 7 included recognition tasks that bore an unavoidable resemblance to one another, the wording of the instructions for the study and test phases in Experiment 6 were made to vary from those of Experiment 7. In addition, the response format in the test phases differed: in Experiment 6, participants used the same 1-6 scale as in Experiments 1-5, while in Experiment 7 they made a binary studied/not studied judgment followed by a rating of confidence on a scale of 1-3 (note that these two response types provide the same information).

Despite the numerous points of departure between Experiments 6 and 7, both took place during the same semester, and the similarity of the recognition tasks across the two experiments was evident to most participants. However, debriefing indicated that none suspected that the two experiments were directly connected, or that their own performance was being compared across the two. The differences in context, time,

materials, and perceived identity of the experiments were meant to produce a situational manipulation of sufficient strength to assess the cross-situational stability of response bias within an individual.

The rationale, methods, and results of the individual experiments are described next, followed by the results from the subset of participants completing both.

Experiment 6

Experiment 3 demonstrated significant positive correlations of bias across recognition tests differing markedly in the to-be-recognized materials, indicating a trait-like robustness of bias with respect to the class of items queried of the recognition system. Nonetheless, a natural assumption is that the nature of the materials does have some impact on the strategic component of a recognition decision independent of an individual's predisposition to bias, such that bias correlations will be stronger when the materials are of the same class across tests. In this respect, the results of Experiment 3 were puzzling: while the WW condition (words used in both study-test cycles) produced the highest inter-test correlation at .81, the PP condition (paintings used in both study-test cycles) produced the lowest (.39), significantly lower than that of the WW condition and directionally lower than the correlations in the two conditions in which materials differed across tests (.45 and .49).

Experiment 6 investigated the unexpectedly modest relationship of bias in the PP condition. Why was the relationship so much weaker across two paintings tests than across two words tests? One possible explanation is that either or both of the correlations were statistical flukes. Indeed, the WW correlation was substantially higher than the analogous correlations observed in Experiments 1 and 2, which suggests that it might

have been inflated by chance. However, the greater stability of response bias across two word recognition tests than across two painting recognition tests may not have been the result of statistical error. One factor potentially driving a difference between the conditions is the relative novelty of processing and attempting to recognize a large number of paintings. Perhaps participants in the PP condition used the experience gained in the first study/test cycle to alter their approach to the second, producing a change in response bias across the tests. For example, participants might have adjusted their encoding strategy in the study phase of the second cycle after learning the nature of the recognition test during the first cycle. Indeed, the observed increase in recognition sensitivity from Test 1 to Test 2 in the PP condition supports the idea that participants used the first study/test cycle to benefit performance in the second. By contrast, lifelong experience processing and recognizing words might preclude further improvement over the course of the experiment in the WW condition; alternatively, word stimuli may not present obvious avenues for improvements in strategy (Postman, 1982).

A straightforward prediction of the above account of the PP condition results is that adding a practice phase before the two study/test cycles should strengthen the relationship of bias across the two tests. That is, if experience with the testing format leads to strategy corrections that alter bias, providing such experience prior to Test 1 should help encourage corrections to be made before, rather than between, Test 1 and Test 2. To test this prediction, half of the participants in Experiment 6 (the “pretest” group) completed an 8-item study list, 16-item test practice phase prior to two study/test cycles with painting stimuli. A control (“preview”) group simply viewed 24 paintings during the practice phase. If test experience drives adjustments of strategy that mediate

bias consistency with paintings, the pretest group should show greater consistency than the preview group; if, however, exposure to paintings is sufficient to enable strategic adjustments, the two groups might be expected to show similar levels of bias stability. If, instead, both groups display correlations at the level seen in Experiment 3, the results will suggest that some other characteristic intrinsic to painting stimuli bears importantly on bias stability.

A second goal of Experiment 6 was to provide a new test of the generality of response bias to tasks beyond recognition memory. Recognition bias was correlated with performance on a "go-no go" (GNG) task that filled the interval between study/test cycles. The details of a GNG task may vary, but the central facet of the paradigm is a test of cognitive control: participants must make simple responses (e.g., hitting the spacebar) to simple stimuli (e.g., letters) as quickly as possible, but must withhold or alter the response on a minority of trials. In the GNG task used in Experiment 6, for example, a series of letters appeared on the screen one-at-a-time. Participants were to hit a key with the right hand as soon as each letter appeared *unless* the letter was 'J,' in which case they were to hit a key with the left hand as quickly as possible. Letters other than J were presented on 75% of trials ("go" trials), priming participants to respond with the right hand. An accurate response on a J ("no go") trial, then, required participants to override this initial impulse and instead respond with the left hand. Because the GNG task is speeded, a classic speed-accuracy tradeoff is observed: the faster the average response, the greater the number of errors committed.

The GNG task is an intuitive one for correlating with recognition bias because performance on both tasks can be described with a signal detection model. Just as

individuals engaged in a recognition task are theorized to establish a response criterion along a continuum of "evidence of oldness" that reflects a tolerance for misses versus false alarms, those engaged in a GNG task are assumed to place a criterion along a continuum of evidence that the majority (in the present case, right-hand) response should be made. The GNG criterion reflects a tolerance for slowness versus error-proneness. Because conservative recognizers are assumed to require more evidence of oldness before making an "old" judgment, they were hypothesized to require more evidence that a given trial is a "go" trial before committing to the right-hand response. Conservatism of recognition bias was thus predicted to correlate positively with accuracy and with reaction time on "no go" trials.

Method

Participants. There were 87 participants in Experiment 6. Participants were randomly assigned to the pretest (N = 46) and preview (N = 41) conditions.

Materials. Twenty-four new paintings were selected from the Art Dealer collection (see Experiment 3 Method) for use in the practice phase. In the pretest condition, eight items were drawn randomly from this set of 24 to compose the practice study list. The practice test list contained the eight studied items plus eight randomly selected new items. In the preview condition, all 24 paintings were shown to participants in a random order. The paintings used in the two main study/test cycles were the same as those in Experiment 3.

Materials for the GNG task consisted of all 26 letters of the alphabet presented in a large, white, boldface font against a black background.

Procedure. The procedure was identical to that of the PP condition of Experiment 3 with two exceptions: the addition of a practice phase at the beginning of the experiment and the use of a GNG filler task between study/test cycles instead of country naming. In the pretest condition, participants were informed that there would be a short practice version of the study list and memory test before the full versions to come. To reduce the potential for the practice phase to set incorrect expectations for the length of the full study list, instructions mentioned that the full study list would contain "approximately 50 items." The practice phase followed the study and test procedure of past experiments. In the preview condition, participants were told they would be viewing a small sample of items to "give an idea of what the paintings look like." The 24 samples were then presented for 1 s each with a 1-s ISI. The procedure was identical for the pretest and preview conditions from the end of the practice phase forward.

Participants next completed the first study/test cycle, followed by the GNG task. Participants first completed a "warm-up" round in which they were asked simply to hit the 'L' key with their right hand as quickly as possible whenever any letter appeared on the screen. Each letter of the alphabet was presented once for a total of 26 warm-up trials. Instructions then informed participants that their task in the round to follow was to hit the 'L' key with their right hand as quickly as possible whenever any letter appeared unless the letter was 'J,' in which case they were to hit the 'A' key with their left hand as quickly as possible. Each letter was presented for up to 2 s. A response triggered a blank 500-ms ITI, followed by the next trial. On the rare occasion that a response was not registered within 2 s, the following trial began automatically. Participants completed three blocks of 100 trials each. Within each block, the letter J was presented 25 times and

each remaining letter was presented three times. The order of the letters was randomized. Participants were given rest breaks of open duration between each test block. The GNG task took approximately 8 minutes to complete, producing roughly the same interval between study/test cycles as Experiments 1 and 3.

The second study/test cycle followed. Finally, participants completed two new personality inventories (discussed on p. 77).

Results and Discussion

The data of one participant who had taken art history and reported familiarity with many of the paintings were removed prior to analysis, leaving 45 pretest and 41 preview participants in the following analyses.

Table 6. Recognition means in Experiment 6.

	H		FA		*c*		*d'*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Pretest Condition								
Test 1	.76	.12	.17	.11	.16	.34	1.83	.59
Test 2	.80	.13	.16	.11	.05	.34	2.04	.85
Preview Condition								
Test 1	.72	.13	.17	.12	.20	.36	1.69	.57
Test 2	.82	.13	.15	.10	.06	.35	2.20	.79
Overall								
Test 1	.74	.12	.17	.11	.18	.35	1.76	.58
Test 2	.81	.13	.16	.11	.05	.34	2.12	.82

Recognition means for Tests 1 and 2 in each condition are presented in Table 6. Across all participants, sensitivity was significantly greater in Test 2 than Test 1, $t(85) = 4.92$, $p < .001$. Mean response bias shifted significantly across tests, $t(85) = 3.66$, $p < .001$, from conservative to neutral. Small correlations of *d'* and *c* were present in both Test 1 ($r = 0.16$) and Test 2 ($r = -0.19$). Therefore, all bias correlations controlled for *d'*. The correlation of bias across tests was highly significant, $r(82) = 0.59$, $p < .001$.

Analyses of sensitivity and bias in the pretest and preview conditions revealed very similar trends. Separate 2 (Test: 1 or 2) x 2 (Condition: Pretest or Preview) mixed factor ANOVAs were conducted on the *d'* and *c* values. For *d'*, there was a main effect of test, $F(1, 84) = 26.1$, $p < .001$, $\eta_p^2 = .237$, reflecting the increased sensitivity on Test 2. There was no main effect of condition ($F < 1$), but a reliable Test x Condition interaction indicated that the increase from Test 1 to Test 2 was greater in the preview condition than in the pretest condition, $F(1, 84) = 4.261$, $p < .05$, $\eta_p^2 = .048$.

Values of *c* shifted significantly from Test 1 to Test 2, $F(1, 84) = 13.4$, $p < .001$, $\eta_p^2 = .137$, but neither overall *c* nor the inter-test shift in *c* varied as a function of condition (both *F*s < 1). The correlation of *c* across tests was significant in the pretest condition, $r(41) = 0.56$, $p < .001$, and in the preview condition, $r(37) = 0.58$, $p < .001$ (see Figure 7). These correlations did not differ from one another ($z < 0.5$). Thus, the only distinction in the results of the pretest and preview conditions was the magnitude of the increase in sensitivity from Test 1 to Test 2.

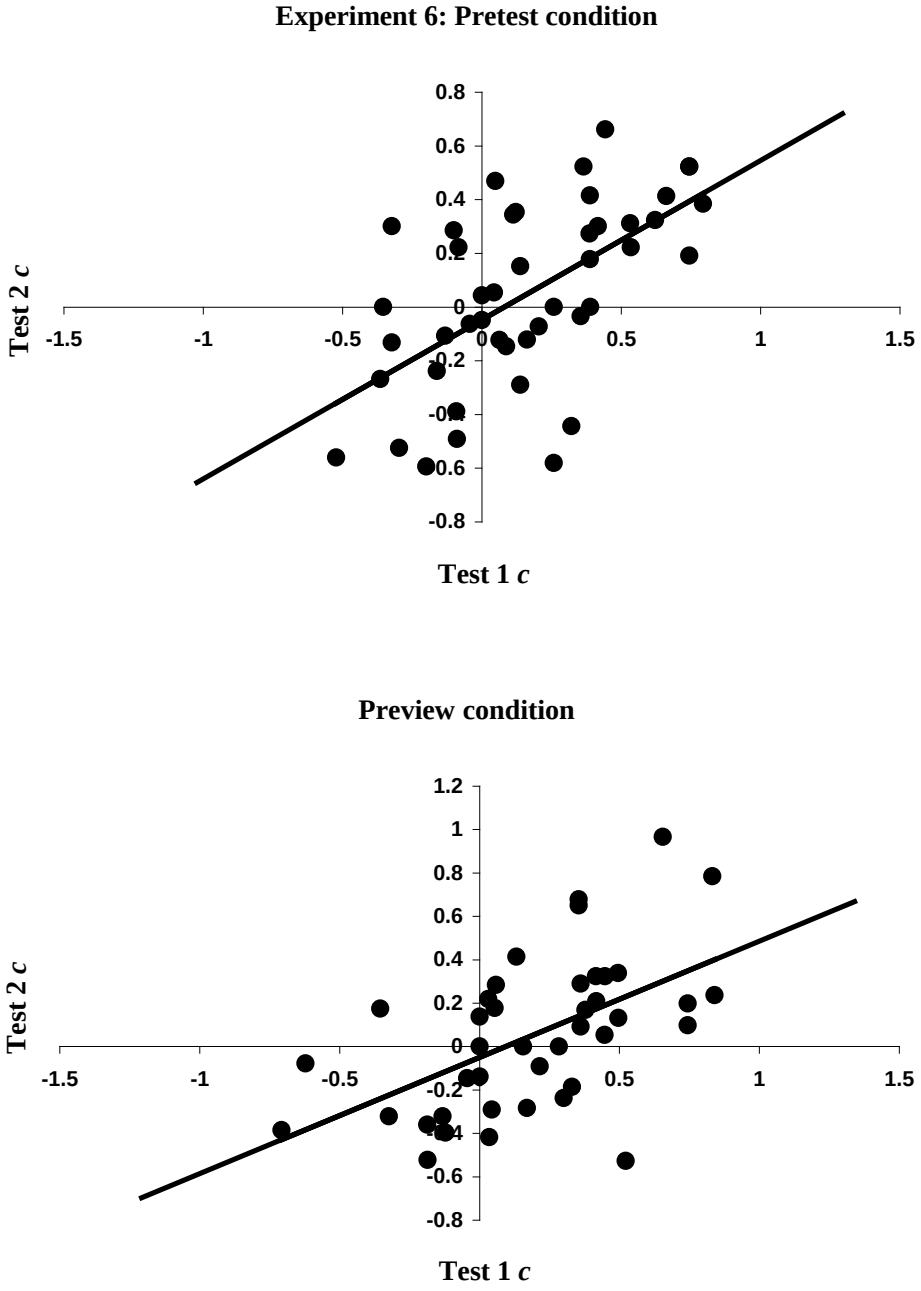


Figure 7. Correlation of recognition bias at Test 1 and Test 2 in Experiment 6.

GNG analyses focused on the minority of trials on which a 'J' was presented ("no go" trials). Mean accuracy on these trials was 78.8% (SD = 12.5%); mean RT was 406 ms (SD = 43). Accuracy and RT were positively correlated, $r(84) = 0.64$, $p < .001$,

reflecting a classic speed-accuracy tradeoff. Correlations were computed between these measures and each participant's mean *c* value across both recognition tests. Neither GNG accuracy nor RT correlated with bias (*r*s = 0.11 and 0.12, respectively). Split-half reliabilities were high for both the accuracy (0.74) and RT (0.89) measures, and correlations corrected for attenuation indicated little if any relationship between recognition bias and GNG performance (both *r*s = 0.15).

The results of Experiment 6 suggest that the weak bias correlation in the PP condition of Experiment 3 – the weakest inter-test correlation observed in the present experiments – was not related to the improvement in recognition sensitivity from Test 1 to Test 2. This improvement was observed in both the pretest and preview conditions, and both showed substantially stronger bias correlations than in Experiment 3. Experiment 6 lacked a condition replicating the PP condition of Experiment 3 (i.e., with no training phase prior to the first study/test cycle). It is possible that this training phase, whether consisting of a model study/test cycle or a sampling of the stimuli, led to a strengthened Test 1-Test 2 bias relationship. However, given that the correlations observed in the pretest and preview conditions were similar to each other and well within the range established by the previous experiments, the results suggest that the magnitude of the paintings-paintings correlation in Experiment 3 was anomalous.

The second goal of Experiment 6 was to test the relationship of recognition bias and two measures of GNG performance. Though the slight positive correlations between the tasks fell in the predicted direction, they did not approach significance, indicating that an individual's criterion for calling an item old was not related to his or her speed/accuracy criterion. An important difference between these two tasks is that GNG

performance is driven by the instruction to respond as quickly as possible, while recognition responses are self-paced. Whether a speeded recognition task would reveal a closer correspondence of bias with GNG performance is a question for future research.

Experiment 7

Experiment 7 was motivated by the expectation that if response bias represents a stable cognitive trait, it should carry predictable consequences for recognition decisions outside the laboratory. A natural testing ground for this assumption is the domain of eyewitness identification from police lineups. When witnesses to a crime are called upon to identify the culprit from a lineup, they are engaged in a real world recognition task, and one with severe ramifications: a false alarm may help convict an innocent person, and a miss may set a criminal free. A sizable literature demonstrated the fallibility of eyewitness memory in lineup-based identification (Brewer & Palmer, 2010) and has led to the development of guidelines for constructing lineups so as to minimize the potential for errors (e.g., Luus & Wells, 1991).

Research using lineups in which the culprit is not present (*target-absent* lineups) has revealed an unsettling potential for false alarms in lineup identification. Although witnesses have the option to (correctly) reject the entire lineup, they often select a suspect (e.g., Luus & Wells, 1994). Given that a liberal response bias in recognition memory is associated with a higher false alarm rate than a conservative bias, a straightforward prediction for accuracy in lineup decisions is that liberal recognizers should be more likely to make a false identification in a target-absent lineup. If response bias is both stable within an individual and predictive of false lineup identification rates, it should be

possible to derive a prediction of the likelihood that a given individual will identify an innocent lineup member based on an assessment of his/her response bias.

In Experiment 7, participants viewed five short video clips, each depicting a person committing a crime. Later, participants were shown five lineups one-at-a-time, each corresponding to one of the crimes witnessed earlier. They were informed that each lineup may or may not contain the culprit, and told that they may reject the lineup if they did not think the perpetrator was present. To increase the potential for false alarms, each lineup contained a foil bearing a strong resemblance to the culprit.

An individual's frequency of choosing a lineup member was correlated with response bias on two recognition tests: one, placed at the start of the experiment, using face stimuli, and another, placed between the viewing of the crime videos and the lineup task, using word stimuli. The inclusion of recognition cycles involving face and word stimuli served three purposes. First, it introduced faces as a rich new domain for testing bias consistency. Given the pervasiveness of face processing in everyday life and the special brain mechanisms often proposed to contribute thereto (e.g., Kanwisher, 2000), faces are arguably a more ecologically valid stimulus domain than words or paintings. Second, it allowed for two new tests of bias consistency across highly distinct materials: faces-words (in Experiment 7) and faces-paintings (for those participating in both Experiment 6 and Experiment 7). Third, it allowed a test of the possibility that the predictive efficacy of response bias for accuracy in target-absent lineup judgments depends on the stimulus used to measure bias. While both face recognition and word recognition are hypothesized to tap an underlying predisposition to bias that should also influence lineup decisions, face recognition bias was expected to be the better indicator of

lineup performance by virtue of the fact that the lineup scenario is itself a face recognition task. The measure of interest was the relationship of response bias in the recognition tests and the number of times (out of five) that a suspect was identified.

Method

Participants. There were 74 participants in Experiment 7.

Materials. Word stimuli were the same as those used in previous experiments. Face stimuli were drawn from a database created at the University of Victoria containing the pictures of 80 undergraduate psychology students with multiple facial expressions. Sample faces appear in Appendix C. Each image depicted the subject from the shoulders up. The shoulders were covered with a black wrap to conceal differences in clothing; hair was in full view. Images used in the experiment depicted faces looking straight ahead. The study list was composed of 38 neutral-expression faces drawn at random from the set of 80. The test list contained those 38 faces with a smiling expression randomly intermixed with 38 new smiling faces. Thus, the images of the faces seen both at study and at test were not the same; participants were required to recognize faces rather than pictures. Four additional faces were included as primacy and recency buffers.

The crime videos and corresponding lineups were obtained from various eyewitness memory researchers. A still frame from one of the crime videos and the corresponding lineup appear in Appendix D. The videos depicted the planting of a bomb, a stalker at an ATM, a breaking and entering into a home, a stalker at a park, and a theft at a store. Videos ranged from 45 to 75 seconds in length. The view of the culprit varied in duration and clarity across videos, but the culprit's face was clearly visible for a portion of each video. Videos were presented with Windows Media Player.

Each lineup contained six members depicted in equal-sized (approximately 4 cm x 6 cm) headshot photographs in the upper half of a Microsoft PowerPoint slide. All six members were of the same apparent age, gender, and ethnicity as the culprit, who was always absent. At least one member of each lineup bore a strong resemblance to the culprit. The ordering of the videos, lineups, and lineup members was constant across participants.

Procedure. The experiment included two recognition study/test cycles differing from those of earlier experiments only in the response format. Instead of entering a number on a six-point, confidence-graded scale, participants first made a binary old/new judgment by pressing '1' to indicate "old" and '2' to indicate "new." Participants were immediately prompted to enter their confidence in the preceding judgment on a scale from 1 to 3 (1 = low, 2 = medium, 3 = high). The decoupling of the recognition and confidence judgments and the use of the terms "old" and "new" in place of "studied" and "not studied" were intended to further differentiate Experiment 7 from Experiment 6 for individuals participating in both (see p. 53).

The experiment began with a face recognition study/test cycle, followed by the viewing of the crime videos. Given the potential sensitivity of lineup performance to subtle instructional differences, instructions preceding the videos and lineups were read verbatim to ensure consistency across participants. Video viewing instructions were as follows: "You are about to watch a series of videos depicting a crime. Later on you will be asked to choose the perpetrator of the crime from a six-person lineup for each video." Each video was preceded by a title screen, during which the experimenter said the number of the video aloud. These steps to mark the beginning of each new video were

taken to ensure that participants would not find it difficult to associate lineups with their corresponding videos later in the experiment.

After watching the five videos, participants completed a recognition study/test cycle with word stimuli. They then proceeded to the lineup task. Instructions were designed to make salient the option to reject lineups without explicitly encouraging any given response, and read as follows: "You will now see a lineup for the crime in which [brief description of crime]. The person who committed the crime may or may not be in the lineup. If you believe that the person who committed the crime is in the lineup, please circle the number of the suspect on the sheet. If the person who committed the crime is not in the lineup, please select 'The perpetrator is not in the lineup.'" The experimenter then displayed the first lineup and gave the participant a sheet of paper and a pen to record his/her judgment. A number between 1 and 6 appeared beneath each lineup member and was used to identify suspects. The title of the associated video appeared at the top of the screen. Participants either selected a member of the lineup by writing his/her number on the sheet or checked a box to indicate that they did not see the perpetrator in the lineup. They then rated the confidence and ease of that decision on separate 1-10 scales. Participants were given as long as needed to make lineup judgments.

Upon completing their judgments for a given lineup, participants handed the form to the experimenter, who then handed them a new form, re-read the first sentence of the lineup instructions (see above), and displayed the next lineup. No verbal feedback was given on judgments. Lineups were displayed in the same order as the corresponding videos had been.

At the end of the experiment, participants completed two personality inventories not included in Experiment 6. These are described in the Personality Measures section.

Results and Discussion

A single target-present lineup was unintentionally included among the lineups given to the first six participants. These participants' lineup data were removed prior to analysis; all other data were retained. The below lineup and recognition analyses therefore include 68 and 74 participants, respectively.

Table 7. Recognition means in Experiment 7.

	H		FA		*c*		*d'*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Word Recognition	.75	.13	.25	.16	.03	.34	1.54	.84
Face Recognition	.66	.13	.26	.12	.12	.30	1.15	.54

Recognition means are displayed in Table 7. A lower hit rate on the face test than on the word test, possibly due in part to the difference in face images at study and test, drove significant differences in both *d'*, $t(72) = 3.563$, $p < .01$, and *c*, $t(72) = 2.256$, $p < .05$, across tests. Values of *d'* and *c* were significantly correlated on the face test, $r(71) = 0.30$, $p < .05$, but not on the word test ($r = 0.06$). Bias correlations controlled for face and word *d'*. The correlation of face and word bias was significant, $r(63) = 0.55$, $p < .001$ (see Figure 8).

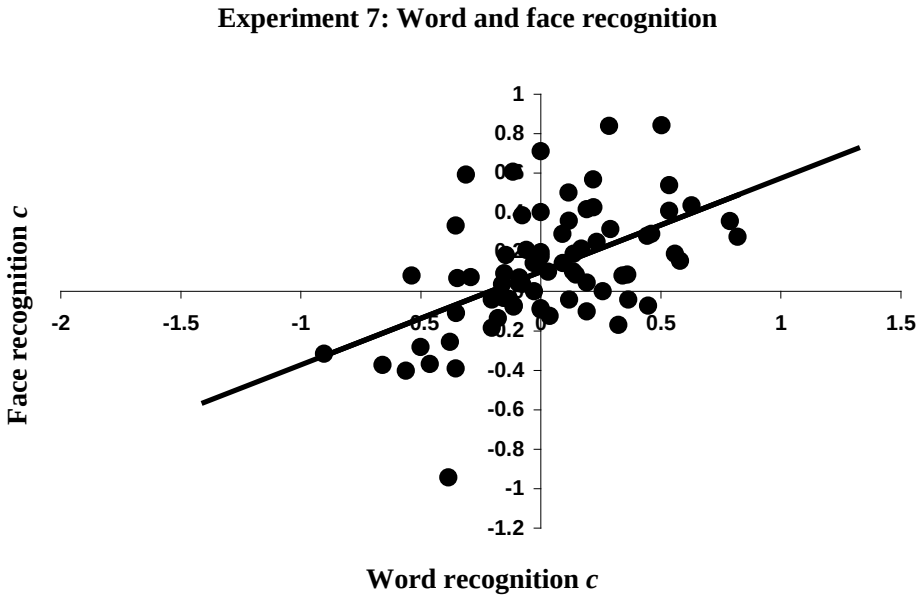


Figure 8. Correlation of word and face recognition bias in Experiment 7.

Every participant identified a suspect in at least one lineup. The mean number of suspects identified was 3.09 (SD = 0.92) out of a possible 5. Response bias is plotted against lineup performance in Figure 9. There was a significant negative correlation between face test bias and number of suspects identified, $r(63) = -0.30$, $p < .05$, indicating that more liberal recognizers tended to make more false identifications. This relationship was slightly weaker for word test bias, but remained significant, $r(63) = -0.24$, $p < .05$. These two correlations did not differ significantly from one another ($z < 0.5$). Neither face recognition nor word recognition *d'* were correlated significantly with frequency of identification (rs = -0.16 and -0.06, respectively; both ps > .21). Thus, recognition sensitivity did not predict lineup decisions.

Mean confidence and ease ratings are displayed as a function of lineup decision (identification or rejection) in Table 8. Neither face nor word bias was significantly related to confidence or ease ratings.

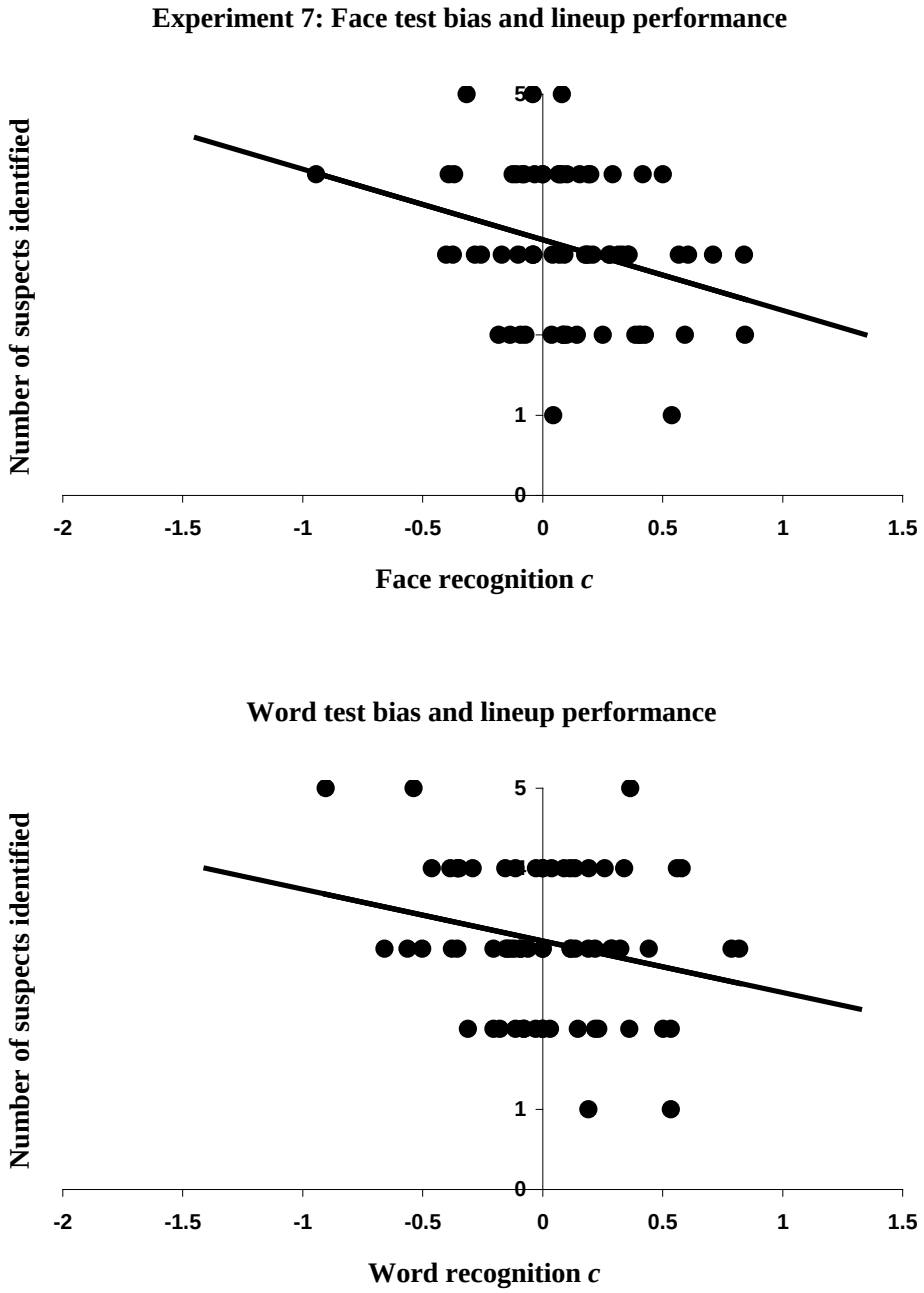


Figure 9. Correlation of recognition bias and frequency of suspect identification in Experiment 7.

Table 8. Confidence and ease ratings in the lineup task and their correlations with recognition bias in Experiment 7.

	M	*SD*	r_{words}	r_{faces}
Confidence Identify	6.0	1.3	-.16	-.17
Confidence Reject	6.0	2.0	-.08	-.14
Ease Identify	5.9	1.3	.01	.11
Ease Reject	5.6	1.9	-.09	-.06

Experiment 7 extended previous support for the stability of bias across materials by demonstrating a strong correlation between word and face recognition bias. Further, both face bias and word bias were found to be moderate but significant predictors of false identification in an eyewitness memory lineup scenario. The implication of this result is that response bias in a typical laboratory recognition test is related to recognition decision making in more complex and realistic settings, and thus bears practical consequences. One striking aspect of this finding was that recognition bias predicted the frequency of false identification while recognition sensitivity did not. These results are considered in detail in the General Discussion.

Experiments 6 and 7 joint participants: Results and Discussion

Thirty-five participants completed both Experiment 6 and Experiment 7. Eighteen participated in Experiment 6 first and 17 participated in Experiment 7 first. The mean interval separating the experiments was 17.0 days (SD = 11.1).

An interview during the debriefing following the second of the two experiments established the extent to which participants were aware of a connection between the two.

All but three participants reported noticing similarities between the two experiments prior to being informed of the connection; the remaining three recognized the similarities immediately after being informed. The majority of participants (28) noticed that the recognition task in the second study resembled that of the first study, either upon reading the instructions or beginning the task. Five participants had noticed a connection while signing up for the second study (e.g., the overlap in the listed investigators, the fact that each study was a potential prerequisite for the other). Three participants, when prompted, said it had occurred to them during the second study that the two experiments might have had overlapping aims. Critically, however, none reported suspecting that there was a direction connection between the two, that their own performance in the two experiments would be compared, or that response bias was of interest in either study.

Mean response bias in Experiment 6 (averaged across the two paintings tests) is plotted against Experiment 7 word bias and Experiment 7 face bias in Figure 10. Experiment 6 painting bias was significantly correlated with Experiment 7 word bias, $r(33) = .44$, $p < .01$, and with Experiment 7 face bias, $r(33) = .39$, $p < .05$. Mean Experiment 6 bias and mean Experiment 7 bias (averaging across word and face bias) were also significantly correlated, $r(33) = .45$, $p < .01$.

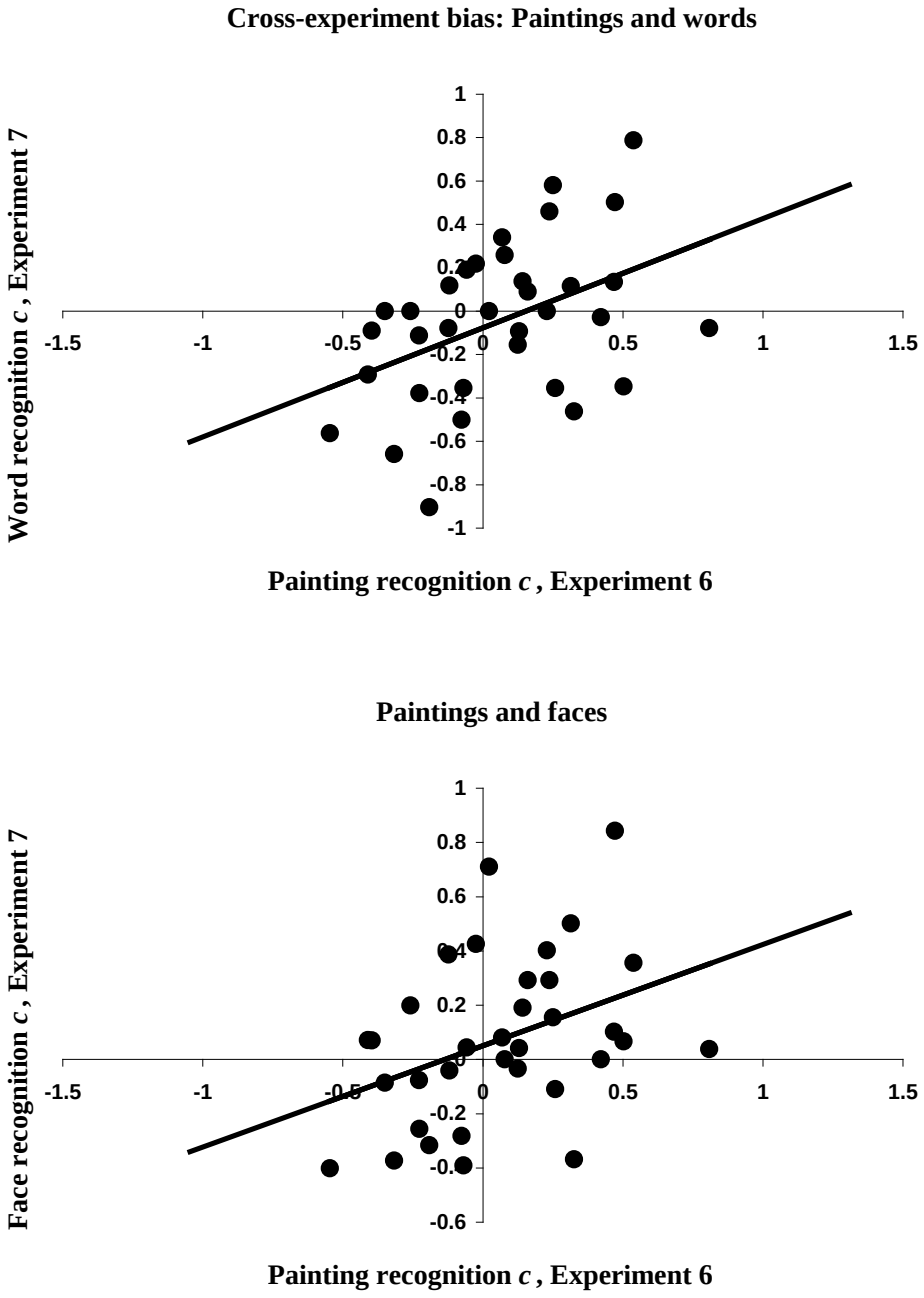


Figure 10. Cross-experiment correlations of recognition bias.

The magnitude of bias shift from Experiment 6 to Experiment 7, calculated as the difference between mean Experiment 7 bias and mean Experiment 6 bias, did not correlate significantly with the number of days between experiments, $r(33) = .16$, $p = .36$.

This result is consistent with the strong within-individual relationship of bias across one week (Experiment 2) in suggesting that time per se does not substantially affect an individual's response bias.

Thus, even recognition tests featuring marked situational contrasts elicit measures of response bias that are stable within an individual. Given the differences in context, materials, and time between the correlated tests, the present analysis provides arguably the strongest current evidence for trait-like consistency in bias.

Experiments 3, 4, 6, and 7: Personality Measures

The rationale for determining the relationships between response bias and established personality traits was two-fold: first, such relationships would suggest that bias possesses the personality trait-like qualities of stability across situations and broad behavioral influence. Second, such relationships would shed light on the nature of this influence. Gillespie and Eysenck's (1980) finding that extraverts show a more liberal response bias than introverts, for example, suggests that a liberal recognition bias is linked to a liberal social bias.

Experiments 3, 4, 6, and 7 included inventories measuring personality traits selected on the basis of their conceptual relevance to response bias. In Experiments 3 and 4, these included a brief index of the "Big Five" personality characteristics (Goldberg, 1992) available as part of the International Personality Item Pool (IPIP; Goldberg, Johnson, Eber, Hogan, Ashton, Cloninger, and Gough, 2006) and the UPPS-P impulsive behavior scale (Whiteside & Lynam, 2001). The Big Five personality characteristics include extraversion, allowing a conceptual replication of Gillespie and Eysenck's (1980) findings and a broad-based assessment of the relationship of response bias to the

dominant personality dimensions (Costa & McCrae, 1992; Goldberg, 1992). The UPPS-P measures five facets of impulsivity: negative urgency, lack of perseverance, lack of premeditation, sensation seeking, and positive urgency. If bias corresponds to a general trait associated with the amount of evidence required before committing to an action (a definition that might be applied to impulsivity as well), a natural prediction is that increasing liberality of response bias should be associated with increasing impulsivity.

Experiments 6 and 7 included the short-form Need for Cognition scale (NFC; Cacioppo & Petty, 1982; Cacioppo, Petty, & Kao, 1984), the Behavioral Inhibition System/Behavioral Activation System scale (BIS/BAS; Carver & White, 1994), and the Maximizing and Regret scales (Schwartz, Ward, Monterosso, Lyubomirsky, White, & Lehman, 2002). The NFC scale measures an individual's tendency to think, to derive enjoyment from thinking, and, conversely, to find dissatisfaction in tasks not involving deep thought. The hypothesis was that participants high in NFC would tend to set a higher evidence criterion in a recognition task, leading to a positive correlation between NFC and conservatism.

The BIS scale measures the extent to which one is sensitive to punishment cues and motivated to avoid punishment; BAS measures the extent to which one is sensitive to reward cues and motivated to seek reward. In a gambling task, for example, a person high on the BAS scale and low on the BIS scale would be expected to make risky decisions in pursuit of reward (e.g., Kim & Lee, 2011), although this pattern is not always observed empirically (e.g., Brand and Altstotter-Gleich, 2008). The inclusion of these scales in the present experiments was based on the possibility that some participants adopt a given valence of bias because they view one type of success (i.e., a hit or a correct rejection) as

more rewarding than another or because they view one type of error (i.e., a false alarm or a miss) as more punishing than another. Participants preferring misses to false alarms but indifferent to the form of their successes, for example, should be conservatively biased. Participants with stronger preferences should exhibit larger biases, with the direction of the bias depending on the preference. If a preference for a given recognition outcome is driven by a sense of punishment (or reward), then individuals high in BIS (or BAS) should exhibit the strongest biases. Thus, the prediction was a curvilinear relationship, with both liberal and conservative recognizers showing high BIS/BAS scores and less biased (more neutral) recognizers showing lower scores.

The Maximization inventory measures one's proneness to carefully consider all options before making a decision (called *maximizing*) versus selecting the first satisfactory option that presents itself (called *satisficing*; Simon, 1956). Intuitively, maximizing represents a more stringent evidence requirement for decision making; thus, the prediction was that maximizers would tend to adopt a more conservative response bias, and satisficers, requiring less evidence before making a decision, would be more liberal recognizers. The Regret scale accompanies the Maximization scale and measures regret proneness (Schwartz et al., 2002). Maximizers are found to be more given to regret than satisficers (Schwartz et al., 2002); like maximization, then, regret was predicted to increase with conservatism in response bias.

Estimates of reliability for each of the personality inventories used in the present experiments were drawn from published validation studies. Reliability varies across the measures but is sufficient to enable correlation with response bias in all cases. For the IPIP test of the Big Five, Cronbach's alphas were .87, .78, .79, .87, and .76 for the

Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Intellect scales, respectively (Gow, Whiteman, Pattie, & Deary, 2005). For the UPPS-P, alpha reliabilities were .87, .89, .85, and .83 for the Premeditation, Urgency, Sensation Seeking, and Perseverance scales, respectively (Whiteside, Lynam, Miller, & Reynolds, 2005). Cronbach's alpha was .90 for the NFC scale (Cacioppo, Petty, & Kao, 1984) and .71 and .67 for the Maximizing and Regret scales, respectively (Schwartz et al., 2002). Finally, alpha levels were .74 for the BIS scale and averaged .72 for the three subscales of BAS (Carver & White, 1994). These reliability estimates were used in the calculations of correction for attenuation reported in Tables 9 and 10.

Method

Participants. One hundred forty-nine participants completed the NEO-FFI and UPPS-P scales as part of Experiments 3 and 4; 85 completed the NFC and BIS/BAS scales as part of Experiment 6; 74 completed the Maximizing and Regret scales as part of Experiment 7.

Materials. The 50-item inventory of Big Five markers was taken from the online International Personality Item Pool (Goldberg et al., 2006; http://ipip.ori.org/New_IPIP-50-item-scale.htm). The 54-item UPPS-P impulsivity inventory was taken from Whiteside and Lynam (2001). The 18-item NFC scale, a validated short form of the original NFC scale developed by Cacioppo and Petty (1982) was taken from Cacioppo, Petty, and Kao (1984). The 7-item BIS and 13-item BAS scales (including, for BAS, the Reward Responsiveness, Drive, and Fun Seeking subscales) were taken from Carver and White (1994). The 13-item Maximization and 5-item Regret scales were taken from Schwartz et al. (2002).

Procedure. The Big Five and UPPS-P inventories were given to participants at the end of Experiments 3 and 4 in a counterbalanced order. They were administered separately but identical instructions and response scales were used with each. Both inventories consisted of short statements such as “Am interested in people” (Big Five) and “I have a reserved and cautious attitude towards life” (UPPS-P). Participants were instructed to answer each question about themselves “as you generally are now, not as you wish to be in the future” and were encouraged to be honest in light of the confidentiality of their answers. Participants responded on a 1-5 scale according to how accurately each statement described them (1 = Very Inaccurate, 2 = Moderately Inaccurate, 3 = Neither Accurate nor Inaccurate, 4 = Moderately Accurate, 5 = Very Accurate). The scale appeared in the middle of the screen at all times with the questions centered above. Responding was self-paced. The questions within each scale were presented in a random order.

The NFC and BIS/BAS scales were administered separately and in a counterbalanced order at the end of Experiment 6; the Maximization and Regret scales were administered in the same fashion at the end of Experiment 7. The procedure for these scales was the same as that of the Big Five and UPPS-P scales except for some slight alterations to the instructions and the use of a 1-9 response scale.

Results and Discussion

Correlations of recognition bias and each subscale of the UPPS-P, overall impulsivity, and the Big Five measures are displayed in Table 9. No significant relationships with bias were observed (all *p*s > .13). Corrections for attenuation are noted

in parentheses beside each value in Table 9. These corrected values did not signal any substantial relationships.

Table 9. Correlations of recognition bias and impulsivity and Big Five personality measures.

	r_{words}	$r_{paintings}$	$r_{overall}$
UPPS-P			
Premeditation	.06 (.08)	.05 (.06)	.04 (.05)
Urgency	-.14 (-.17)	-.02 (-.02)	-.07 (-.09)
Sensation Seeking	.08 (.10)	-.02 (-.03)	.02 (.03)
Perseverance	.10 (.13)	.00 (.00)	.07 (.09)
Overall Impulsivity	.06 (.08)	.00 (.00)	.03 (.04)
Big Five			
Agreeableness	.04 (.05)	-.15 (-.20)	-.08 (-.11)
Conscientiousness	.05 (.07)	-.03 (-.04)	.02 (.03)
Emotional Stability	.12 (.15)	-.02 (-.03)	.07 (.09)
Extraversion	-.09 (-.11)	.01 (.01)	-.05 (-.06)
Intellect/Imagination	-.08 (-.11)	-.03 (-.04)	-.05 (-.07)

Note. Correlations corrected for attenuation appear in parentheses.

Correlations involving the NFC, BIS/BAS, Maximizing, and Regret scales are displayed in Table 10, with corrected correlations in parentheses. Significant and suggestive relationships are highlighted in Figures 11-13. As predicted, participants

higher in Need for Cognition tended to use more conservative recognition criteria (see Figure 11), though this relationship did not reach significance, $r(83) = .18$, $p < .10$.

Table 10. Correlations of recognition bias and NFC, BIS/BAS, Maximizing, and Regret measures.

	$r_{paintings}$		r_{words}	r_{faces}	$r_{overall}$
NFC	.18 (.22)	Maximizing	-.05 (-.07)	-.25 (-.34)	-.15 (-.21)
BAS	-.06 (-.08)	Regret	-.09 (-.13)	-.23 (-.33)	-.15 (-.21)
BIS	-.18 (-.25)				

Note. Correlations in parentheses are corrected for attenuation.

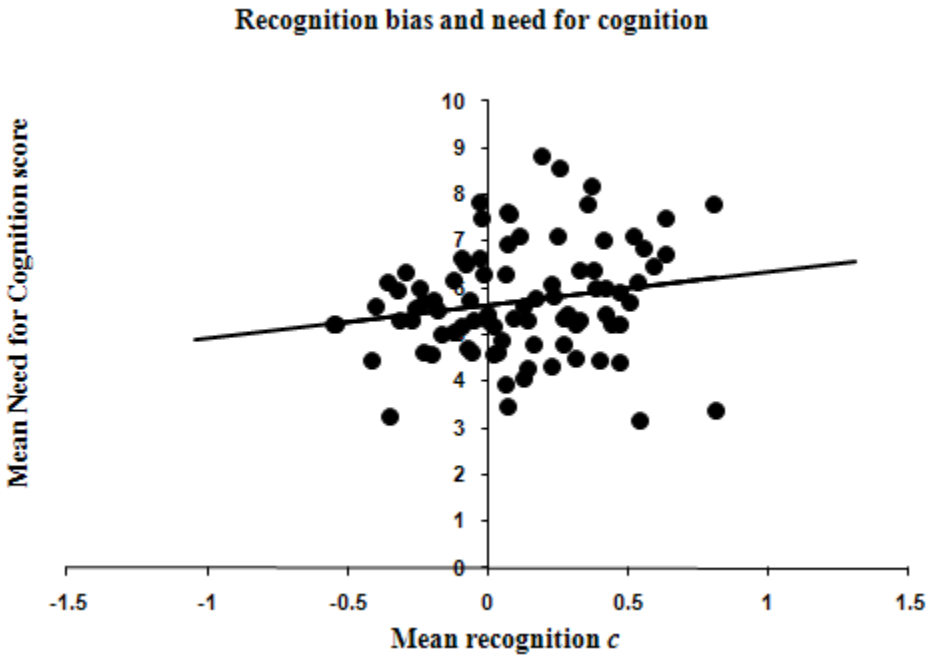


Figure 11. Correlation of recognition bias and scores on the Need for Cognition scale.

Scores on the BAS and BIS scales were not linearly related to bias. While the -.25 adjusted correlation of bias and BIS appears suggestive, examination of the associated scatterplot revealed that the strength of the relationship was driven in large part by four participants with extremely low BIS scores; thus, it is not considered further here. The predicted curvilinear relationship was not evident from examination of the scatterplots. No further BIS/BAS analyses were conducted.

Word recognition bias predicted neither Maximizing nor Regret scores. Interestingly, however, face recognition bias was modestly related to both measures (maximizing: $r(63) = -.22$, $p < .08$; regret: $r(63) = -.20$, $p < .11$; see Figures 12 and 13). Though these correlations are not statistically reliable, the correction for attenuation raises both substantially (to -.34 and -.33, respectively). One final correlation, though not directly involving recognition bias, was worthy of note: Regret scores significantly predicted false lineup identifications, $r(63) = .30$, $p < .05$ (see Figure 14). Contrary to prediction, then, more regret-prone individuals rejected fewer lineups, not more.

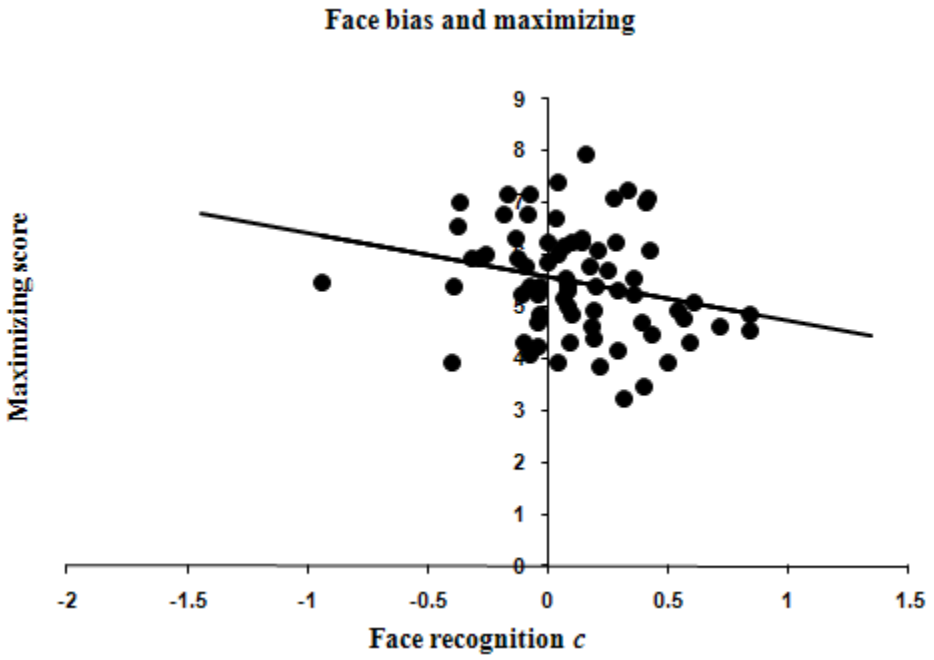


Figure 12. Correlation of recognition bias and scores on the Maximizing scale.

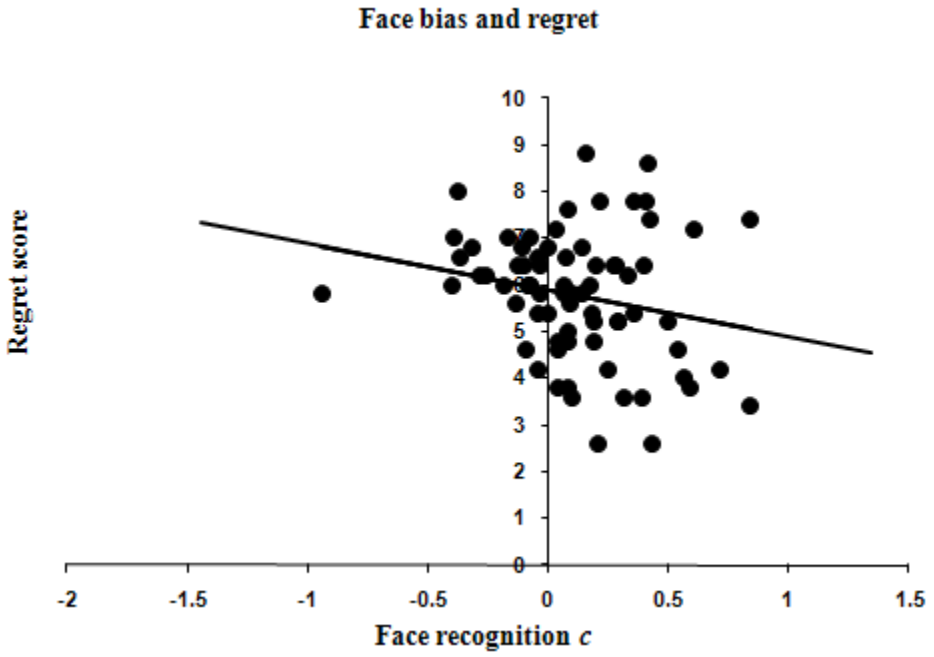


Figure 13. Correlation of recognition bias and scores on the Regret scale.

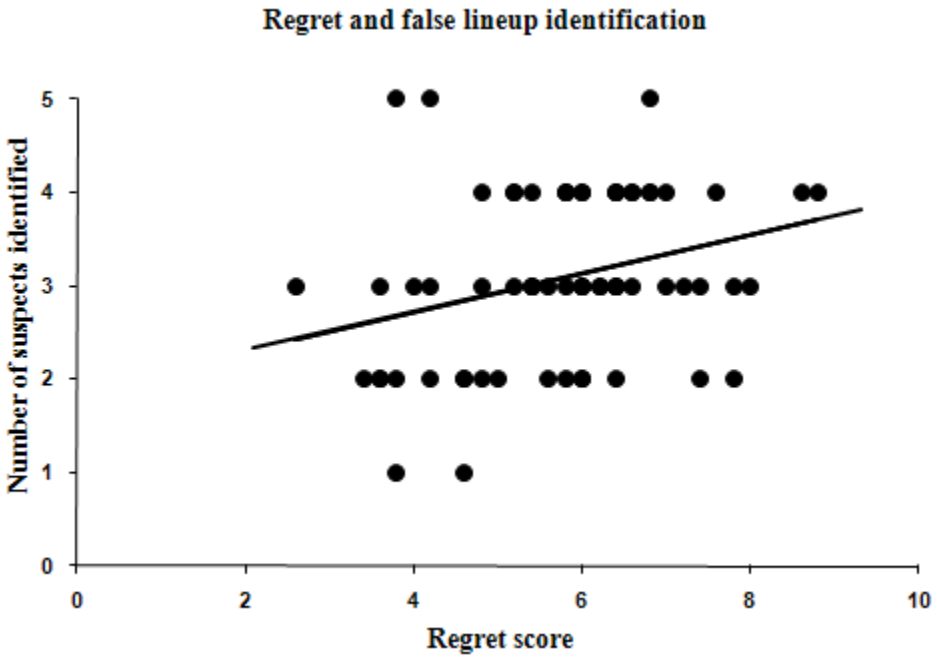


Figure 14. Correlation of scores on the Regret measure and the number of suspects identified in the lineup task, Experiment 7.

If the personality measures used in the present experiments yielded some indications of relationships between recognition bias and broader personality traits, they raised more questions than answers. The tendency for participants higher in need for cognition to exercise a more conservative response bias, for example, is consistent with the characterization of conservative recognizers as more cautious and thoughtful in their approach to tasks, such that they would require more memory evidence before committing to an "old" judgment. It in unclear, however, why liberal recognizers should score higher in maximization, since this measure also taps evidence-gathering tendencies before committing to a decision. Nor is it clear why face bias, but not word bias, predicted maximization. Similarly, the intriguing finding that regret proneness and false lineup identifications were correlated does not present an immediate explanation, although it is consistent with the observed negative correlation between regret and face

bias. The lack of any sizeable relationships between bias and the impulsivity measures is surprising, given that both pertain to an individual's required level of evidence before action is taken. Finally, the Gillespie and Eysenck (1980) finding of a recognition bias-extraversion relationship was not replicated. These findings are considered below.

General Discussion

The construct of response bias is indispensable to recognition memory theory: it provides a means of characterizing numerous phenomena marked by a tendency for the recognition system to change its apparent standard for generating an "old" response without changing its accuracy in doing so. Moreover, bias provides a basis for understanding the recognition decision itself: how it is reached under conditions of uncertainty, and how it is affected by factors unrelated to memory. The present experiments were, in essence, designed to examine whether response bias also provides a basis for understanding individual recognizers. The perspective taken was a departure from that of most previous research on bias: instead of asking what factors influence bias from without (e.g., task and stimulus manipulations), the present work asked to what extent bias is founded within an individual as a stable cognitive trait.

The experiments reported here approached the question of trait bias based on several separate but interrelated markers of within-individual stability: consistency across time, materials, tasks, and situations, correspondence with personality traits, and manifestation in memory judgments of an applied nature. These lines of evidence varied in their strength of support for the notion of bias as a trait, but collectively they provided numerous indications of compelling within-individual consistency.

Consistency across time. Experiment 1 established the foundational observation that one's response bias does not vary freely from one recognition test to the next; rather, bias on an initial test is highly predictive of bias on a subsequent test. Experiment 1 used neutral, everyday words as stimuli and provided no incentive for participants to adopt any particular response criterion. Patterns of responding, then, can be thought to reflect an individual's predisposition to liberality, neutrality, or conservatism. The high correlation between Test 1 and Test 2 bias reflected a robust tendency for this predisposition to remain intact, at least under highly consistent testing circumstances.

Experiment 2 extended this finding by demonstrating a similar correlation of bias on two tests one week apart. As argued above, this similarity is surprising in light of the nature of the intervals separating Tests 1 and 2 in the two experiments. A typical participant in Experiment 2 completed a recognition cycle, left the laboratory, spent a week eating, sleeping, engaging in activities, attending classes, interacting with numerous others, and ranging through moods and stress levels, and then returned to the laboratory and displayed a level of bias consistency typical of participants who had taken Test 1 ten minutes earlier. This finding suggests that time alone does not substantially affect an individual's response bias, another facet of trait-like stability.

Consistency across materials. While the results of Experiment 2 were striking, the conclusions were restricted by the fact that the testing conditions and environment were the same across tests; perhaps these factors, rather than an inherency of bias, produced the observed correlations. Experiment 3, 6, and 7 addressed this possibility. In Experiment 3, significant bias correlations obtained even with substantial differences in the stimuli across two tests (the PW and WP conditions). The magnitude of the

relationship in these conditions (.49 and .45, respectively) was smaller than in Experiments 1 and 2 and the WW condition of Experiment 3 (.69, .67, and .81, respectively), indicating that stimulus match may contribute to bias consistency. However, the presence of large, significant correlations across stimulus classes sharing few common features continued to suggest that participant predisposition accounts for a good deal of the variance in bias.

Experiments 6 and 7 provided further evidence of bias consistency within and across stimulus domains. In Experiment 6, correlations of .55 and .59 between painting tests suggested that the relatively low correlation in the PP condition of Experiment 3 (.39) was anomalous (though Experiment 6 included a training phase and lacked a pure replication condition). In Experiment 7, faces and words showed a strong, positive bias relationship despite possessing few shared features and varying in the degree of match between studied and tested items: studied words were identical at study and test while studied faces displayed different expressions at study and test.

Consistency across situations. Jointly, Experiments 6 and 7 represented the most stringent test of bias consistency. Participants signing up for both experiments completed recognition cycles not only involving highly distinct stimuli (i.e., paintings in Experiment 6, words and faces in Experiment 7) but separated by an average interval of 2.5 weeks as part of what participants believed were two entirely separate experiments. For these participants, time, materials, and situations all varied across the two recognition tests, yet bias was still positively and significantly correlated. Thus, the analysis of joint Experiment 6 and 7 participants revealed the most compelling evidence that recognition

bias is, to an extent, inherent to an individual and independent of prevailing test circumstances.

The observed cross-experiment correlations (.43 between Experiment 6 paintings and Experiment 7 words; .38 between Experiment 6 paintings and Experiment 7 faces) were notably smaller than those within single experiments. That is, when time, materials, and situations did not differ across tests, Test 1 bias was more predictive of Test 2 bias than when these factors did differ. This trend converges with the observation in Experiment 3 that bias was generally less consistent when stimuli differed across tests than when they did not, even though significant stability was evident in both cases. While these comparisons run across experiments and should be interpreted with caution, particularly when the differences between correlation coefficients are not large, they point to a combined influence of time, materials, situations, and individual predisposition on recognition memory response bias. It is not surprising that attributes external to the recognizer should play a role in determining bias; indeed, previous research has identified such attributes (e.g., asymmetrical payoffs, emotionality of stimuli). Moreover, malleability in response bias as conditions change is an adaptive feature of the recognition system, particularly when target and lure frequencies differ (e.g., Kantner & Lindsay, 2010; Rhodes & Jacoby, 2007) or when one type of error is perceived to bear a greater cost than another (e.g., a false alarm versus a miss; Van Zandt, 2000). The central contribution of the current experiments is the suggestion that observed bias is tethered to a predisposition internal to the recognizer, and that a complete account of recognition memory performance must include this factor.

Consistency across tasks. Is the bias observed in recognition tests a special case of a general proclivity to act on the basis of a certain level of evidence? Experiments 2 and 4-6 explored this question by allowing recognition bias to be correlated with performance on non-recognition tasks. These experiments produced mixed results. Bias failed to predict DRM false recall in Experiment 2, but the interpretation of this result was clouded by the fact that many participants had very recently learned of the DRM paradigm in classroom lectures. Experiment 4 was timed to resolve this concern and did show a significant tendency for participants with a more liberal recognition bias to recall more critical lures in the DRM task. These findings are noteworthy in two respects: first, they represent a correspondence of recognition bias with performance on a task outside of recognition memory (i.e., free recall), suggesting a common cognitive substrate. Second, they suggest that liberality in response bias is related to general false memory proneness, which has itself been discussed as a possible trait (e.g., Geraerts et al., 2009).

Recognition bias was not, however, associated with degree of conservatism in answering general knowledge questions. Whether conservatism was operationalized as the use of wide ranges in numerical estimates (Experiment 2) or the frequent use of the "pass" option to avoid the risk of an incorrect response (Experiments 4 and 5), there was no apparent relationship with conservatism in recognition. As noted earlier, these null findings may result from a true lack of relationship between recognition bias and risk tolerance in estimation. Based on the current findings, it might be the case that bias correlates with performance on memory tasks that are episodic in nature (e.g., DRM), but not with other forms of decision making, such as estimation, risk-taking, or GNG (with which bias was uncorrelated in Experiment 6).

However, the null effects may also have arisen from the nature of the general knowledge task. Perhaps this task did not elicit decisions based on a criterial amount of risk because participants did not perceive their decisions as risky. The frequent use of the pass option by many participants and the low confidence associated with answers (approximately 58% in both Experiment 4 and Experiment 5) suggest that participants were well aware of their uncertainty; if they did not perceive risk, then, it was likely because they did not consider errors to be consequential. A straightforward solution would be increasing the gains or losses associated with responses, but a greater impact might be achieved by making participants feel that they have something to lose in the experiment (in Experiments 4 and 5, participants were aware that they stood to lose nothing even in a worst-case scenario). Endowing participants with a sum of money at the start of the experiment and making retention of the sum contingent on performance might provoke loss aversion (Kahneman & Tversky, 1984), increasing the subjective feeling of risk associated with responses. Another possibility is that trial-by-trial feedback of the type found in gambling scenarios would increase investment in the outcome of each response, bringing out individual differences in risk tolerance as a decisional factor.

Relationship to personality characteristics. Personality measures provide an opportunity to discover relationships between response bias and established characteristics of an individual, and some of the correlations observed in the present experiments encourage further investigation of such relationships. However, significant correlations between personality and cognitive variables can be elusive (e.g., Salthouse & Siedlecki, 2007). In the present experiments, some measures expected to correlate with bias did not (e.g., impulsivity, extraversion), while two of the three relationships that did

approach significance were opposite the predicted direction. Given the moderate magnitude of these correlations, it will be important to establish their replicability. Nonetheless, the present data suggest that NFC, maximizing, and regret warrant further investigation as predictors of recognition bias. In addition, some individual differences measures (not tested here) have been associated with DRM performance (e.g., dissociative experiences, fantasy proneness, hypnotizability). Future work should determine whether these measures are also linked to recognition bias.

Relationship to eyewitness memory. Bias in a traditional laboratory-format recognition test correlated significantly with performance in an applied recognition scenario: judging target-absent lineups after watching videos depicting crimes. In Experiment 7, liberal recognizers were associated with an increased frequency of incorrectly identifying lineup members as culprits, suggesting that the willingness of such participants to endorse probes as old on the basis of less memory evidence bears consequences outside of traditional recognition tests.

While the evidence connecting recognition bias with lineup performance should be considered preliminary, it is conceivable that a bias measure taken from a standard recognition test could be used in conjunction with other measures (e.g., a personality profile including the Regret scale) to help determine whether an eyewitness has a predisposition that favors identifying or rejecting lineup members. Such information could not, of course, be used to determine the likelihood of accuracy (because the target-present versus target-absent status of a lineup is not known *a priori*); it could, however, give an indication as to whether a witness is more likely to err by selecting an innocent person (a liberal bias) or by failing to select the guilty party (a conservative bias). The

results of Experiment 7 suggest that face recognition test bias might be a better predictor of lineup bias than word recognition, although the difference in correlations was not large. Importantly, recognition sensitivity did not predict lineup performance. Response bias therefore appears to be the most informative recognition measure for the present purposes.

Implications

The central implication of the thesis presented here is that individuals can be placed along a continuum of recognition bias from more conservative to more liberal, and, critically, that this placement characterizes the *individual*, not just his or her performance on a given test. The idea that people are inherently disposed to some tendency or other is commonplace in the domain of personality psychology; indeed, the very notion of a personality trait implies that people possess enduring attributes that guide their thinking and behavior across different situations. The notion of a memory trait is much less explored. However, the proposal that response bias represents a cognitive trait is related in spirit to work identifying particular populations with especially liberal response bias and especially poor performance in the DRM paradigm (see Previous Evidence Suggestive of Trait Bias). The proposal here is that recognition bias and DRM performance are both partial manifestations of a trait corresponding to the amount of evidence an individual requires to take a given action.

Limitations

A number of limitations of the present work mark avenues for future investigation. Perhaps the most obvious is that the overwhelming majority of the

evidence presented for trait response bias has been correlational. In such situations the burden often remains on the investigator to establish a causal link between the variables in question. In the present work, of course, no implication is made that Test 1 bias *causes* Test 2 bias; rather, the implication is that both Test 1 and Test 2 bias are caused by a common underlying trait. The objective was to test for the presence or absence of a relationship between biases on the two tests so as to establish grounds for inferring the trait. For this initial set of inquiries into trait bias, then, correlations were generally the appropriate measure.

Nonetheless, future work should pursue more direct evidence of the existence of the trait. There at least two general means by which this might be done. First, a functional relationship of response bias with non-recognition tasks or personality traits could be better established by manipulating a correlate of response bias and measuring the effect of the manipulation on bias, or vice versa. Second, the neural correlates of bias could be investigated. Specific proposals for both approaches are discussed in the Future Directions section.

A second limitation concerns the inference of response bias from signal detection measures. As noted earlier, many different combinations of distributional shifts can account for a given pattern of hit and false alarm rates, and assumptions are generally required in order to conclude that sensitivity but not bias (or vice versa) differs across tests, individuals, or conditions. This issue may be highlighted by examining the spread of criterion values in Figure 2. These criteria represent values of *c* and were claimed to reflect the variety in biases that participants display as an empirical regularity. As noted in the introduction, however, it is possible that each participant in that experiment had an

identical response bias but varied in how familiar they tended to find both old and new test items. Individual differences in the global familiarity of test probes would be modeled by signal detection theory as shifts in the location of the old-item and new-item distributions along the evidence continuum. Such differences could predict the same variety of *c* values across participants as actual differences in criterion location across participants.

Unfortunately, signal detection theory offers no definitive means of determining which of the above two scenarios is accurate. Hence, the claim that differences in obtained *c* values reflect differences in bias assumes that the locations of the old and new distributions differ negligibly across participants. This assumption, however, is implicit in virtually all research using signal detection measures. In addition, individual differences in the locations of both distributions would presumably arise from differential experience with or ability to encode stimuli in everyday life, since the new-item distribution contains stimuli not presented earlier in the experiment. Such differences seem unlikely, especially with stimuli as ubiquitous in everyday life as the commonplace words used in several of the present experiments.

Although the concern over extra-experimental sources of variation in distribution locations may be argued to be minimal, experimentally induced variation was more likely, and, as noted earlier, can also obscure the interpretation of signal detection measures. In the case of the PW condition in Experiment 3, for example, mean *c* across participants was highly conservative on Test 1 (with painting stimuli) and neutral on Test 2 (with word stimuli). Because the paintings used in the experiment were certainly much less familiar to participants from extra-experimental sources than the words, it is

reasonable to suppose that distributions for both old and new paintings tended to be farther to the left on the evidence continuum than those for words. As noted above, a criterion placed at a given point along the continuum will yield a more conservative *c* value in the former case than the latter even though bias does not differ across the two stimuli. However, any distributional shifts driven by background familiarity with the materials should have been similar across participants, leaving correlations of bias across tests, the critical measure, unaffected.

Differential shifting of the old-item distribution only was also possible in the present experiments, and might have occurred for at least two types of comparisons: between Test 1 and Test 2 (e.g., because participants utilized an improved encoding strategy in the second study/test cycle) and between participants (e.g., because some participants encoded study items more effectively than others). Correlations of *c* and *d'* were usually very weak in the current experiments, and partial bias correlations controlling for *d'* were used when necessary, but it remains possible that sensitivity influenced obtained values of bias (see Current Experiments: Measurement of Response Bias). If anything, however, a change in d' masquerading as a change in bias would generally have worked *against* the hypothesis of trait bias by producing an underestimation of bias consistency.

A third limitation of the current experiments concerns the ecological validity of the time and context manipulations used to test bias stability. The combination of Experiments 6 and 7, which provided numerous situational changes across which to correlate response bias, was nonetheless confined to a laboratory environment, with attending cross-situational similarities (e.g., procedures, the experimenter-participant

relationship) that may have inflated bias consistency relative to what would be observed in situations posed by everyday life. Similarly, future research should assess within-individual consistency across wider time intervals than those tested in the current experiments. The Experiment 6/7 combination contained the longest average inter-test gap (ranging from a few days to nearly two months across participants); if bias is a trait, however, consistency should be observable over spans of years.

A further concern regarding the external validity of the current work arises from the homogeneity of the samples. While variation in some demographic factors is present in the pool from which participants were drawn (e.g., course of study, ethnicity, country of origin), all participants were University of Victoria undergraduates enrolled in a psychology course and taking part in the experiment for bonus credit, and nearly all were between the ages of 18 and 24. Sample homogeneity is a tolerated limitation of much university-based research in experimental psychology, but is arguably of particular concern in establishing evidence for a generalized cognitive trait. Although the current experiments encourage a conceptualization of response bias as a trait, future tests should include the recruitment of community members varying in age, occupation, and other factors.

Future Directions

Correlational studies such as those reported here should continue to provide insight into the factors that mediate within-individual bias consistency and the range of non-recognition tasks and personal characteristics related to bias. As noted above, however, an important direction for future research is to establish more direct forms of evidence for trait bias. One way of doing so is experimental manipulation: response bias

may be manipulated and the effect on a related task measured, or a correlate of bias may be manipulated and the effect on bias measured. A straightforward method for influencing bias is the use of corrective feedback at test, which can falsely reinforce either “old” or “new” responses (Han & Dobbins, 2008) or correctly tune participants to unequal base rates of old and new items (Estes & Maddox, 1995; Kantner & Lindsay, 2010). After training participants to tighten or relax their recognition criterion, performance on a second task (e.g., DRM recall) could be assessed. The alternative, measurement of bias after the manipulation of a correlated variable, would be possible if a manipulable cognitive measure was found to have a relationship with bias. If a positive mood, for example, was associated with a more liberal recognition bias, a mood manipulation should impact observed bias on a subsequent recognition test (e.g., Rotteveel & Phaf, 2007).

The most direct form of evidence for trait bias might be obtained by investigating the neural markers of individual differences in bias. The neural correlates of bias have been investigated using electroencephalography (EEG) and functional magnetic resonance imaging (fMRI). Results from these techniques have dovetailed with experimental work (reviewed in Previous Evidence Suggestive of Trait Bias) in implicating the PFC in criterion setting. Windmann et al. (2002) divided participants into liberal and conservative groups on the basis of response bias in a recognition task. The critical measure was an event-related potential (ERP) component, maximal over prefrontal scalp locations between 300 and 500 ms post-stimulus onset, thought to index the familiarity of recognition test probes (e.g., Curran, 2000; Yonelinas, Otten, Shaw, & Rugg, 2005). Windmann et al. (2002) found that test items called “old” elicited a larger

familiarity component in conservative recognizers than in liberal recognizers, presumably because the former require more memory evidence (i.e., familiarity) before they will call an item "old." Azimian-Fardiani and Wilding (2006) replicated this result with an instructional manipulation of bias. In an fMRI study, Miller, Handy, Cutler, Inati, and Wolford (2001) found activation in areas of the dorsolateral PFC (as well as other areas) associated with criterion shifting.

The identification of brain regions involved in criterion setting suggests that the neural locus of individual differences in bias should also be identifiable. Given its spatial resolution, brain imaging with fMRI presents a promising means of localizing the origins of trait bias. A straightforward starting point would be an experiment similar to those reported above (e.g., two recognition study/test cycles, perhaps separated by a week) but conducted while participants are in a scanner. Based on previous findings, a positive correlation should be observed between PFC activity and conservatism of response bias on both tests. Of particular interest is whether PFC activity and response bias would be positively correlated across the two tests. The finding that an individual's behavioral and neurobiological indices of response bias on Test 1 predict those indices one week later would provide evidence that trait-like stability in levels of PFC activity underlie intra-individually stable measures of response bias.

More refined hypotheses may then be tested. For example, the fact that response bias is known to be influenced by various task manipulations (e.g., instructional motivation, payoffs) raises the possibility that such task-driven/situational bias and trait/dispositional bias are regulated by distinct cortical regions. The PFC is proposed to serve at least two biasing functions in decision making: an early, "covert" bias, supported

by the ventromedial PFC (VMPFC), that is impenetrable to conscious influence (Bechara, Damasio, Tranel, & Damasio, 1997), and a later bias, supported by the adjacent dorsolateral PFC (DLPFC), that accrues from conscious knowledge of task-appropriate response strategies (Burgess & Shallice, 1996). An intriguing possibility is that these two areas of PFC underlie trait and situational recognition response biases, respectively. Future work could thus test the hypothesis that VMPFC activity during a recognition test will reflect stable individual differences in bias while DLPFC activity will vary according to explicit task demands.

Finally, though the emphasis in the present work has been on a context-independent predisposition to response bias, it is clear from past research and the present experiments that an array of influences – both internal and external to the recognizer – combine to produce the single measure of bias obtained in a given recognition test. Modeling techniques such as multiple regression will be valuable in determining the range of these influences and the extent to which each drives bias. Such factors are likely to include the payoffs and rewards associated with different outcomes, the presence or absence of feedback, the stimuli, the experimental context, state and trait characteristics of the recognizer, and, based on the present research, a predisposition to liberality, conservatism, or neutrality.

Conclusion

Treisman and Williams (1984) stated that “It is regrettable that criterion setting should generally be treated as a given but unexamined fact, because its proper understanding could clarify important psychological aspects of decision in all areas…” (p. 68). The current work was undertaken in the same spirit. Here the study of criterion

setting was approached from an individual differences perspective, based on the largely unexamined phenomenon of wide variation in criterion placement across participants in a recognition task. The basic question of *why* some individuals are conservative and others liberal on a straightforward memory task in which bias is neither encouraged nor inherently advantageous has driven the present research. The results suggest that bias is, at least to an extent, a stable characteristic of a recognizer, and one with implications for recognition theory, personality theory, decision making, eyewitness memory, and general cognition.

References

Ackerman, R., & Goldsmith, M. (2008). Control over grain size in memory reporting--with and without satisficing knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1224-1245.

Azimian-Faridani, N., & Wilding, E. L. (2006). The influence of criterion shifts on electrophysiological correlates of recognition memory. *Journal of Cognitive Neuroscience, 18*, 1075-1086.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science, 275*, 1293-1294.

Benjamin, A., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language, 51*, 159-172.

Beth, E. H., Budson, A. E., Waring, J. D., & Ally, B. A. (2009). Response bias for picture recognition in patients with alzheimer disease. *Cognitive and Behavioral Neurology, 22*, 229-235.

Blair, I. V., Lenton, A. P., & Hastie, R. (2002). The reliability of the DRM paradigm as a measure of individual differences in false memories. *Psychonomic Bulletin & Review, 9*, 590-596.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 105*, 14325-14329.

Brand, M., & Altstötter-Gleich, C. (2008). Personality and decision-making in laboratory gambling tasks--evidence for a relationship between deciding advantageously under risk conditions. *Personality and Individual Differences, 45*, 226-231.

Brewer, N., & Palmer, M. A. (2010). Eyewitness identification tests. *Legal and Criminological Psychology, 15,* 77-96.

Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, word frequency and negative recognition. *The Quarterly Journal of Experimental Psychology, 29,* 461-473.

Brown, S., Steyvers, M., & Hemmer, P. (2007). Modeling experimentally induced strategy shifts. *Psychological Science, 18,* 40-45.

Budson, A. E., Todman, R. W., Chong, H., Adams, E. H., Kensinger, E. A., Krangel, T. S., et al. (2006). False recognition of emotional word lists in aging and alzheimer disease. *Cognitive and Behavioral Neurology, 19,* 71-78.

Burgess, P. W., & Shallice, T. (1996). Response suppression, initiation and strategy use following frontal lobe lesions. *Neuropsychologia, 34,* 263-272.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42,* 116-131.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48,* 306-307.

Carlin, M. T., Toglia, M. P., Wakeford, Y., Jakway, A., Sullivan, K., & Hasel, L. (2008). Veridical and false pictorial memory in individuals with and without mental retardation. *American Journal on Mental Retardation, 113,* 201-213.

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.

Clancy, S. A., McNally, R. J., Schacter, D. L., Lenzenweger, M. F., & Pitman, R. K. (2002). Memory distortion in people reporting abduction by aliens. *Journal of Abnormal Psychology, 111*, 455-461.

Clancy, S. A., Schacter, D. L., McNally, R. J., & Pitman, R. K. (2000). False recognition in women reporting recovered memories of sexual abuse. *Psychological Science, 11*, 26-31.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 33A*, 497-505.

Costa, P. T., Jr., & McCrae, R. R. (1992). *NEO-PI-R: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Curran, T. (2000). Brain potentials of recollection and familiarity. *Memory & Cognition, 28*, 923-938.

Curran, T., DeBuse, C., & Leynes, P. A. (2007). Conflict and criterion setting in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 2-17.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58*, 17-22.

Delis, D.C., Kramer, J.H., & Kaplan, E. (2001). *Delis-Kaplan Executive Function System Examiner's Manual*. The Psychological Corporation, a Harcourt Assessment Co.: San Antonio, TX.

Dobbins, I. G., & Han, S. (2008). What constitutes a model of item-based memory decisions? In B. H. Ross (Ed.), *Skill and strategy in memory use.* (pp. 95-144). San Diego, CA US: Elsevier Academic Press.

Dougal, S., & Rotello, C. M. (2007). 'Remembering' emotional words is based on response bias, not recollection. *Psychonomic Bulletin & Review, 14*, 423-429.

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

Estes, W. K., & Maddox, W. T. (1995). Interactions of stimulus attributes, base rates, and feedback in recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*, 1075-1095.

Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition, 38*, 833-848.

Geraerts, E., Lindsay, D. S., Merckelbach, H., Jelicic, M., Raymaekers, L., Arnold, M. M., et al. (2009). Cognitive mechanisms underlying recovered-memory experiences of childhood sexual abuse. *Psychological Science, 20*, 92-98.

Gillespie, C. R., & Eysenck, M. W. (1980). Effects of introversion–extraversion on continuous recognition memory. *Bulletin of the Psychonomic Society, 15*, 233-235.

Gold, C. A., Marchant, N. L., Koutstaal, W., Schacter, D. L., & Budson, A. E. (2007). Conceptual fluency at test shifts recognition response bias in Alzheimer's disease: Implications for increased false recognition. *Neuropsychologia, 45*, 2791-2801.

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment, 4*, 26-42.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., et al. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84-96.

Goldsmith, M., Koriat, A., & Weinberg-Eliezer, A. (2002). Strategic regulation of grain size memory reporting. *Journal of Experimental Psychology: General, 131*, 73-95.

Gordon, S. K., & Clark, W. C. (1974). Application of signal detection theory to prose recall and recognition in elderly and young adults. *Journal of Gerontology, 29*, 64-72.

Gow, A. J., Whiteman, M. C., Pattie, A., & Deary, I. J. (2005). Goldberg's 'IPIP' Big-Five factor markers: Internal consistency and concurrent validation in Scotland. *Personality and Individual Differences, 39*, 317-329.

Han, S., & Dobbins, I. G. (2008). Examining recognition criterion rigidity during testing using a biased-feedback technique: Evidence for adaptive criterion learning. *Memory & Cognition, 36*, 703-715.

Harkins, S. W., Chapman, C. R., & Eisdorfer, C. (1979). Memory loss and response bias in senescence. *Journal of Gerontology, 34*, 66-72.

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

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

Hockley, W. E. (2011). Criterion changes: How flexible are recognition decision processes? In P. Higham & J. Leboe (Eds.), *Constructions of remembering and metacognition: Essays in honor of Bruce Whittlesea*. Houndmills, UK: Palgrave Macmillan.

Huh, T. J., Kramer, J. H., Gazzaley, A., & Delis, D. C. (2006). Response bias and aging on a recognition memory task. *Journal of the International Neuropsychological Society, 12*, 1-7.

Johansson, M., Mecklinger, A., & Treese, A. (2004). Recognition memory for emotional and neutral faces: An event-related potential study. *Journal of Cognitive Neuroscience, 16*, 1840-1853.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist, 39*, 341-350.

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

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience, 3*, 759-763.

Kim, D., & Lee, J. (2011). Effects of the BAS and BIS on decision-making in a gambling task. *Personality and Individual Differences, 50*, 1131-1135.

Koriat, A., & Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: Distinguishing the accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of Experimental Psychology: General, 123*, 297-315.

Kramer, J. H., Rosen, H. J., Du, A., Schuff, N., Hollnagel, C., Weiner, M. W., et al. (2005). Dissociations in hippocampal and frontal contributions to episodic memory performance. *Neuropsychology, 19*, 799-805.

Lindsay, D. S., & Kantner, J. (2010). *Metamemorial influences on recognition memory response bias*. Paper presented at the 51st Annual Meeting of the Psychonomic Society, St. Louis, MO.

Lindsay, D. S., & Kantner, J. (2011). A search for influences of feedback on recognition of music, poetry, and art. In P. Higham & J. Leboe (Eds.), *Constructions of remembering and metacognition: Essays in honor of Bruce Whittlesea*. Houndmills, UK: Palgrave Macmillan.

Luus, C. A. E., & Wells, G. L. (1994). The malleability of eyewitness confidence: Co-witness and perseverance effects. *Journal of Applied Psychology, 79*, 714-723.

Luus, C. E., & Wells, G. L. (1991). Eyewitness identification and the selection of distracters for lineups. *Law and Human Behavior, 15*, 43-57.

Macmillan, N. A. (1993). Signal detection theory as data analysis method and psychological decision model. In G. Keren, C. Lewis, G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues.* (pp. 21-57). Hillsdale, NJ England: Lawrence Erlbaum Associates, Inc.

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

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin, 98*, 185-199.

Meyersburg, C. A., Bogdan, R., Gallo, D. A., & McNally, R. J. (2009). False memory propensity in people reporting recovered memories of past lives. *Journal of Abnormal Psychology, 118*, 399-404.

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/Revue Canadienne De Psychologie Expérimentale, 55, 162-173.

Miller, M. B., Guerin, S. A., & Wolford, G. L. (in press). The strategic nature of false recognition in the DRM paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Miller, M. B., & Wolford, G. L. (1999). Theoretical commentary: The role of criterion shift in false memory. *Psychological Review, 106*, 398-405.

Moritz, S., Woodward, T. S., Jelinek, L., & Klinge, R. (2008). Memory and metamemory in schizophrenia: A liberal acceptance account of psychosis. *Psychological Medicine: A Journal of Research in Psychiatry and the Allied Sciences, 38*, 825-832.

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

Murchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement, 56*, 63-75.

Parks, T. E. (1966). Signal-detectability theory of recognition-memory performance. *Psychological Review, 73*, 44-58.

Patterson, C. M., & Newman, J. P. (1993). Reflectivity and learning from aversive events: Toward a psychological mechanism for the syndromes of disinhibition. *Psychological Review, 100*, 716-736.

Postman, L. (1982). An examination of practice effects in recognition. *Memory & Cognition, 10*, 333-340.

Qin, J., Ogle, C. M., & Goodman, G. S. (2008). Adults' memories of childhood: True and false reports. *Journal of Experimental Psychology: Applied, 14*, 373-391.

Ratcliff, R., Sheu, C., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99*, 518-535.

Rhodes, M. G., & Jacoby, L. L. (2007). On the dynamic nature of response criterion in recognition memory: Effects of base rate, awareness, and feedback. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 305-320.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 803-814.

Rotello, C. M., & Macmillan, N. A. (2008). Response bias in recognition memory. In B. H. Ross (Ed.), *Skill and strategy in memory use.* (pp. 61-94). San Diego, CA US: Elsevier Academic Press.

Rotteveel, M., & Phaf, R. H. (2007). Mere exposure in reverse: Mood and motion modulate memory bias. *Cognition and Emotion, 21*, 1323-1346.

Salthouse, T. A., & Siedlecki, K. L. (2007). An individual differences analysis of false recognition. *American Journal of Psychology, 120*, 429-458.

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing versus satisficing: Happiness is a matter of choice. *Journal of Personality and Social Psychology, 83*, 1178-1197.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*, 129-138.

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

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117,* 34-50.

Stadler, M. A., Roediger, H. L. III, & McDermott, K. B. (1999). Norms for word lists that create false memories. *Memory & Cognition, 27,* 494-500.

Starns, J. J., Lane, S. M., Alonzo, J. D., & Roussel, C. C. (2007). Metamnemonic control over the discriminability of memory evidence: A signal detection analysis of warning effects in the associative list paradigm. *Journal of Memory and Language, 56,* 592-607.

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1379-1396.

Titus, T. G. (1973). Continuous feedback in recognition memory. *Perceptual and Motor Skills, 37,* 771-776.

Trahan, D. E., Larrabee, G. J., & Levin, H. S. (1986). Age-related differences in recognition memory for pictures. *Experimental Aging Research, 12,* 147-150.

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review, 91,* 68-111.

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

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

Whiteside, S. P., & Lynam, D. R. (2001). The five factor model and impulsivity: Using a structural model of personality to understand impulsivity. *Personality and Individual Differences, 30*, 669-689.

Whiteside, S. P., Lynam, D. R., Miller, J. D., & Reynolds, S. K. (2005). Validation of the UPPS impulsive behaviour scale: A four-factor model of impulsivity. *European Journal of Personality, 19*, 559-574.

Whittlesea, B. W. A. (2002). False memory and the discrepancy-attribution hypothesis: The prototype-familiarity illusion. *Journal of Experimental Psychology: General, 131*, 96-115.

Windmann, S., & Krüger, T. (1998). Subconscious detection of threat as reflected by an enhanced response bias. *Consciousness and Cognition: An International Journal, 7*, 603-633.

Windmann, S., Urbach, T. P., & Kutas, M. (2002). Cognitive and neural mechanisms of decision biases in recognition memory. *Cerebral Cortex, 12*, 808-817.

Wixted, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 681-690.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*(1), 152-176.

Woodard, J. L., Axelrod, B. N., Mordecai, K. L., & Shannon, K. D. (2004). Value of signal detection theory indexes for wechsler memory scale-III recognition measures. *Journal of Clinical and Experimental Neuropsychology, 26*, 577-586.

Worthen, J. B., & Wood, V. V. (2001). Memory discrimination for self-performed and imagined acts: Bizarreness effects in false recognition. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 54A*, 49-67.

Yaniv, I., & Foster, D. P. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *Journal of Experimental Psychology: General, 124*, 424-432.

Yonelinas, A. P., Otten, L. J., Shaw, K. N., & Rugg, M. D. (2005). Separating the brain regions involved in recollection and familiarity in recognition memory. *The Journal of Neuroscience, 25*, 3002-3008.

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin, 133*, 800-832.

Appendix A: Sample word stimuli used in Experiments 1-4 and 7

deer

lamp

shirt

monkey

moon

joke

jets

watch

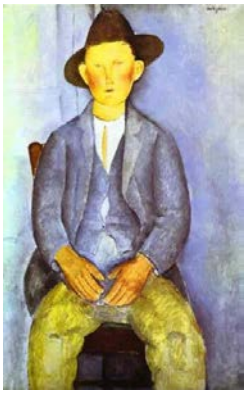
pound

paper

steel

buzzard

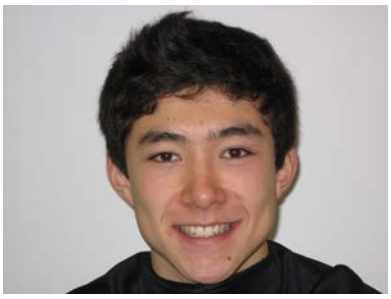
Appendix B: Sample painting stimuli used in Experiments 3 and 6



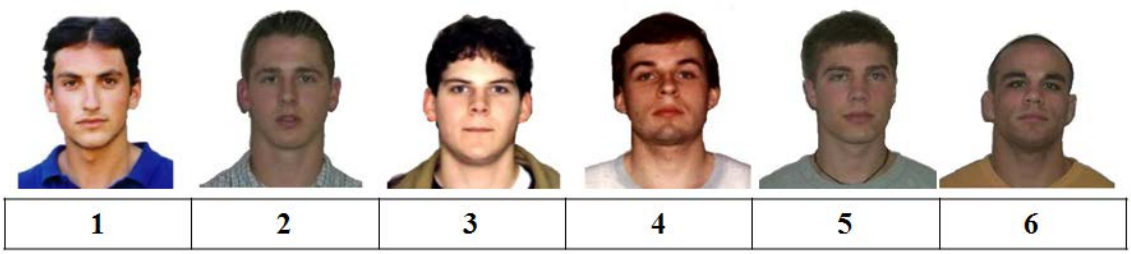
Appendix C: Sample face stimuli used in Experiment 7



Appendix D: Sample still from crime video and lineup used in Experiment 7

Breaking & Entering



1	2	3	4	5	6