Question format shifts bias away from the emphasised response in tests of recognition memory

Ravi D. Mill and Akira R. O'Connor

School of Psychology and Neuroscience, University of St Andrews

Ravi D. Mill, School of Psychology and Neuroscience, University of St Andrews; Akira R. O'Connor, School of Psychology and Neuroscience, University of St Andrews.

Ravi Mill is supported by a PhD studentship from the Scottish Imaging Network: A Platform for Scientific Excellence (SINAPSE).

Correspondence concerning this article should be addressed to Akira O'Connor, School of Psychology, University of St Andrews, St Mary's College, South Street, St Andrews, Fife, KY16 9JP, Scotland, UK. E-mail: aro2@st-andrews.ac.uk

Abstract

The question asked to interrogate memory has potential to influence response bias at retrieval, yet has not been systematically investigated. According to framing effects in the field of eyewitness testimony, retrieval cueing effects in cognitive psychology and the acquiescence bias in questionnaire responding, the question should establish a confirmatory bias. Conversely, according to findings from the rewarded decision-making literature involving mixed incentives, the question should establish a disconfirmatory bias. Across three experiments (*n*s = 90 [online], 29 [laboratory] and 29 [laboratory]) we demonstrate a disconfirmatory bias - "old?" decreased old responding. This bias is underpinned by a goal-driven mechanism wherein participants seek to maximise emphasised response accuracy at the expense of frequency. Moreover, we demonstrate that disconfirmatory biases can be generated without explicit reference to the goal state. We conclude that subtle aspects of the test environment influence retrieval to a greater extent than has been previously considered.

Keywords: episodic memory; recognition; decision-making; evaluation; response bias; goal-directed cognition

1. Introduction

Each time we decide whether or not we recognise something in our environment, the environment itself informs both why we are attempting to tell old from new, and the consequences of making such a decision. When making memory decisions in the real world, it is therefore sensible to consider all aspects of our memory decisions, from cause to consequence, so as to set appropriate goals that enable more strategic use of the available memory evidence. To illustrate this, consider a memory user faced with two tasks: the first being to identify a criminal from a suspect line-up; the second being to decide whether or not to greet a potential acquaintance at the supermarket. If both the suspect and the potential acquaintance evoke equal memory evidence in their respective scenarios, it is conceivable that the markedly different consequences of incorrect “old” decisions in each situation will lead to different memory outcomes. In the line-up the available memory evidence may be deemed insufficient to inform the police that the criminal is present, whereas in the supermarket, the same level of evidence may be more than enough to greet the potential acquaintance. In this way, even subtle manipulations of standard recognition testing environments may influence memory evaluations.

How the testing environment biases memory has been the subject of extensive laboratory investigation, which has fuelled the formalisation of the source monitoring framework – the collection of monitoring/control processes that underpin the evaluation of retrieved memory content (Jacoby, Kelly & McEllrey, 1999; Johnson, Hashtroudi & Lindsay, 1993). For instance, numerous studies have demonstrated influences of the testing environment on the likelihood of endorsing false memories, as manipulated both in the Deese-Roediger-McDermott paradigm (DRM; Deese, 1959; Johnson et al., 1997; Mather, Henkel & Johnson, 1997; Roediger &

McDermott, 1995) and in laboratory analogues of eyewitness testimony (Lindsay & Johnson, 1989; Loftus, Donders, Hoffman & Schooler, 1989; Loftus, 1996). Testing bias effects in the standard single item recognition paradigm have also been observed, albeit with a predominant focus on one particular manipulation – cueing. Cueing participants that a word presented at test is likely to be old or new has been shown to bias decision-making in a confirmatory direction e.g. a “likely old” cue increases the likelihood of “old” responding. This effect is consistent across many methods of cue delivery, ranging from informing participants of the relative proportion of *old* and *new* items at test (Ratcliff, Sheu, & Gronlund, 1992; Van Zandt, 2000), providing valid and invalid performance feedback (Estes & Maddox, 1995; Han & Dobbins, 2009; Rhodes & Jacoby, 2007), and trial-by-trial cueing by text, presentation location and colour (Aminoff et al., 2012; Jaeger, Cox, & Dobbins, 2012; O'Connor, Han, & Dobbins, 2010; Rhodes & Jacoby, 2007). As long as cues can be easily assimilated into the memory decision-making process, they lead to a confirmatory bias that increases endorsements of the cued decision (albeit often to a lesser degree than would be optimal according to the cue; Cox & Dobbins, 2011; Healy & Kubovy, 1978; Wallace, 1980).

Despite the large volume of research on how explicit cues bias memory, whether implicit cues embedded within traditional laboratory-based recognition experiments also introduce bias has remained largely overlooked. This is surprising given the extensive research in analogous applied domains such as questionnaire design (e.g. response acquiescence; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003) and eyewitness testimony (e.g. leading questions; Loftus, 1996) in which strong influences of question format on decision-making have been observed. These effects appear particularly relevant to laboratory-based memory research which often

employs the single item recognition paradigm (often termed yes/no recognition) in which the question “old?” is presented with a to-be-judged item and is responded to with either “yes” or “no” keypresses (e.g. Donaldson, 1996). If this format implicitly emphasises the importance of making an “old” decision, and this implicit emphasis biases memory performance, then this bias is likely presenting itself in most laboratory-based memory research where this canonical “old?”-cued recognition test is employed.

The direction of any bias observed in recognition memory research would provide insight into memory decision-making processes that ultimately determine responding. Two competing predictions of bias direction are apparent, one derived from the laboratory-based memory cueing and applied framing literatures summarised above, and an alternative derived from research into rewarded decision-making. Findings from the cueing/framing literatures suggest that the question “old?” sets up an expectation of encountering an *old* item, thereby leading to a bias *towards* making the emphasised memory decision. A bias in the opposite direction would be anticipated from findings in the rewarded decision-making literature. In mixed incentive research, increasing both the monetary payoff for a correct response and the monetary punishment for an incorrect response, for one of two available response options, motivates a bias against making that monetized response (e.g. Newman, Widom & Nathan, 1985). To clarify, if the aim was to instantiate similar incentives for a single item recognition task, this might involve offering a £1 reward for each correct “old” decision, as well as a £1 punishment for each incorrect “old” decision, whilst not providing any incentives for correct or incorrect “new” decisions. The anticipated effect of this manipulation would be to instil a bias against making an “old” decision (as has been observed previously, Han, Huettel, Raposo, Adcock &

Dobbins 2010). This reluctance is putatively goal-driven, as participants seek to maximise the accuracy of endorsing the decision that is monetarily emphasised as a higher-order goal, even if this entails making it less often (e.g. preferring 10 responses with 90% accuracy to 20 responses with 65% accuracy). If the format of the test question serves to impart similar albeit more implicit emphasis to particular memory decisions, we would hence expect a different bias direction to the cueing hypothesis outlined above: the question “old?” should reduce the likelihood of responding “old”, thereby leading to a bias *against* making the emphasised memory decision.

Below we report three recognition experiments which assess whether the format of the test question biases recognition memory responding, and in what direction. We used both an established measure (signal detection estimates of response criterion, which bin responses by *new* or *old* item status) and a somewhat unorthodox measure (‘decision accuracy’, which bins responses by decision status; Duncan, Sadanand & Davachi, 2012) of recognition performance. Across these experiments, we replicated a novel form of memory bias sensitive to the decision goals implicitly emphasised by the test environment.

2. Experiment 1

We first used an online experiment to explore whether bias in yes/no recognition memory is influenced by the question presented alongside to-be-judged words at test. Following an incidental encoding procedure, the test question was varied on an item-by-item basis, between “old?” (old question emphasis condition) and “new?” (new question emphasis condition), to which participants could respond “yes” or “no”.

2.1. Method

2.1.1. Participants. Participants were 90 self-reported native English speakers who reached the minimum performance threshold of $d' > 0.1$ in each within-subjects experimental condition[1] (58 female; mean age = 34.7; age range = 20 to 66), from a full sample of 120 completing the online experiment (25 were excluded on the basis of being self-reported non-native English speakers[2], 5 for poor performance). Participants were recruited via links to the experiment posted on the laboratory website and social networking sites (i.e. Facebook and Twitter), and were told that they would get feedback on their memory performance (i.e. *d'* and *c* values) as a participation incentive. Informed consent was obtained in accordance with the University Teaching and Research Ethics Committee at the University of St Andrews.

2.1.2. Stimuli. For each participant, a different set of words was randomly sampled from a pool of 2199 singular, common nouns from the English Lexicon Project (Balota et al., 2007), after the removal of low frequency words (Hyperspace Analogue to Language HAL frequency high-pass cut-off: 7.70). This served to exclude highly distinctive items, which could lead to extreme variations in memory strength and consequently reduce the bias effects of primary interest (final word list characteristics: mean HAL frequency = 8.98, mean word length = 7.24, mean number of syllables = 2.43). One hundred and twenty words were used in the single study-test block administered to participants, with 60 words presented at both study and test (*old* words) and 60 words presented only at test (*new* words). The

[1] This performance threshold was maintained across all three experiments.
[2] We allowed all website visitors to take part in the study regardless of their self-reported native English speaking status. This gave us the option to exclude non-native speakers more reliably than if we had gated participation by language status, thereby encouraging misreporting of this demographic information amongst those who wished to take part in the study but were actually non-native English speakers.

experiment was programmed using JavaScript and presented to participants via their internet browsers.

2.1.3. Procedure. Onscreen instructions were followed by a single study-test block. During the study phase, participants counted the syllables in each of 60 serially presented words. For each self-paced study trial, single words were presented above the cue “syllables?” and participants used their mouse to indicate a response by pressing a button corresponding to “1” through “6+” (study phase RT *M* = 2.25s, *SD* = 1.02). A 0.5 s fixation cross preceded each study trial. The test phase immediately followed the study phase. A 0.5 s fixation cross preceded each self-paced test trial (test phase RT *M* = 2.62s, *SD* = 0.75), during which the question for the upcoming trial was presented: either “old?” or “new?”. The fixation was replaced by the to-be-judged word and participants used their mouse to indicate a response by pressing one of two onscreen buttons corresponding to “yes” or “no”. Thirty *old* words and 30 *new* words were presented in each question emphasis condition and stimuli were randomly intermixed throughout the test phase (see Fig. 1a).

2.1.4. Calculation

Our analyses for all reported experiments were conducted on sensitivity (*d'*) and response bias (*c*) parameters, derived from assumptions of the equal variance signal detection model (Green & Swets, 1966; Macmillan & Creelman, 2005). As in Snodgrass and Corwin (1988), a correction for errorless responding was made by taking the numbers of hits (*Hn*; the number of *old* items correctly judged “old”), misses (*Mn*; the number of *old* items incorrectly judged “new”), correct rejections (*CRn*; the number of *new* items correctly judged “new”) and false alarms (*FAn*; the

number of *new* items incorrectly judged “old”), and calculating adjusted hit (*H′*) and false alarm rates (*FA′*) as follows:

$$H' = (Hn + 0.5) / (Hn + Mn + 1) \quad (1)$$

$$FA' = (FAn + 0.5) / (FAn + CRn + 1) \quad (2)$$

These adjusted measures were then used to compute *d′* and *c*:

$$d' = z(H') - z(FA'), \quad (3)$$

$$c = -0.5 \times [z(H') + z(FA')] \quad (4)$$

The signal detection model assumes a continuous normal distribution of memory strength and hence the sensitivity and bias parameters could theoretically span an infinite range of values. However, in practice *d′* typically ranges from 0 to 3 (for above chance performance) and *c* ranges from -1 to 1 (with a value of 0 reflecting unbiased performance). As highlighted in the introduction, the *absolute* value of criterion on the memory strength continuum in any given recognition test reflects the influence of multiple sources of recognition bias. Hence, bias effects specifically attributable to the present manipulation of question format were inferred via scrutiny of the *relative* criterion shifts between question conditions in positive or negative directions (rather than via interpretation of the *absolute* criterion values relative to 0 in each isolated condition). To summarise, larger *d′* parameters indicate increased sensitivity (better discrimination of *old* from *new*) and increasingly positive *c* parameters indicate a more conservative bias (reduced tendency towards responding “old”).

We also conducted more unorthodox analyses of the accuracy of “old” and “new” decisions (old_{corr} and new_{corr} respectively). These decision accuracy measures represent the proportion correct out of all “old” or “new” decisions made by each subject, with responses binned according to “subjective” decision status i.e. the total number of “old” or “new” decisions made (which varied across subjects). This is not to be confused with the standard Hit and Correct Rejection rates which bin responses according to objective “item status” categories i.e. the proportion correct out of all *old* or *new* items presented in the test list (which remained fixed across subjects). As examples of these decision accuracy measures, consider a recognition test phase that presents 30 *old* and 30 *new* items. If Participant A makes a total of 20 “old” decisions during their experimental run, with 15 of these decisions being correct in identifying *old* items, this would lead to an old_{corr} value of .75 (i.e. 15 correct out of a total of 20 “old” decisions). Similarly, if the same participant makes a total of 10 “new” decisions and 9 of these decisions are correct in identifying *new* items, this would lead to a new_{corr} value of .90 (i.e. 9 correct out of a total of 10 “new” decisions). If Participant B also makes 9 correct “new” decisions but attempts a greater overall number of 20 “new” decisions, this would lead to a lower new_{corr} value of .45 (9 correct out of 20 made) – a performance difference with Participant A that would be obscured by comparison of Correct Rejection rates alone (which would be 9 correctly identified *new* items out of a possible 30 for both participants i.e. .30). The decision accuracy measures therefore provide further insight into participants’ subjective experience in making memory decisions under varying question emphasis conditions.

2.2. Results and Discussion

2.2.1. Sensitivity and bias. For all experiments, means and standard deviations are presented in Table 1. We first established that *d'* did not differ across question emphasis conditions, $t(89) = 0.29$, $p = .771$, $d = 0.03$. Participants' sensitivity in discriminating *old* from *new* words was not influenced by the test question.

Bias introduced by question emphasis was then assessed by comparing criterion placement across the two emphasis conditions, each comprising a test wordlist with an equal proportion of *old* and *new* items (see Fig. 1b). Estimates of *c* were significantly higher in the old emphasis than the new emphasis condition, $t(89) = 2.141$, $p = .035$, $d = 0.23$. Participants were biased against endorsing the decision emphasised by the question i.e. the question "old?" reduced the likelihood of responding "old" relative to the question "new?" Notably the *counter-emphasis* direction of this bias contrasts that observed following cueing manipulations, which act to increase endorsement of the cued decision (e.g. O'Connor et al., 2010). Rather these findings favour a goal-driven interpretation, wherein the test question serves to emphasise one memory decision as a goal and participants are consequently more cautious in how they make that emphasised decision.

2.2.2. Decision accuracy analyses. To further elucidate the mechanisms underlying the observed question bias, we also analysed how the accuracy of "old" and "new" decisions varied across emphasis conditions (old_{corr} and new_{corr} respectively; see Fig. 2a). A 2 (decision type: old or new) x 2 (question emphasis: "old?" or "new?") repeated measures ANOVA was conducted on decision accuracy, revealing no significant main effects of decision type and question emphasis, both *F*s < 1. Crucially, the decision type x question emphasis interaction was significant,

$F(1,89) = 6.07$, $p = .016$, $\eta_p^2 = .064$.[3] Question emphasis improved the accuracy of endorsements of the emphasised decision i.e. “old?” improved the accuracy of “old” decisions and “new?” improved the accuracy of “new” decisions (see Table 1 for decision accuracy means). Post-hoc comparisons revealed nonsignificant accuracy effects in counter-emphasis directions for both “new” and “old” decisions, $t(89) = 1.63$, $p = .106$, $d = 0.16$ and $t(89) = 0.94$, $p = .348$, $d = 0.10$ respectively (see Fig. 2a). These findings support the goal-driven interpretation of the observed bias, suggesting that the reduction in endorsements of the decision emphasised by the question may in fact reflect the instilling of a more stringent monitoring strategy that serves to improve the accuracy in making that decision.

_____

TABLE 1 ABOUT HERE

_____

_____

FIGURE 1 ABOUT HERE

_____

_____

FIGURE 2 ABOUT HERE

[3]Given the large online sample recruited, we also conducted the same question emphasis analyses for *d'*, *c* and the decision accuracy measures, after the exclusion of participants whose observed response bias values were identified as potential outliers. Outliers were searched for on the basis of both standard deviation (outliers classified as any response bias value greater than 3 standard deviations from the mean) and interquartile range methods (outliers classified as any bias value exceeding the upper or lower quartiles by a degree of 1.5 times the interquartile range). No outliers were identified via the first method, and exclusion of one potential outlier via the second method did not alter the pattern of results.

---

3. Experiment 2

Given the numerically small effects observed in Experiment 1, we first sought to replicate these exploratory findings in Experiment 2. We also aimed to follow-up observation of the question bias with targeted attempts at clarifying its underlying mechanism. We hence repeated the question emphasis manipulation from Experiment 1 in a more controlled laboratory setting, albeit with the “old?” and “new?” questions varied in block-wise fashion rather than the previously instantiated intermixed approach (to ease participants’ cognitive load during the longer laboratory run time). A levels-of-processing (LOP; Craik & Lockhart, 1972) manipulation was also incorporated to assess whether the question bias is attenuated under conditions of high overall memory strength. Two incidental encoding procedures were employed in separate blocks: a deep LOP pleasantness judgement task; and a shallow LOP letter case judgement task. To clarify, the case judgement task was expected to instil an even lower level of overall memory strength than the syllable-counting task in Experiment 1 (i.e. expected LOP pleasantness > LOP syllable counting > LOP case judgement), and therefore enhance the observed bias. Evidence of bias in shallow but not deep LOP conditions would further indicate that the effect is caused by higher order decision strategies whose influence is greater in situations of low overall memory strength and associated subjective uncertainty (Hirshman, 1995; Kahneman, Slovic & Tversky, 1982).

3.1. Method

3.1.1. Participants. Participants were 29 native English speakers who reached the minimum performance threshold (20 female; mean age = 22.1; age range = 19 to 28)

from a full sample of 31 completing the experiment (2 were excluded for poor performance). Informed consent was obtained in accordance with the University Teaching and Research Ethics Committee at the University of St Andrews and all participants were compensated for their time at the rate of £5/hr.

3.1.2. Stimuli. The pool of stimuli from which words were drawn was identical to that used in Experiment 1. Each of four study-test blocks utilised 120 words (60 *old* and 60 *new* at each test phase) such that participants were presented with 480 words over the course of the experiment. The experiment was presented and responses recorded using PCs running MATLAB (The MathWorks Inc., Natick, MA, 2000) and Psychophysics Toolbox (Brainard, 1997).

3.1.3. Procedure. After the presentation of on-screen instructions and a practice phase, participants completed four self-paced study-test blocks comprising blocked combinations of study LOP (shallow, deep) and test question emphasis (“old?”, “new?”). In the two shallow LOP study phases, 60 serially presented words were shown in either uppercase or lowercase (50% uppercase) below the prompt “uppercase?” and participants used the keyboard to indicate a “yes” or a “no” response (shallow LOP RT M = 1.31s, SD = 0.62). In the two deep LOP study phases, all words were shown in lowercase and the response prompt was “pleasant?” with “yes” and “no” response options (deep LOP RT M = 1.87s, SD = 0.84). Each test phase, comprising 60 *old* and 60 *new* words, immediately followed the preceding study phase. In the two old emphasis condition test phases, single words were presented below the question “old?”, which was onscreen throughout the test block, and participants used the keyboard to give a “yes” or a “no” response (test phase RT M = 1.77s, SD = 0.62). After participants rendered their recognition

decision, they received the prompt “confidence?” and provided a “low”, “medium” or “high” confidence rating using the keyboard. The only difference in the new emphasis condition was that the question “new?” was onscreen throughout the test block. A 0.5 s fixation cross preceded each trial across all study and test blocks. The confidence assessment for each test trial was preceded by a 0.25 s fixation cross.

Across participants, the study-test block order was pseudo-randomised such that only one level of a factor would change in any block transition. Thus, while it was possible for a participant to experience a shallow-“old?” to shallow-“new?” block transition, it was not possible for a participant to experience a shallow-“old” to deep-“new” block transition. Following completion of all four study-test blocks, a short battery of standardized cognitive measures was administered - results from the battery are not presented here.

3.2. Results and Discussion

3.2.1. Sensitivity and bias. To confirm that the LOP manipulation affected memory performance as anticipated, a 2 (LOP: shallow or deep) x 2 (question emphasis: “old?” or “new?”) repeated measures ANOVA was conducted on estimates of *d′* (see Table 1 for condition means). The ANOVA revealed a main effect of LOP, $F(1,28) = 109.14$, $p = .001$, $\eta_p^2 = .796$. As expected, sensitivity was lower in the shallow LOP ($M = 1.33$, $SD = 0.73$) than the deep LOP condition ($M = 2.70$, $SD = 0.52$). There was also an unexpected main effect of question emphasis, $F(1,28) = 5.84$, $p = .022$, $\eta_p^2 = .173$, such that *d′* was higher in the old ($M = 2.10$, $SD = 0.55$) than the new emphasis condition ($M = 1.92$, $SD = 0.57$). There was no significant LOP x emphasis interaction, $F(1,28) = 1.06$, $p = .311$, $\eta_p^2 = .037$.

The primary analysis on *c* was carried out using a 2 (LOP) x 2 (question emphasis) repeated measures ANOVA. We found a main effect of question emphasis on *c*, with significantly higher *c* estimates observed in the old (*M* = 0.09, *SD* = 0.24) than the new (*M* = -0.03, *SD* = 0.32) emphasis condition, $F(1,28) = 6.65$, $p = .015$, $\eta_p^2 = .192$. Although there was no significant interaction effect, $F(1,28) = 3.13$, $p = .088$, $\eta_p^2 = .101$, a numerical trend towards an interaction was observed, and planned pairwise comparisons revealed that the question bias achieved statistical significance only in the shallow LOP condition, $t(28) = 3.05$, $p = .005$, $d = 0.62$, but not in the deep LOP condition, $t(28) = .70$, $p = .487$, $d = 0.13$ (see Figure 1c). These findings replicate the bias observed in Experiment 1, and support our prediction of an attenuated bias under conditions of high overall memory strength.

From the same ANOVA on *c*, we found a main effect of LOP on criterion placement, such that *c* estimates were significantly lower in the deep (*M* = -0.11, *SD* = 0.29) than the shallow LOP conditions (*M* = 0.17, *SD* = 0.30), $F(1,28) = 23.66$, $p = .001$, $\eta_p^2 = .458$. Considered with the prior unexpected effects of emphasis on *d',* this suggests a complex relationship between higher-order and memory-specific processes in the deep LOP condition which may warrant clarification in future research[4].

3.2.3. Decision accuracy analyses. As before, we also analysed how the accuracy of “old” and “new” decisions was influenced by question emphasis, separately for shallow and deep LOP conditions. A 2 (decision type) x 2 (question emphasis) repeated measures ANOVA for the shallow LOP conditions revealed a significant main effect of decision type, $F(1,28) = 9.77$, $p = .004$, $\eta_p^2 = .259$ (see Fig, 2b).

[4] Eliciting response confidence in Experiment 2 and Experiment 3 enabled us to also model unequal variance estimates of sensitivity and response bias (Egan, 1975; Ratcliff, Sheu, & Gronlund, 1992), for which corroborative effects of question emphasis were observed.

Participants' "old" decisions were characterised by greater overall accuracy (*M* = .76, *SD* = .10) than their "new" decisions (*M* = .71, *SD* = .09). Though not of primary relevance to our question bias effects, similar disparities in the response profiles of "old" and "new" decisions have been reported previously (e.g. Jaeger, Cox & Dobbins, 2012). No main effect of question emphasis was observed, $F(1,28) = 1.46$, $p = .237$, $\eta_p^2 = .050$. Crucially however, the decision type x question emphasis interaction was significant, $F(1,28) = 9.04$, $p = .006$, $\eta_p^2 = .244$. This replicated the decision accuracy finding from Experiment 1 - question emphasis improved accuracy in endorsing the emphasised decision i.e. "old?" improved the accuracy of "old" decisions and "new?" improved the accuracy of "new" decisions (see Table 1 for decision accuracy means). Post-hoc comparisons revealed a significant difference in "old" decision accuracy across question conditions, $t(28) = 2.23$, $p = .034$, $d = 0.41$, with a nonsignificant difference obtained for "new" decision accuracy, $t(28) = 0.42$, $p = .679$, $d = 0.07$.

A 2 x 2 ANOVA was also conducted for decision accuracy in the deep LOP condition, yielding a nonsignificant main effect of decision type, $F(1,28) = 3.80$, $p = .061$, $\eta_p^2 = .119$, and a significant main effect of question emphasis, $F(1,28) = 5.65$, $p = .025$, $\eta_p^2 = .168$ (see Fig, 2c). Further, the interaction effect was nonsignificant, $F(1,28) = 1.00$, $p = .326$, $\eta_p^2 = .034$. The question "old?" (*M* = .91, *SD* = .05) hence led to greater decision accuracy than the question "new?" (*M* = .89, *SD* = .05), a finding supported by the previously presented effect of old emphasis in improving *d'* in the same deep LOP condition. Collectively, these findings raise the possibility that the question "old?" instilled a more rigorous source monitoring strategy that prioritised the recovery of recollected content as a basis for responding. In conditions of high overall memory strength (as in the deep LOP condition), this directly

improved performance, given the higher proportion of recollected *old* items likely to be encountered at test. This contrasts the lack of a main effect of emphasis on decision accuracy or *d'* in the shallow LOP condition, where *old* items were less likely to be associated with recollected content and hence the adoption of a monitoring strategy that exclusively prioritised recollection exerted a less beneficial influence on performance.

To summarise, the observed effects of the test question manipulation support the instilment of a goal-driven bias which is particularly prominent under conditions of low memory strength. This reflects the heightened influence of an underlying strategic process that sets a higher threshold when evaluating diagnostic evidence in favour of the decision option emphasised as goal salient. When memory evidence is weakly diagnostic overall, this leads to fewer instances of this heightened emphasis threshold being surpassed and this served to enhance the counter-emphasis behavioural tendency, as indexed by effects on bias and decision accuracy. There also appear to be test question-driven effects on memory sensitivity, raising the potential for question format to impact directly on the assessment of memory evidence. Although not the primary emphasis of this set of experiments, it may be a finding which warrants separate investigation.

4. Experiment 3

The effects of test question on recognition performance in Experiments 1 and 2 have thus far been suggestive of a counter-emphasis bias, wherein one dimension of the test environment, the question, implicitly emphasises a particular memory decision as a salient goal and leads to reduced endorsement of that decision. However, in both experiments, the observed bias effect could alternatively reflect a tendency to

respond “no”, irrespective of the presented question and the implicitly emphasised decision goals (see Fig. 3b).

To adjudicate between these interpretations, we independently manipulated question format (similar to emphasis in Experiments 1 and 2) and response format. The response format manipulation was effectuated by providing only one response option at test, with the alternative response indicated by withholding from using that response option. This is similar to laboratory-based go/no go procedures (e.g. Bruin & Wijers, 2002) and applied eye-witness identification procedures where participants are given the option of not-responding in perpetrator-free line-ups (e.g. Weber & Perfect, 2012) Across study-test blocks, this single response option denoted either a "yes" or a "no" response, and was varied independently of question format. This allowed us to test whether bias shifts according to the decision emphasised as a goal by the combination of the question and response formats, or whether it followed the decision which maps onto a "no" response irrespective of emphasis (see Fig. 3b).

_____

FIGURE 3 ABOUT HERE

_____

_____

FIGURE 4 ABOUT HERE

_____

4.1. Method

4.1.1. Participants. Participants were 29 native English speakers who reached the minimum performance threshold (21 female; mean age = 20.6; age range = 18 to 25) from a full sample of 35 completing the experiment (4 were excluded for failing to understand task instructions, 2 for poor task performance). Informed consent and compensation procedures were identical to those in Experiment 2.

4.1.2. Stimuli. The pool of stimuli from which words were drawn was identical to that used in Experiments 1 and 2. Each of five study-test blocks utilised 120 words (60 *old* and 60 *new* at each test phase) such that participants were presented with 600 words over the course of the study. The experiment was presented and responses recorded using PCs running MATLAB and Psychophysics Toolbox.

4.1.3. Procedure. After the presentation of on-screen instructions and a practice phase, participants completed five study-test blocks. All study phases employed a self-paced case judgement task presented and responded to in a manner identical to that described for Experiment 2 (study phase RT M = 1.09s, SD = 0.58), and were immediately followed by test phases comprising 60 *old* words and 60 *new* words. The first test phase comprised a recognition test with neutral emphasis (no single decision was emphasised by the question or response format) and was used to calibrate the response window in which recognition decisions could be made in subsequent test phases. In this calibration test phase, to-be-judged words were presented below the question “old/new?”, and participants used the keyboard to give an “old” or a “new” response. Following each recognition decision, participants received the prompt “confidence?” and provided a confidence rating of “low”, “medium” or “high”. Both responses were self-paced. The response window for recognition decisions in subsequent test phases was set, on a participant-by-

participant basis, as the 80th percentile of the slowest category of recognition decisions broken down by confidence (typically “low” confidence decisions). This response window ensured that participants had sufficient time to make paced recognition decisions in subsequent test phases (*M* = 3.84 s, *SD* = 1.69).

The next four test phases comprised blocked combinations of question format (“old?” or “new?”) and response format (“yes” keypress or “no” keypress; see Fig. 3a). Question format was manipulated across blocks as described for Experiment 2. Response format was manipulated across blocks by allocating either “yes” or “no” to a single keyboard response. Participants submitted the allocated response with a keypress or endorsed the opposing response by withholding the keypress for the duration of the paced trial. To prevent preferential keypress responding to hasten completion of the task, the second stage of responding was initiated after the full response window had elapsed, irrespective of whether a keypress was made or withheld. Subsequently, participants received the prompt “confidence?” and provided a self-paced confidence rating of “low”, “medium” or “high”. Participants could alternatively make a “discard” response if they wanted their previous recognition decision to be ignored. This option was intended for when participants had failed to render a response due to an attentional lapse and prevented such trials from being coded as deliberately withheld responses. A 0.5 s fixation cross preceded each recognition assessment and a 0.25 s fixation cross preceded each confidence assessment.

Across the four study-test blocks, participants completed two test phases with an *old* decision emphasis (“old?”-“yes” and “new?”-“no”), and two test phases with a *new* decision emphasis (“old?”-“no” and “new?”-“yes”). The four tests were presented in a

pseudorandomised order that minimised the combined switching of question and response formats across study-test blocks.

4.2. Results and Discussion

4.2.1. Sensitivity and bias. A 2 (response format: "yes" keypress or "no" keypress) x 2 (question format: "old?" or "new?") repeated measures ANOVA on estimates of *d'* revealed no significant main or interaction effects, all *F*s < 1. As expected, sensitivity was unaffected by the test format manipulations.

For the main analysis, a 2 (response format) x 2 (question format) repeated measures ANOVA on response criterion *c* found no main effects of response or question format, $F(1,28) = 3.73$, $p = .064$, $\eta_p^2 = .117$ and $F < 1$ respectively. Crucially, the interaction of response format and question format was significant, $F(1,28) = 8.50$, $p = .007$, $\eta_p^2 = .233$ (see Table 1 for means). Planned pairwise comparisons revealed significant increases in *c* (i.e. a shift towards conservative "old" responding) when old was emphasised by the combination of response and question formats (see Fig. 3c). For the "yes" response format, *c* was placed higher for "old?" than "new?" questions, $t(28) = 2.41$, $p = .023$, $d = 0.45$. For the "no" response format, *c* was lower for "old?" than "new?" questions, $t(28) = 2.77$, $p = .010$, $d = 0.52$. Note that in the "no" response format condition, old emphasis is imparted by a "new?" question and new emphasis is imparted by an "old?" question (see Fig. 3a for further details). These findings replicate the biases observed in Experiments 1 and 2, and suggest that the effect is indeed driven by which decision is emphasised as a goal, and not by a general tendency to respond "no".

4.2.3. Decision accuracy analyses. We also analysed how the accuracy of "old" and "new" decisions was influenced by question emphasis, which are presented

separately for each response format condition. A 2 (decision type) x 2 (question format) repeated measures ANOVA for the “yes” response condition yielded no significant effects of decision type, $F(1,28) = 1.98$, $p = .171$, $\eta_p^2 = .066$, and question format, $F < 1$. Importantly, the decision type x question format interaction was significant, $F(1,28) = 14.19$, $p = .001$, $\eta_p^2 = .336$, with “old?” improving the accuracy of “old” decisions and “new?” improving the accuracy of “new” decisions (see Fig. 4a; see Table 1 for means). Post-hoc comparisons revealed a numerical difference approaching significance in “old” decision accuracy across question conditions, $t(28) = 2.01$, $p = .054$, $d = 0.37$, with a nonsignificant difference in the same emphasis-consistent direction obtained for “new” decision accuracy, $t(28) = 0.48$, $p = .637$, $d = 0.09$.

A 2 (decision type) x 2 (question format) repeated measures ANOVA for the “no” response condition yielded no significant main effects of decision type and question format, both Fs < 1. Crucially, the decision type x question format interaction was significant, $F(1,28) = 14.19$, $p = .001$, $\eta_p^2 = .336$, with “new?” improving the accuracy of “old” decisions and “old?” improving the accuracy of “new” decisions (as mentioned before, this bias is in a goal-driven direction, given the inverse emphasis imparted by the question in the “no” response condition; see Fig. 4b). Post-hoc comparisons revealed an emphasis-consistent trend approaching significance for new decision accuracy across question conditions, $t(28) = 2.04$, $p = .051$, $d = 0.39$, with a nonsignificant effect in the same emphasis-driven direction obtained for old decision accuracy, $t(28) = 1.12$, $p = .271$, $d = 0.21$. The decision accuracy findings from Experiments 1 and 2 were hence replicated in both response format conditions - question emphasis improved the accuracy of endorsing the emphasised decision.

5. General Discussion

Across three recognition experiments, the format of the test question acted as a consistent source of bias in memory decision-making. Experiment 1 recruited a large online sample and provided some preliminary evidence of bias effects introduced by the test question typically presented in recognition research. Experiment 2 effectuated a more targeted laboratory follow-up to these exploratory findings, by demonstrating that the question bias is most prominent under conditions of low memory strength. Experiment 3 tested two competing explanations for the bias, finding evidence that it was driven by a goal-driven mechanism rather than a preference for “no” responses. In all three experiments, we replicated the novel finding of a counter-emphasis bias – a decrease in endorsement of the memory decision emphasised by the memory test itself. Our analyses involving the more unorthodox decision accuracy measures also recovered a consistent effect of question emphasis improving the accuracy of endorsements of the emphasised decision. Considered with the response bias effects, the pattern of results suggests that the counter-emphasis serves to improve the accuracy of making emphasised responses, at the expense of their frequency. We now discuss these results in light of previous findings in the fields of memory and decision-making.

Prior studies observing effects of test format on memory performance have often employed recognition paradigms primarily tailored for the study of false memories, such as the Deese-Roediger-McDermott paradigm (DRM; Deese, 1959; Johnson et al., 1997; Mather, Henkel & Johnson, 1997; Roediger & McDermott, 1995). Further, those studies that have investigated effects of test format in the standard single item recognition paradigm have typically focused on measures of source memory or

subjective experiences of memory (Bastin & Van der Linden, 2003; Dodson & Johnson, 1993; Hicks & Marsh, 1999; Khoe, Kroll, Yonelinas, Dobbins, & Knight, 2000; Marsh & Hicks, 1998). For example, Hicks and Marsh (1999) reported an increase in "old" responding in a recognition test phase comprising three response options with two "old" subcategories ("remember *old*"/"know *old*"/"*new*") compared to a test phase with two equally weighted response options ("*old*"/"*new*", followed by "remember"/"know"; though cf. Bruno & Rutherford, 2010). Here we present evidence of a test format bias in the primary measure of recognition performance, which manifests in the absence of unequally weighted response options. Nevertheless, our findings coalesce with previously reported test format biases in false memory and source memory paradigms, in highlighting the salient influence of monitoring and control processes in constraining the outcomes of memory evaluations (Johnson et al., 1993; Jacoby et al., 1999).

Further, the observed bias was in the opposite direction to previously reported test format biases, such as the acquiescence bias in questionnaire responding (Podsakoff, et al., 2003), the Loftus framing effect (Loftus, 1996) and decision cueing in recognition research (e.g. Egan, 1958; O'Connor, et al., 2010; Ratcliff, et al., 1992; Rhodes & Jacoby, 2007). We suggest that our counter-emphasis bias differs from cue-driven biases in that it establishes a decision goal against which retrieved memory evidence is evaluated, rather than contributing to the evidence itself. This aspect also differentiates the present question effects from the classic Loftus framing effect, in which questions that framed details of potentially encoded contexts ("leading questions") led to a confirmatory response tendency that enhanced fallacious endorsement of contexts as having been previously experienced (Loftus, 1996). Our questions lacked any such allusion to explicit aspects of encoding, which

would putatively impact upon memory decision-making by directly interfering with the assessment of memory evidence. Rather, our question manipulations impacted via a more meta-memorial pathway, by implicitly emphasising a particular memory decision as a higher order goal, the endorsement of which demanded a higher level of memory evidence relative to the non-emphasised decision. Our findings therefore inform a tentative distinction between three major strategic biases wrought by different aspects of the retrieval environment on memory decision-making: decision cueing (confirmatory), evidence framing (confirmatory) and goal emphasis (disconfirmatory). Future research will be needed to corroborate this distinction.

The present effects hence reveal a strategic influence on conscious memory use consistent with findings in the rewarded decision-making literature, particularly mixed incentive research. In mixed incentive research, monetising one of two available response options leads to a reluctance to endorse the monetised decision (Newman et al., 1985). A similar disconfirmatory bias was reported in a prior recognition experiment that differentially provided mixed incentives for “old” or “new” decisions (Han et al., 2010). The present findings are notable for inducing comparable caution without the provision of incentives, and demonstrate that even in the absence of any explicit mention of reward, participants display a goal-directed bias that is driven by implicit information gleaned from the testing environment. Future research is necessary to test the correspondence of the present question emphasis effects with more explicit emphasis mechanisms involving incentives.

We also found evidence that the counter-emphasis bias, whilst decreasing the frequency of the emphasised decision, improved the accuracy of these decisions when they were made. Most manipulations enacted solely at retrieval have little or

no impact on *old/new* discrimination sensitivity, which is largely determined by encoding processes (Craik & Lockhart, 1972; Stretch & Wixted, 1998; albeit with some exceptions; Dobbins & McCarthy, 2008; Marsh & Hicks, 1998). Indeed, across all three experiments our retrieval manipulations had little effect on measures of *d′* sensitivity that binned responses by 'objective' item categories and collapsed across old and new decisions. However, the reported effects on decision accuracy (old_{corr} and new_{corr}; measures that binned responses by 'subjective' decision categories) highlight the efficacy of these more unorthodox analyses in elucidating a strategic effect which would otherwise be obscured by standard *d′* sensitivity analysis. Analogous influences of the test environment on the strategies adopted in monitoring retrieved memory evidence have been highlighted previously in the study of source monitoring and false memories (Johnson et al., 1997; Lindsay & Johnson, 1989; Mather et al., 1997). In the present context, it furthers our suggestion of a goal-directed mechanism underlying the counter-emphasis bias, in that participants were seemingly prompted by the question to adjust their response criterion towards a more rigorous evaluation of the retrieved memory evidence supporting the emphasised decision. The question therefore served to instil the goal of getting a greater proportion of the emphasised decisions correct, even if that entailed making fewer endorsements of that decision.

Whilst the counter-emphasis bias was reliably elicited across three experiments, further research is undoubtedly necessary to corroborate our interpretations of its underlying mechanism. Beyond the highlighted need to test both the proposed differentiation of retrieval format biases and the correspondence between implicit and explicit forms of goal emphasis, future research might also seek to clarify whether *old* and *new* items are differentially biased by goal emphasis mechanisms

(as has been recently suggested e.g. Kafkas & Montaldi, 2014), as well as the potential mediating role of individual differences in bias tendency (Aminoff et al., 2012; Kantner & Lindsay, 2012). Another intriguing avenue for future research could examine participants' awareness of the variations in question emphasis, and how manipulating this awareness (e.g. by dividing attention at test) impacts on the evoked bias. Nonetheless, the reliable effect of question format on recognition bias presented here reflects the role that subtle aspects of the test environment can have on the precise quantification of established memory measures. Importantly for laboratory-based memory research, we show that the canonical yes-no memory experiment itself provides a source of bias which has hitherto received scant systematic investigation.

As with many psychological phenomena, it is useful to think of the findings reported here as functional or adaptive when placed in the context of the real world in which we make almost all of our memory decisions. An eyewitness confronted with a line-up of potential suspects is cautious of making an "old" decision as this initiates an action sequence that might end with the prosecution of the identified individual; the alternate "new" decision is comparatively less consequential, as no conclusive step is taken to end the ongoing suspect search. Alternatively, a guard tasked with ensuring security at restricted event may be cautious about making a "new" decision, as this initiates apprehension of the identified individual, whereas rendering the alternate "old" decision maintains the status quo. In sum, the present findings add to emerging research highlighting the close coupling of memory and environmental factors which may be the reason the evaluation is being made in the first place. Above all, it is necessary to remember that memory function, like any cognitive function, can only be isolated from the environment in which it is carried out with

great difficulty. Even when we believe we have this isolation, the inferences our participants make which carry over from their adaptive use in hundreds of day-to-decisions, may have pervasive and unexpected influences which it is important to investigate and understand.

References

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

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445-459. doi: 10.3758/BF03193014

Bastin, C., & Van der Linden, M. (2003). The contribution of recollection and familiarity to recognition memory: a study of the effects of test format and aging. *Neuropsychology, 17*(1), 14-24. doi: 10.1037//0894-4105.17.1.14

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

Bruin, K. J., & Wijers, A. A. (2002). Inhibition, response mode, and stimulus probability: A comparative event-related potential study. *Clinical Neurophysiology, 113*, 1172-1182.

Bruno, D., & Rutherford, A. (2010). How many response options? A study of remember-know testing procedures. *Acta Psychologica, 134*(2), 125-129.

Cox, J. C., & Dobbins, I. G. (2011). The striking similarities between standard, distractor-free, and target-free recognition. *Memory and Cognition, 39*(6), 925-940. doi: 10.3758/s13421-011-0090-3

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684. doi: 10.1016/S0022-5371(72)80001-X

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58*(1), 17-22.

Dobbins, I. G., & McCarthy, D. (2008). Cue-framing effects in source remembering: A memory misattribution model. *Memory & Cognition, 36*(1), 104-118. doi: 10.3758/MC.36.1.104

Dodson, C. S., & Johnson, M. K. (1993). Rate of false source attributions depends on how questions are asked. *The American journal of psychology*, 541-557.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition, 24*(4), 523-533. doi: 10.3758/BF03200940

Duncan, K., Sadanand, A., & Davachi, L. (2012). Memory's penumbra: episodic memory decisions induce lingering mnemonic biases. *Science, 337*(6093), 485-487. doi: 10.1126/science.1221936

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

Egan, J. (1975). *Signal detection theory and ROC analysis*. New York: Academic press.

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

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Han, S., & Dobbins, I. G. (2009). Regulating recognition decisions through incremental reinforcement learning. *Psychonomic Bulletin & Review, 16*(3), 469-474. doi: 10.3758/PBR.16.3.469

Han, S., Huettel, S. A., Raposo, A., Adcock, R. A., & Dobbins, I. G. (2010). Functional significance of striatal responses during episodic decisions: recovery or goal attainment? *Journal of Neuroscience, 30*(13), 4767-4775. doi: 10.1523/jneurosci.3077-09.2010

Healy, A. F., & Kubovy, M. (1978). The effects of payoffs and prior probabilities on indices of performance and cutoff location in recognition memory. *Memory and Cognition, 6*(5), 544-553. doi: 10.3758/BF03198243

Hicks, J. L., & Marsh, R. L. (1999). Remember-know judgments can depend on how memory is tested. *Psychonomic Bulletin & Review, 6*(1), 117-122. doi: 10.3758/BF03210818

Hirshman, E. (1995). Decision processes in recognition memory: criterion shifts and the list strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(2), 302-313. doi: 10.1037//0278-7393.21.2.302

Jacoby, L. L., Kelley, C. M., & McElree, B. D. (1999). The role of cognitive control: Early selection versus late correction. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (Vol. xii, pp. 383-400). New York, NY, US: Guilford Press.

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

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*(1), 3-28.

Johnson, M. K., Nolde, S. F., Mather, M., Kounios, J., Schacter, D. L., & Curran, T. (1997). The similarity of brain activity associated with true and false

recognition memory depends on test format. *Psychological Science, 8*(3), 250-257. doi: 10.1111/j.1467-9280.1997.tb00421.x

Kafkas, A., & Montaldi, D. (2014). Two separate, but interacting, neural systems for familiarity and novelty detection: A dual-route mechanism. *Hippocampus*. doi: 10.1002/hipo.22241

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgement under uncertainty: Heuristics and biases*. New York, NY: Cambridge University Press.

Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition, 40*(8), 1163-1177. doi: 10.3758/s13421-012-0226-0

Khoe, W., Kroll, N. E. A., Yonelinas, A. P., Dobbins, I. G., & Knight, R. T. (2000). The contribution of recollection and familiarity to yes-no and forced-choice recognition tests in healthy subjects and amnesics. *Neuropsychologia, 38*(10), 1333-1341. doi: 10.1016/s0028-3932(00)00055-5

Lindsay, D. S., & Johnson, M. K. (1989). The eyewitness suggestibility effect and memory for source. *Memory & Cognition, 17*(3), 349-358.

Loftus, E. F., Donders, K., Hoffman, H. G., & Schooler, J. W. (1989). Creating New Memories That Are Quickly Accessed and Confidently Held. *Memory & Cognition, 17*(5), 607-616. doi: 10.3758/Bf03197083

Loftus, E. F. (1996). *Eyewitness testimony*: Harvard University Press.

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

Marsh, R. L., & Hicks, J. L. (1998). Test formats change source-monitoring decision processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(5), 1137. doi: 10.1037//0278-7393.24.5.1137

Mather, M., Henkel, L. A., & Johnson, M. K. (1997). Evaluating characteristics of false memories: Remember/know judgments and memory characteristics questionnaire compared. *Memory & Cognition, 25*(6), 826-837. doi: 10.3758/Bf03211327

Newman, J. P., Widom, C. S., & Nathan, S. (1985). Passive avoidance in syndromes of disinhibition: psychopathy and extraversion. *Journal of Personality and Social Psychology, 48*(5), 1316-1327. doi: 10.1037//0022-3514.48.5.1316

O'Connor, A. R., Han, S., & Dobbins, I. G. (2010). The inferior parietal lobule and recognition memory: Expectancy violation or successful retrieval? . *Journal of Neuroscience, 30*(8), 2924-2934. doi: 10.1523/JNEUROSCI.4225-09.2010

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879. doi: 10.1037/0021-9010.88.5.879

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing Global Memory Models Using ROC Curves. *Psychological Review, 99*(3), 518-535. doi: 10.1037/0033-295X.99.3.518

Rhodes, M. G., & Jacoby, L. L. (2007). On the dynamic nature of response criterion in recognition memory: effects of base rate, awareness, and feedback. *Journal of Experimental Psychology Learning Memory and Cognition, 33*(2), 305-320. doi: 10.1037/0278-7393.33.2.305

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

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of Measuring Recognition Memory: Applications to Dementia and Amnesia. *Journal of Experimental Psychology: General, 117*(1), 34-50. doi: 10.1037/0096-3445.117.1.34

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(6), 1397-1410. doi: 10.1037/0278-7393.24.6.1397

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

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

Weber, N. & Perfect, T. (2012). Improving eyewitness identification accuracy by screening out those who say they don't know. *Law and Human Behavior, 36*(1), 28-36. doi: 10.1037/H0093976.

Tables

Table 1.

		Experiment 1		**Experiment 2**				**Experiment 3**			
Encoding task		syllables		case		pleasantness		case			
Question format		old	new	old	new	old	new	old	new	old	new
Response format		-	-	-	-	-	-	yes	yes	no	no
Emphasis		**old**	**new**	**old**	**new**	**old**	**new**	**old**	**new**	**new**	**old**
H	*M*	.77	.79	.64	.70	.92	.91	.69	.72	.74	.68
	SD	.15	.12	.17	.13	.08	.08	.13	.15	.14	.14
CR	*M*	.81	.79	.81	.74	.89	.86	.75	.71	.69	.73
	SD	.10	.12	.12	.15	.08	.09	.13	.16	.16	.15
d'	*M*	1.71	1.73	1.37	1.28	2.83	2.57	1.26	1.28	1.26	1.18
	SD	0.61	0.58	0.70	0.87	0.63	0.54	0.55	0.82	0.80	0.66
c	*M*	0.06	-0.00	0.26	0.07	-0.09	-0.13	0.12	0.00	-0.07	0.08
	SD	0.36	0.37	0.41	0.26	0.34	0.32	0.38	0.34	0.30	0.35
old_{corr}	*M*	.81	.80	.78	.73	.90	.87	.74	.71	.70	.72
	SD	.09	.10	.11	.12	.07	.07	.11	.12	.12	.12
new_{corr}	*M*	.79	.80	.71	.71	.92	.91	.70	.71	.72	.68
	SD	.10	.09	.10	.11	.07	.07	.08	.10	.12	.10

Note: *M*, mean; *SD*, standard deviation.

Figures

Figure 1.

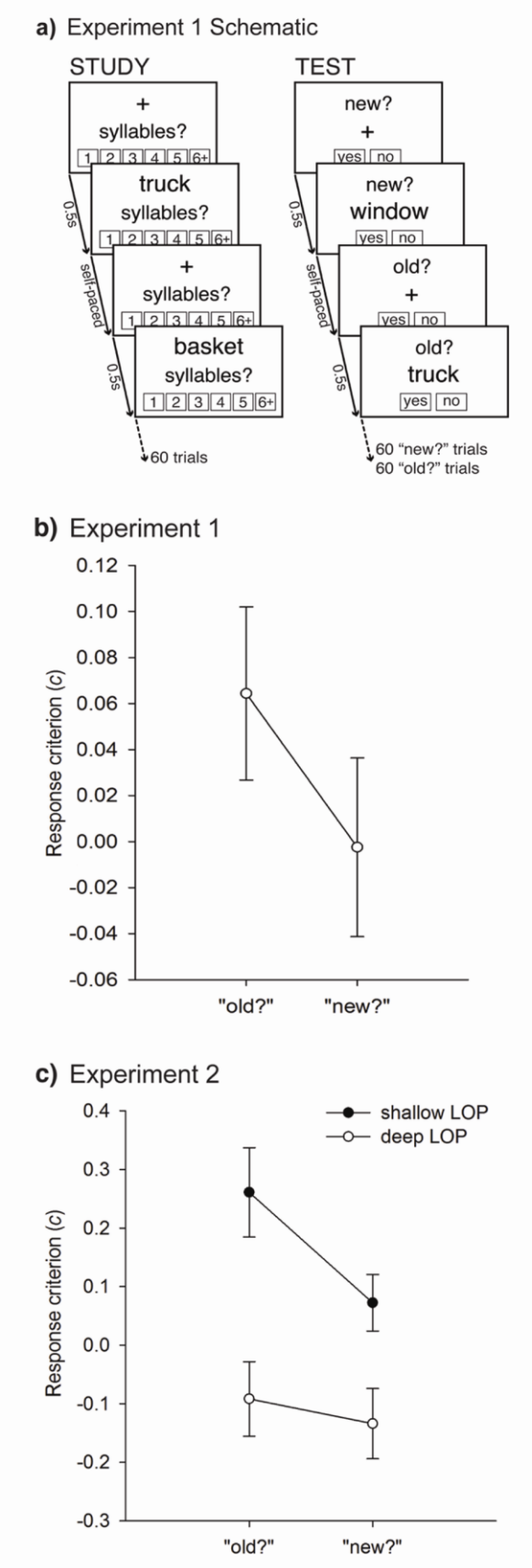


Figure 2.

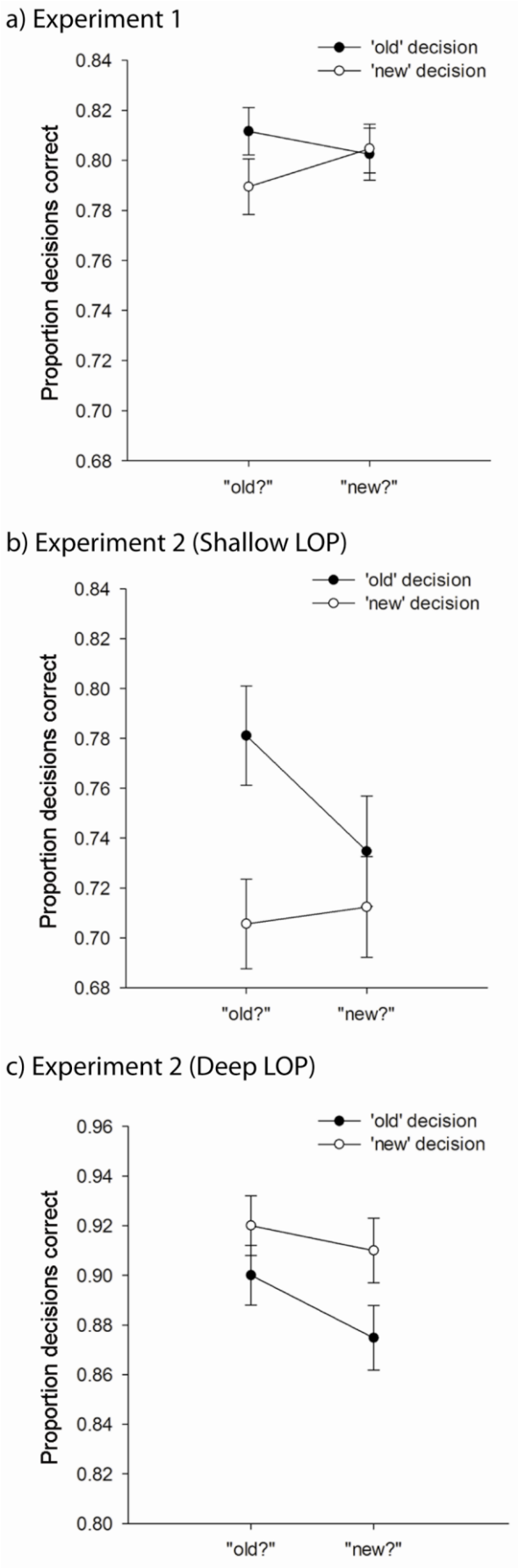

a) Experiment 1
'old' decision
'new' decision
Proportion decisions correct
0.84
0.82
0.80
0.78
0.76
0.74
0.72
0.70
0.68
"old?"
"new?"
b) Experiment 2 (Shallow LOP)
'old' decision
'new' decision
Proportion decisions correct
0.84
0.82
0.80
0.78
0.76
0.74
0.72
0.70
0.68
"old?"
"new?"
c) Experiment 2 (Deep LOP)
'old' decision
'new' decision
Proportion decisions correct
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.82
0.80
"old?"
"new?"

Figure 3.

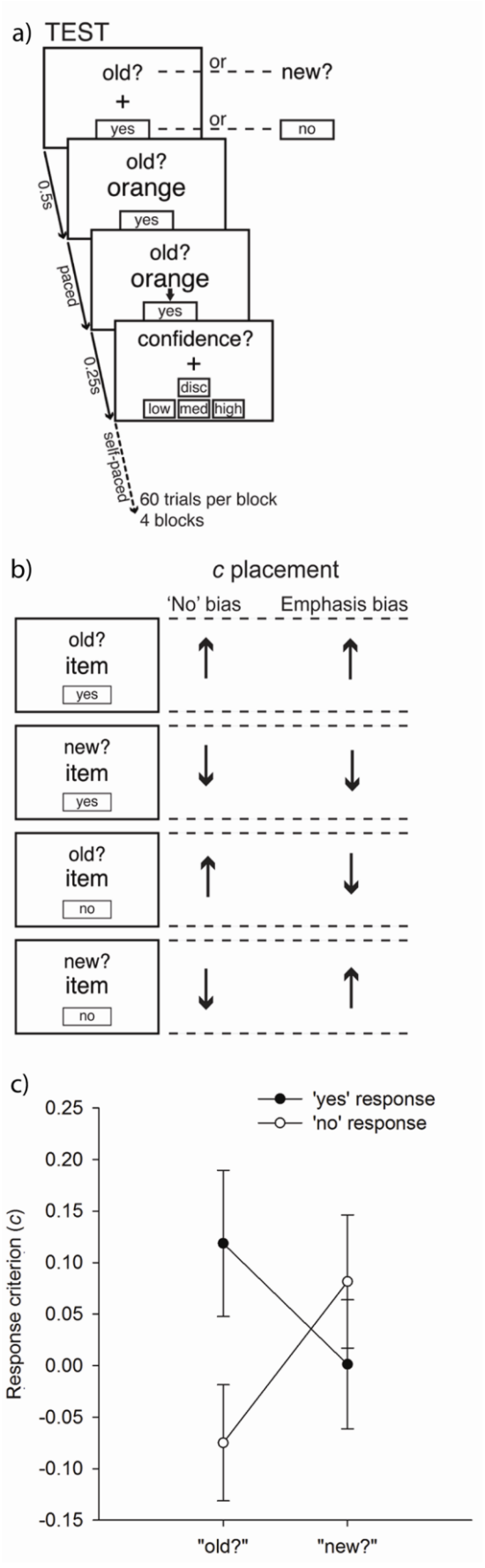
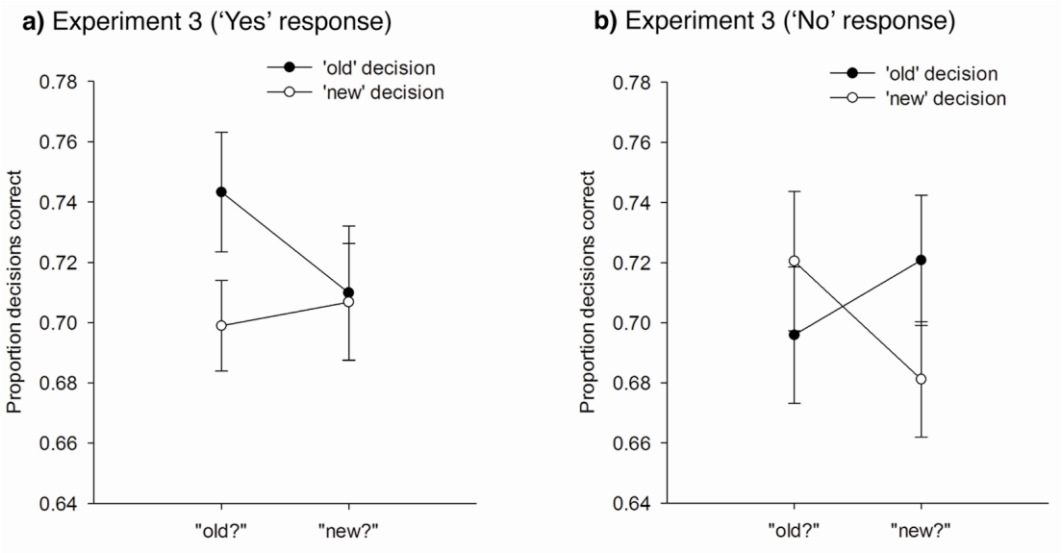

a) TEST
old? + yes
or
new?
or
no
old? orange
yes
old? orange
yes
confidence? +
disc
low med high
0.5s
paced
0.25s
self-paced
60 trials per block
4 blocks
b)
c placement
'No' bias
Emphasis bias
old? item yes
new? item yes
old? item no
new? item no
c)
'yes' response
'no' response
Response criterion (c)
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
"old?"
"new?"

Figure 4.


a) Experiment 3 ('Yes' response)
'old' decision
'new' decision
Proportion decisions correct
0.78
0.76
0.74
0.72
0.70
0.68
0.66
0.64
"old?"
"new?"
b) Experiment 3 ('No' response)
'old' decision
'new' decision
Proportion decisions correct
0.78
0.76
0.74
0.72
0.70
0.68
0.66
0.64
"old?"
"new?"

Captions

Table 1. Design key and descriptive statistics for unadjusted hit rate (*H*), unadjusted correct rejection rate (*CR*), sensitivity (*d'*), response bias (*c*), "old" decision accuracy (old_{corr}) and "new" decision accuracy (new_{corr}), as observed across all experiments.

Figure 1. a) Design schematic for Experiment 1 showing study and test phases. b) Experiment 1 results for response criterion (*c*) across question emphasis conditions. c) Experiment 2 results for *c* across question emphasis conditions, with separate lines denoting LOP condition. Error bars represent standard error of the mean.

Figure 2. Results of the accuracy by decision type analyses for: a) Experiment 1, b) the Experiment 2 shallow LOP condition and c) the Experiment 2 deep LOP condition. Separate lines denote decision type ("old" or "new"; responses binned by decision status) and error bars represent standard error of the mean.

Figure 3. a) Design schematic for Experiment 3 test phase. Within each 120-trial test block, participants responded to words presented alongside a question (either "old?" or "new?"; varied blockwise) by endorsing or withholding a single response (either "yes" or "no"; varied blockwise). The response window for endorsing or withholding recognition responses was equated, and calibrated on a participant-by-participant basis. After each recognition response, confidence was assessed on a three-point scale, with an additional "discard" option allowing participants to discard missed responses. b) Schematic representation of criterion placement (*c*) as generated by

competing bias predictions for Experiment 3. Arrows denote deviations from optimal criterion placement at 0, with increasing *c* associated with decreased “old” responding (‘conservative’ “old” responding) and decreasing *c* associated with increased “old” responding (‘liberal’ “old” responding). According to the “no” attraction prediction, participants consistently prefer responding “no”, and hence *c* should decrease when an “old” decision maps to the “no” response option (“new?”-“yes” and “new?”-“no” test blocks), and increase when it maps to the ‘yes’ response option (“old?”-“yes” and “old?”-“no” test blocks). According to the counter-emphasis prediction, the *combination* of the question and response formats serves to emphasise one class of decision, and participants decrease endorsement of that decision. Hence, *c* should increase when “old” decisions are emphasised (“old?”-“yes” and “new?”-“no” test blocks) and decrease when “new” decisions are emphasised (“new?”-“yes” and “old?”-“no” test blocks). c) Experiment 3 results for *c* across question format conditions, grouped by response format conditions. Error bars represent standard error of the mean.

Figure 4. Results of the accuracy by decision type analyses for the two response format conditions in Experiment 3: a) “Yes” response condition and b) “No” response condition. Note that in the ‘no’ response condition, emphasis is reversed i.e. “old?”-“no” emphasises “new” decisions and “new?”-“no” emphasises “old” decisions. Separate lines denote decision type (old” or “new”; responses binned by decision status) and error bars represent standard error of the mean.