Characterizing the Materials-Based Bias Effect: A Robust yet Mysterious Conservative Response Bias in Recognition Memory for Paintings

by

Kaitlyn Fallow
BA, from the University of New Brunswick, 2012
BSc, from the University of New Brunswick, 2012

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

MASTER OF SCIENCE

in the Department of Psychology

Supervisory Committee

Characterizing the Materials-Based Bias Effect: A Robust yet Mysterious Conservative Response Bias in Recognition Memory for Paintings

by

Kaitlyn Fallow
BA, from the University of New Brunswick, 2012
BSc, from the University of New Brunswick, 2012

Supervisory Committee

Dr. D. Stephen Lindsay, Department of Psychology
Supervisor

Dr. Michael Masson, Department of Psychology
Co-Supervisor

Abstract

Supervisory Committee
Dr. D. Stephen Lindsay, Department of Psychology
Supervisor
Dr. Michael E. J. Masson, Department of Psychology
Co-Supervisor

A series of recognition memory experiments using masterwork paintings and words are reported in which participants were reliably conservative in endorsing images of paintings as "studied". The current paper establishes the historical context of this materials-based bias effect (MBBE) and presents two new experiments aimed at characterizing the underlying mechanisms. Nine previous experiments are reviewed to illustrate the MBBE's robustness to various encoding and test manipulations and the insufficiency of two prior hypotheses in accounting for its origins. Meta-analyses of response bias and sensitivity and analysis of these measures by test quartile are presented and discussed along with receiver operating characteristics and response time data for all of these experiments.

In one new experiment, the response scale on the recognition test was modified to allow participants to choose from not only "studied" or "not studied" options, but also options indicating uncertainty due to the similarity among test items. The hypothesis that these similarity/confusability-related responses would be chosen more for paintings was not supported. A second new experiment aimed to better characterize the time course of the MBBE by implementing a 1-s respond deadline, which was hypothesized to reduce the effect, but this hypothesis was also not supported. Results of all experiments are discussed in the context of unequal variance and dual process models of recognition memory.

Table of Contents

List of Figures

Acknowledgments

I owe a great deal of gratitude to my supervisor, Dr. Steve Lindsay, for his guidance, patience, understanding, and support throughout this project and my MSc program in general. His enthusiasm for research is contagious and inspiring, and I'm incredibly fortunate to have him as a mentor. I would also like to thank my co-supervisor, Dr. Michael Masson, for his kindness and sound statistical advice; his ability to straightforwardly convey difficult ideas and understand what aspects I'm struggling with before I even say it is remarkable and has triggered many a light-bulb moment. Finally, I wish to thank Justin Kantner, who has been involved in this research line since its inception and could always be relied upon to answer questions, offer a new idea, or track down an elusive data file.

Chapter 1: Introduction and Background

The most common approach to testing recognition memory comprises two phases. In the study or learning phase, participants are asked to study a series of stimuli such as words, photographs, or sounds, following which they proceed to the test phase, wherein they view some combination of studied and novel items and are asked to make some kind of old/new judgment for each. Recognition memory data have been approached from a vast range of theoretical and modeling perspectives, including choice theory (Luce, 1959), a variety of threshold theories (Egan, 1958; Luce, 1963), and diffusion modeling (Ratcliff, 1978), to name a few. Many of these strategies and/or variants thereof remain in current use (e.g., Starns, Ratcliff, & McKoon, 2012), but the predominant approach with two-phase recognition experiments is to conceptualize the data in terms of signal detection theory (SDT; Green & Swets, 1966), which has left its mark on the vocabulary used to describe various measures of performance. The standard approach is to calculate participants' hit (correct "old" responses) and false alarm (incorrect "old" responses) rates, which are then typically used to calculate two types of derived scores – sensitivity and response bias – that characterize different aspects of performance.

Sensitivity (sometimes called discrimination) represents the extent to which a subject was successful in endorsing old items (i.e., achieving hits) and rejecting new ones (i.e., avoiding FAs) on a recognition memory test (Macmillan & Creelman, 2005). This is probably most often the measure of primary interest in studies of recognition memory and the factors that influence it. However, researchers who disregard the other type of derived score – response bias – risk missing potentially interesting and important effects. Measures of response bias characterize a subject's general tendency toward calling items "old" or "new," irrespective of accuracy. In SDT terms, response bias is thought to represent the decision criterion, which can be roughly imagined as a point along a continuum of strength of evidence of "oldness" above which an individual is willing to endorse an item as having been studied. Response bias can be neutral, meaning the participant shows no consistent proclivity toward choosing one response over the other; liberal, meaning the participant tends to call items "old" and produces a response profile comprising mostly hits and false alarms; or conservative, reflecting a tendency toward rejecting items and producing mostly correct rejections and misses. The direction and extent of response bias can differ among individuals, experimental conditions, and item classes, even when sensitivity is effectively equivalent.

Variables that Influence Response Bias

Experimental Factors

There are several well-established means of influencing response bias via experimental manipulation, perhaps the simplest of which is to explicitly ask participants to be more or less cautious or lenient in endorsing old items; evidence that individuals can readily adhere to such instructional motivation dates back to work undertaken by Egan (1958). Another possibility is to vary test list composition by increasing or decreasing the proportion of old items, manipulations which will tend to produce more liberal and more conservative responding, respectively (Healy & Kubovy, 1978; Van Zandt, 2000). A third consistently effective method is to manipulate the payoff structure of the experiment by, for example, offering a greater reward – or alternatively, a lesser punishment – for one type of response versus another. Such manipulations affect bias in a predictable way; offering participants a greater reward for correct rejections than for hits will typically encourage a conservative approach, while reversing this reward structure will likely lead to more liberal responding (e.g., Van Zandt, 2000).

Other manipulations have been shown to affect response bias in somewhat less intuitive ways. For example, increasing the delay between study and test is generally associated with more liberal responding (Gehring, Toglia, & Kimble, 1976), while altering or removing stimulus context between study and test can increase conservativeness (Feenan & Snodgrass, 1990).

Individual Differences

An observed relationship between some factor and response bias or criterion shifts may not necessarily generalize across groups of individuals. To complicate things even further, response bias – like most dependent variables – is often analyzed as a mean calculated across participants, but there is evidence to suggest that some of the between-subjects variability in bias may in fact be predictable, and that attributing such differences to mere chance fluctuations is an oversimplification that may obscure interesting effects. A recent study by Kantner and Lindsay (2012) found within-individual response bias to correlate strongly across recognition memory tasks separated by ten minutes or a week as well as between tasks using different stimuli, suggesting that some element of trait-like stability may underlie a given individual's response bias in any experiment, independent from any effects of experimental conditions.

In a related vein, Aminoff and colleagues (2012) reported intriguing individual differences with respect to the willingness to shift criterion between tests as appropriate. By manipulating the proportion of targets (i.e., old items) on the test, the authors created conditions under which liberal (70% targets) or conservative (30% targets) biases would produce higher performance. Participants exhibited remarkable variability in the degree to which they were able to shift criteria, with some individuals shifting appropriately and others shifting excessively or not at all. Within individuals, however, the extent and direction of criterion shifting appeared somewhat stable, showing a significant correlation between word and face recognition tests (Aminoff et al., 2012).

Beyond such baseline differences, factors like aging and disease can influence response bias as well. Alzheimer's patients, for example, tend to show a markedly liberal response bias (e.g., Balota, Burgess, Cortese, & Adams, 2002). Findings for healthy older adults have been comparatively mixed; some studies have found normal aging to be associated with increasingly liberal responding (e.g., Huh, Kramer, Gazzaley & Delis, 2006), but Marquie and Baracat (2000) reported increasingly conservative bias with age, although only for the most highly educated adults in their study. Interestingly, when conditions are set up such that Alzheimer's patients and healthy older and younger adults are matched with respect to discrimination – either by manipulating list length (Budson, Wolk, Chong & Waring, 2006) or increasing the study/test delay (Deason, Hussey, Ally & Budson, 2012) – all groups tend to exhibit a comparably equivalent liberal bias.

Item-Related Effects

Even if individuals show some consistency with respect to response bias and shifts therein across experiments using different types of stimuli, stimulus characteristics themselves are far from trivial in this regard. Myriad item attributes have been found to be related to bias in recognition memory experiments, and unlike manipulations such as varying the proportion of old items that influence response bias in intuitively sensible ways, item effects on bias are often difficult to explain. For example, Brodeur, Chauret, Dion-Lessard, and Lepage (2011) found that participants responded more liberally to figures and photos that were symmetrical relative to their asymmetrical counterparts. The authors reported a number of other esoteric associations between stimuli and bias, such as more liberal biases for figures with lower "evocative scores" and greater contour length, and more conservative responding to small and meaningful figures relative to large and meaningless figures, respectively (Brodeur et al., 2011).

Emotion – perhaps better characterized as a factor bridging both item and individual characteristics than a straightforward stimulus attribute – has been a topic of considerable interest in the memory literature, and its possible role in recognition decision processes is no exception. As noted by Dougal and Rotello (2007), investigations into the effects of emotion on sensitivity have yielded mixed results and there is, thus far, no real consensus as to the role, if any, of emotional valence and arousal in determining recognition accuracy. In contrast, the authors' review of the existing literature found that all recognition studies in which emotion was manipulated yielded corresponding response bias differences (Dougal & Rotello, 2007). This is not to say that such effects are entirely understood (and in fact, as mentioned below, there have been cases since wherein no such differences were found), but their consistency relative to effects on sensitivity measures emphasizes the importance of considering response bias whether or not it is central to the research question. When emotional effects on bias are found, they tend to be in the form of more liberal responding to emotional than neutral stimuli, particularly when the emotional stimuli are associated with negative arousal (e.g., Dougal & Rotello, 2007).

Windmann and Chmielewski (2008) also found some evidence for this tendency to adopt a lower criterion with emotionally salient stimuli. Their participants responded significantly more liberally to emotionally laden words than to neutral words; interestingly, however, the authors did not find the same effect for photographs depicting emotional and neutral facial expressions (Windmann & Chmielewski, 2008). Beth and colleagues (2009) compared response bias for words and pictures more directly in both healthy older adults and Alzheimer's patients, and found that although both groups showed better discrimination for pictures of common objects than the corresponding words, only the healthy group exhibited a more liberal bias when the stimuli were pictures. While a substantial number of studies support the observation that pictures tend to be more memorable than words, an effect often referred to as “picture superiority”, literature regarding response bias differences between the two types of stimuli is comparatively scant, and – as the experiments to be described in the current paper have demonstrated with remarkable consistency – a clear picture has yet to emerge.

Unexpected Response Bias in an Accuracy Feedback Study

Findings reported by Lindsay and Kantner (2011) further illustrate the complexity of the mosaic of possible influences on response bias as well as the importance of considering the possibility of bias effects, even when they are not the focus of interest. The authors designed a series of experiments aimed at investigating the influence of accuracy feedback on recognition memory for complex and relatively novel stimuli, specifically poetry, paintings, and Korean melodies. The results were inconclusive with respect to the effects of feedback on recognition accuracy, but 26 of 32 tests revealed a directionally conservative response bias, and in 22 of these cases this bias differed significantly from zero (Lindsay & Kantner, 2011). This effect was especially marked for paintings, for which response bias was conservative in all experiments for both feedback and control groups (see Figure 1 for these results collapsed across feedback and control groups; figures referred to in this chapter are more thoroughly explained in Chapter 2).

The Materials-Based Bias Effect (MBBE)

Comparing Response Bias for Paintings and Words

The unanticipated finding that response bias in several recognition memory experiments using paintings as stimuli was overwhelmingly and consistently conservative motivated Lindsay and Kantner (2011) to compare paintings directly with words in a within-subjects design. Words are probably the stimulus type used most often in studies of recognition memory, and although manipulations such as those outlined in the previous section can certainly yield bias differences in such studies, words in general are not associated with any consistent response bias pattern. Average response bias in this initial within-subjects experiment (Experiment 1) was again significantly conservative for paintings, while words yielded a liberal bias (Figure 2). However, sensitivity was also higher for paintings than words in this study (Figure 3). Although the measures of sensitivity (d') and response bias (c) the authors used are statistically uncorrelated (e.g., Macmillan & Creelman, 2005) and there was therefore no obvious reason to assume the bias effect was attributable to differences in sensitivity, Lindsay and Kantner launched a second experiment (Experiment 2) with some of the most subjectively distinctive paintings removed in an attempt to make discrimination more comparable for paintings and words.

This follow-up experiment and a series of others have, without exception, replicated this conservative bias for paintings when participants study and are tested on these stimuli, whether paintings are the only item type (as in the five experiments originally described by Lindsay and Kantner, and Experiments 8-9 in this paper) or are intermixed with words (as in this paper's Experiments 1-7, one group in Experiment 8, and Experiments 10-11). This effect has been termed the Materials-Based Bias Effect (MBBE), a name that reflects both the differences in bias between paintings and words and the conservatism seen in experiments with only paintings (Lindsay & Kantner, 2011). This designation also allows for the likely possibility that the effect is not exclusive to paintings but will eventually be found to extend to other stimuli that share some critical attribute(s).

The majority of the MBBE experiments conducted thus far have shared a common overall structure. This is described in more detail in Chapter 2, but generally speaking, the experiments all included a study phase comprising randomly intermixed words and paintings and a test list, also randomized, containing all studied items and an equal number of new items of each type. In other words, the test phase in each experiment included 25% studied paintings, 25% new paintings, 25% studied words, and 25% new words, such that the "ideal" response bias would be none at all, and there was no obvious incentive to lean toward one response or the other. In the two between-subjects MBBE studies (Experiments 8 and 9), the overall setup was the same, but stimuli were either all words or all paintings. These experiments have yielded varying patterns of sensitivity differences between paintings and words, but the response bias pattern is unwavering and has persisted despite various manipulations at study and test. The mechanism underlying the effect, however, remains elusive; several hypotheses regarding the origin of conservatism on paintings have been tested, but none has received strong support.

The Subjective Memorability Hypothesis

One seemingly plausible explanation was that people expect paintings to be more memorable than words and therefore set a higher criterion for endorsing them as old. This possibility was not only intuitively compelling, but also has precedent in the recognition memory literature dating back to work conducted by Brown and colleagues (Brown, 1976; Brown, Lewis, & Monk, 1977; but see Wixted, 1992, for a challenge to these ideas). These researchers theorized that items judged by subjects to be memorable – a judgment that could be influenced by a wide range of variables, such as personal relevance or repeated presentations – would lead them to expect a strong sense of "oldness" for such items at test. Correspondingly, subjects might take a conservative approach, believing this stringent criterion will facilitate accurate rejection of new items and be exceeded only by items that are truly old. A critical point here is that the items do not need to be more memorable in an objective sense to produce such an effect. The important thing is that the subject expects better memory for the class of items in question; this may or may not be an appropriate expectation.

The subjective memorability hypothesis was tested in four studies (Experiments 3-6) by asking participants to indicate, following the study phase, the percentages of words and paintings they expected to correctly recognize on the test. Two of these experiments incorporated an orienting task in the study phase, requiring a pleasantness judgment (on a 3-point scale in Experiment 4 and a 2-point scale in Experiment 5) for each item, with the goal of minimizing potentially confounding word memorization strategies that some participants reported in earlier studies. The only other difference was that after Experiment 3, wherein only one memorability estimate was requested for each item type, the question was changed such that participants were asked to estimate the percentages of studied words and paintings for which they expected to experience strong, fair, weak, and no memory at test. The two specific hypotheses related to memorability were as follows: (1) subjects would expect to remember more paintings than words, and (2) the extent of this tendency would be related to the extent of the difference in response bias between paintings and words (specifically, c for paintings minus c for words).

Although the significant difference in bias between paintings and words emerged as expected in all experiments (see Experiments 3-6 in Figure 2), the memorability hypotheses were not supported. Participants did not tend to anticipate better memory for paintings than words, being roughly evenly split with respect to which stimulus type they expected to recognize more successfully. Furthermore, there was no correlation between differences in memorability estimates and response bias. In other words, expecting better memory for paintings than words did not predict increasingly more conservative responding to paintings as compared to words. The results of these four studies effectively sealed the fate of the proposal that inflated expectations of the memorability of paintings – at least as measured in this particular manner – cause people to approach these items conservatively.

The Reminding Hypothesis

With each subsequent experiment further cementing the robustness of the conservative bias for paintings and yet revealing no apparent explanation, another hypothesis was pursued. This second hypothesis was based on the idea that if certain items are more conducive to study phase retrieval, also known as reminding (e.g., Hintzman, 2009), such items may produce confusion when encountered on the test. Hintzman conceptualizes reminding as a spontaneous process whereby encountering some stimulus evokes retrieval of a previously encountered stimulus, leading to encoding of not only the second stimulus, but also the experience of being reminded of the previous one. Perhaps paintings, being arguably more visually striking than words, more often remind people of previously viewed paintings or increase the likelihood that people will become aware of such events when they occur. This could conceivably lead people to adopt a conservative criterion for paintings in a few different ways. For example, reminding experiences might make the presence of similarities among items more salient. Having many such experiences for paintings might produce a sense that many paintings in this stimulus set share similarities, leading subjects to be cautious in endorsing paintings at test. Reminding might also promote uncertainty as to whether a test painting that seems initially familiar was actually on the list or has merely called other paintings to mind, if such experiences were common at study.

The reminding hypothesis for conservatism on paintings was tested in Experiment 7, in which participants were asked to press the spacebar each time an item on the study list reminded them of a previously presented item. The predictions were that on average, people would press the spacebar more often for paintings, and on an individual level, the magnitude of the difference between reported remindings on words and paintings would be correlated with the magnitude of the difference in response bias between the two types of items. Participants were, as usual, more conservative on paintings than words, and they did demonstrate a significant tendency to report remindings more often on paintings. However, contrary to expectations, these two tendencies were uncorrelated, suggesting that although paintings may be more conducive to reminding than words, this tendency cannot adequately explain the response bias difference between the two item types. The mechanism underlying the MBBE, therefore, remained an open question meriting more extensive investigation.

Chapter 2: Meta-analysis of Experiments 1-9

The Materials-Based Bias Effect (MBBE) research line has produced a veritable cornucopia of data. This is true with respect to not only the number of participants thus far (which, at the time of writing, well exceeds one thousand), but also the numerous types of information recorded during each experimental session and the multitude of possible combinations thereof. To name only a few examples, data collected from previous experiments include response times, confidence in old/new decisions at test, and judgments of item pleasantness in the case of experiments that included an orienting task at study. Furthermore, all of the experiments discussed in the current paper had 96 items in the study phase and 192 items at test, making for a lot of data points per participant. The number of potential different ways one could analyze the available data is colossal, and – while it is of course not appropriate to draw bold conclusions or make claims regarding causation based on such *post hoc* analyses – these existing data may yet hold clues that could suggest promising avenues for future investigation. To anticipate, the cross-experimental data presented in this chapter include average measures of sensitivity and response bias for each study; comparable analyses of sensitivity and bias divided by test quartile; confidence data in the form of receiver operating characteristics (ROCs), and response time data.

Data collection for most of the experiments included in the meta-analyses described below (specifically, all except Experiments 10 & 11) had been completed prior to the initiation of these analyses. For this reason, exact details regarding the total number of participants in each experiment prior to exclusions (if any), reasons for these exclusions, and demographic information for the participant pool at the time data were collected (which ranged from 2006 to

2014) were not always available. As much detail as is known is given below, and cases wherein details are inexact will be indicated.

Materials and procedures were similar across experiments and are therefore described in fairly broad terms. Experiment-specific details are mentioned where they are relevant to the hypotheses being tested or likely to influence the results; things like changes to the study phase task will be described, for example, while differences in equipment or slight changes to the stimulus pool will not. Most of the analyses described in this chapter were based on data from the MBBE experiments only, the sole exception being the forest plot analyses, wherein response bias data from the five pre-MBBE paintings studies described by Lindsay and Kantner (2011) were also used. These experiments are not the focus of the current investigation and are only mentioned to establish the context for the subsequent MBBE research line and further emphasize the remarkable consistency of conservative responding in recognition memory for paintings. As such, the methodological details of these experiments will not be described (but see Lindsay and Kantner, 2011, for more information).

Method: General

Participants

Participants in all studies – both the MBBE line and the preceding paintings-only experiments – were undergraduate students at the University of Victoria who participated voluntarily, generally for bonus course credit (but possibly a small payment in some earlier experiments). These participants were drawn from a pool in which most individuals are female (2013 estimate: 69%) and the vast majority are between 18 and 25 years of age.

Analyses for the five pre-MBBE paintings experiments were based on a total of 233 participants, with Ns for individual experiments ranging from 20 to 57. At least 554 participants were involved in the first nine MBBE experiments. Four hundred and eight of these participants completed one of eight within-subjects versions of the experiment wherein the test and study lists comprised both words and paintings (individual Ns = 21-84). Three of these participants were excluded from analyses – one due to experimenter error, and two for unknown reasons – such that meta-analyses were based on 405 participants. The remaining 146 completed one of two experiments comprising only words or paintings (individual Ns = 66 and 80; exactly half in each condition). A list of all experiments that contributed data to the following meta-analyses is presented in Appendix A and includes more experiment-specific sample size information.

Materials

All experiments were administered using E-Prime software (Schneider, Eschman, & Zuccolotto, 2002a & 2002b) on a desktop PC. High-resolution digital scans of relatively obscure masterwork paintings by renowned artists were used as stimuli in all experiments. With the exception of one pre-MBBE paintings study that used only portraits, these paintings encompassed various styles and themes (e.g., portraits, landscapes, still lifes, etc.). The exact set of images used differed somewhat among experiments, but all were selected from a larger collection assembled by Jeffrey Toth for the purpose of developing a memory-training video game. In experiments that used words, these were 4- to 8-letter medium- to high-frequency nouns obtained from the MRC psycholinguistic database (http://www.psych.rl.ac.uk; Coltheart, 1981).

All MBBE experiments had 96-item study lists bookended by three primacy and three recency buffers. In within-subjects experiments (1-7, & one group in 8), 48 of these study items were words and 48 were paintings; in between-subjects experiments (8 & 9), all 96 items were of the same stimulus type. The test list always included all studied items and 96 new items (48 of each stimulus type in within-subjects experiments) for a total of 192 items. Both study and test lists were randomly generated anew for each participant from the previously mentioned sets of words and paintings, the only constraint aside from list length and overall composition being that a given participant's test list always included all items from that participant's study list. In other words, old and new items were randomly intermixed throughout the test list, as were words and paintings in both the study and test lists in the within-subjects studies.

Procedure

Words and/or paintings were presented one at a time in the centre of a white background, with words typically displayed in size 14 black font and paintings ranging in size from roughly 200×200 to 350×360 pixels. Study items were presented for 1 s each following display of a 250-ms fixation cross, with a 1-s interstimulus interval (ISI) in all but Experiments 4, 5, and 7, which all effectively had 2-s ISIs for reasons explained below.

At the beginning of the study phase, participants were instructed to attend to the items and try to remember them as well as possible for a later memory test. In addition to these instructions, participants in Experiments 4 and 5 were instructed to make pleasantness judgments for each item on a 3- (Experiment 4) or 2-point scale (Experiment 5), and had 2 s to make these judgments. Participants in the reminding study, Experiment 7, were also asked to report each occasion on which they found themselves spontaneously reminded of a previous word or painting in the list by pressing the spacebar during the 2-s ISI following the item that elicited this experience.

Between the study and test phases, participants always completed a 5-min distractor task unrelated to the experiment itself, such as listing countries on a sheet of paper as they came to mind. In Experiments 3-6, participants were also asked for subjective memorability estimates at this time. Test phase instructions were largely the same for all experiments: participants were informed that they would again see a series of words and/or paintings – some of which would be items that had been on the study list, and others that would be new in the context of the experiment – and that they would be asked to judge each as old/studied or new/unstudied, and to provide an estimate of how confident they were in this decision. These decisions were made on a 6-point scale ranging from 1 ("definitely new") to 6 ("definitely old"). In Experiment 2, some participants also received accuracy feedback throughout the test phase, but this manipulation was not of interest for current purposes and data were collapsed across conditions accordingly. Participants were always debriefed at the end of the experiment, and in some experiments this phase also included additional questions (e.g., about self-perceived accuracy), but these responses are not reported.

Meta-analysis of Sensitivity and Response Bias

As is common in investigations of recognition memory, calculations of sensitivity and response bias from hit and false alarm rates have typically been the first data analysis step upon completion of each MBBE experiment. Both calculations are critical – response bias for obvious reasons, given its centrality to this line of research, but also sensitivity, for the purposes of establishing the response bias difference between words and paintings as an entity independent from any differences in discriminability between the two stimulus types.

There are numerous options, graphical and otherwise, for presenting bias and sensitivity data. As with any experimental result, researchers may prefer one method to another depending on the aspects of the data they wish to emphasize. In the case of the MBBE, there are several such aspects: the tendency for response bias to be significantly conservative for paintings when stimuli are either only paintings, or both paintings and words; the marked differences in bias between words and paintings in experiments using the latter approach, and the tendency for liberal responding to words in this context but approximately neutral responding when they are the only stimulus type; the remarkable consistency of the above effects across experiments; and the apparent independence of the above effects from materials-based differences in recognition sensitivity.

With these important aspects of the data in mind, as well as the fact that the nine experiments described above represent variations on the same experimental structure and could therefore be sensibly combined in a meta-analysis, it was decided that forest plots would be the most effective means of conveying sensitivity and response bias values. Forest plots generally depict the individual means or effect sizes obtained in a series of studies, their corresponding confidence intervals, a summary measure representing the result of the overall meta-analysis, and some representation of the weighting of each individual mean in calculating this final measure (Lewis & Clarke, 2001). The resultant plot paints a comprehensible picture of the main meta-analytic result and how it was obtained that can be easily described and understood largely intuitively by the viewer.

Forest plots were constructed for both response bias and sensitivity using results from the nine previously described MBBE studies (Experiments 1-9), with values for words and paintings calculated separately but displayed in a single plot to facilitate comparison. Experiments 10 and 11, which are described in detail in Chapters 3 and 4, respectively, were also included in these two meta-analyses. Analyses were also conducted separately for within- and between-subjects studies, yielding a total of four such plots for the MBBE experiments. One additional forest plot was constructed to show the response bias results for the five pre-MBBE experiments conducted with only paintings (Lindsay & Kantner, 2011).

Method

Overall hit and false alarm rates were calculated for each participant in all experiments. All calculations were done separately for words and paintings where applicable (i.e., in the within-subjects experiments), and occasional false alarm rates of 0 and hit rates of 1 were replaced according to Macmillan and Kaplan (1985). Hit rates (HRs) and false alarm rates (FARs) were then used to determine sensitivity and response bias. The sensitivity measure used in this and all subsequent analyses was d', calculated as the difference between the normalized (z-transformed) values of hit and false alarm rates; the corresponding response bias measure is c, which is calculated as the negative of half the sum of these normalized values divided by two (i.e., -[z(HR) + z(FAR)/2]; Macmillan & Creelman, 2005). These two measures were averaged across participants, but within experiments, and the resultant values were used in the meta-analysis.

Meta-analysis calculations and construction of forest plots were accomplished using Cumming's (2001) ESCI software. In addition to the abovementioned means, these calculations require corresponding standard deviation values and sample sizes. The required values were inputted in ESCI, which completed the meta-analyses using a fixed effects model, calculated 95% confidence intervals for individual and meta-analysed means, and displayed the results in forest plots.

Results

All sensitivity and response bias meta-analyses are presented as forest plots (Figures 1-3). The size of each square is proportional to the weighting of that particular mean in the meta-analysis calculation based on the associated variance and sample size. In other words, larger squares represent means that were more heavily weighted in the final calculation and tend to come from larger samples with relatively low variance. Error bars are 95% confidence intervals.

The diamonds at the bottom of each plot represent the result of the associated meta-analysis and include the corresponding 95% CI.

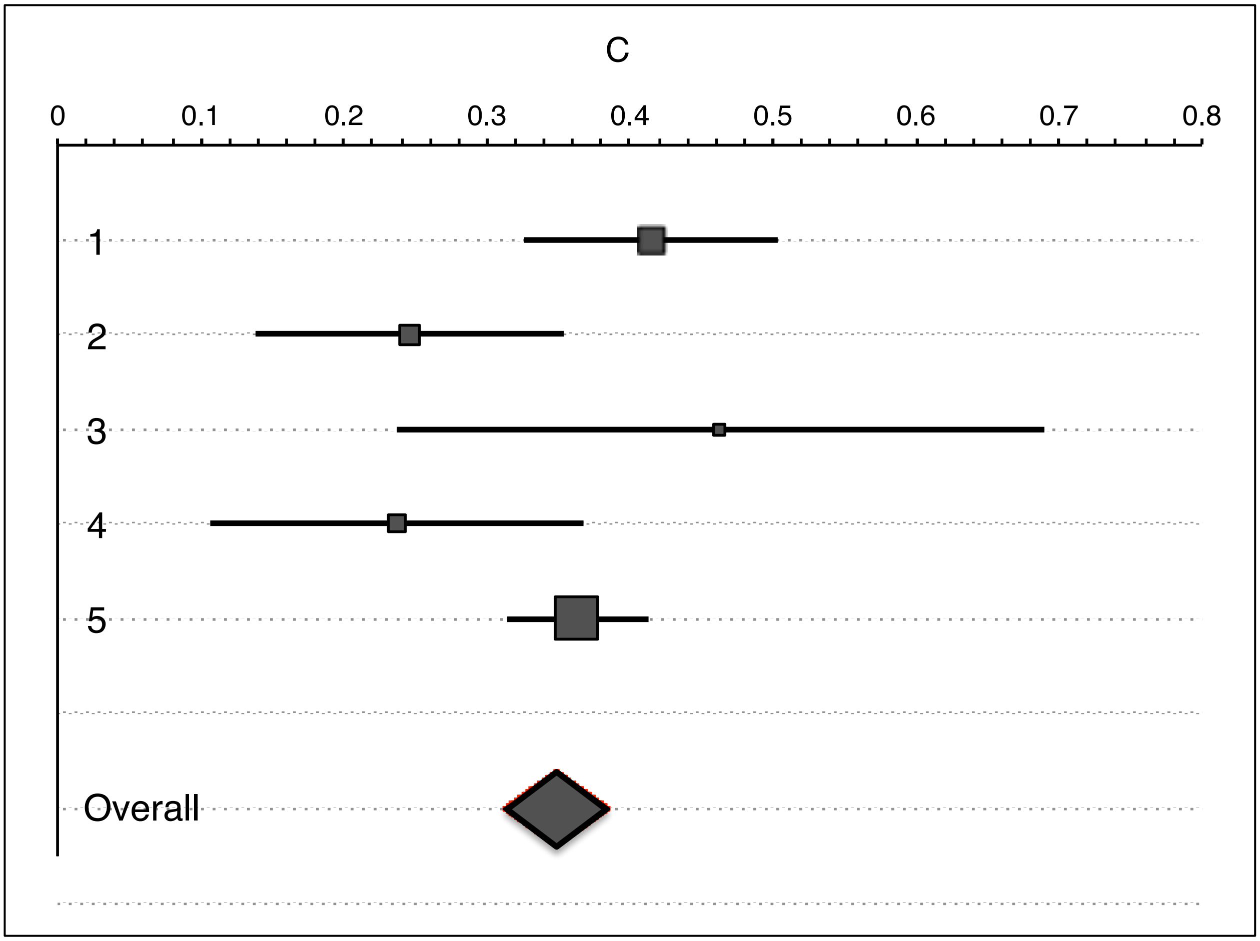

Figure 1. **Forest plot depicting mean response bias (c) values obtained by Lindsay and Kantner (2011).**

Data are from five recognition memory experiments using paintings as stimuli, with the overall estimate of C based on a meta-analysis of the five experiments shown at the bottom. Error bars are 95% confidence intervals (CIs), and square sizes are proportional to the weighting of each experiment's mean in the meta-analysis calculation based on sample size and variance. The 95% CI for the overall estimate is represented by the edges of the diamond. Plot constructed using ESCI (Cumming, 2001).

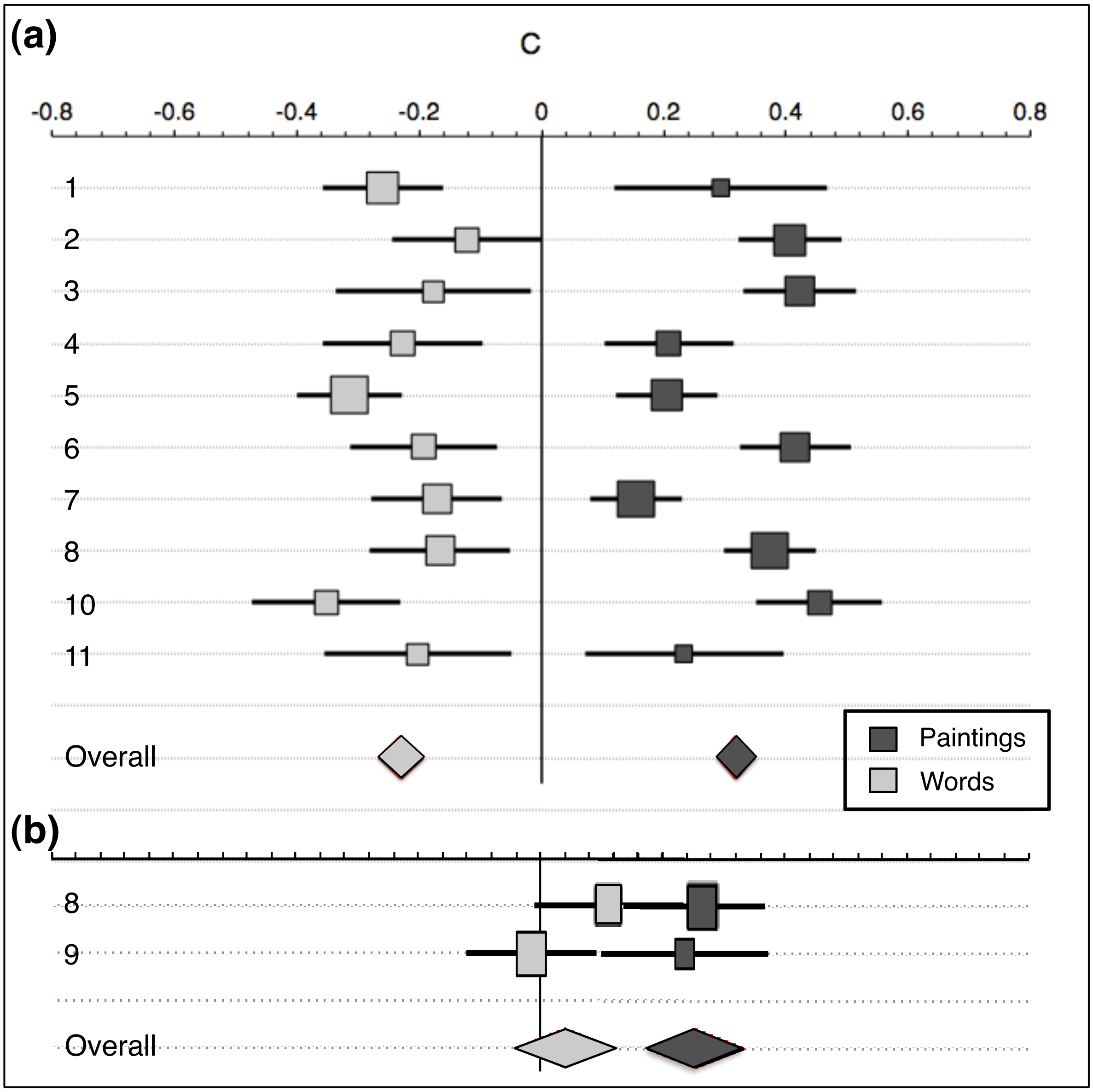

Figure 2. **Forest plot depicting mean response bias (c) values obtained for paintings (■) and words (■) for Experiments 1-11 and overall estimates of c (shown as diamonds) from four meta-analyses.**

Item type was manipulated either within- (a) or between-subjects (b). Numbers on the left hand side refer to specific experiments (see Appendix A). Error bars are 95% confidence intervals (CIs), and square sizes are proportional to the weighting of each experiment's mean in its corresponding meta-analysis calculation based on sample size and variance. 95% CIs for overall estimates are represented by the left and right corners of each diamond. Plots constructed using ESCI (Cumming, 2001).

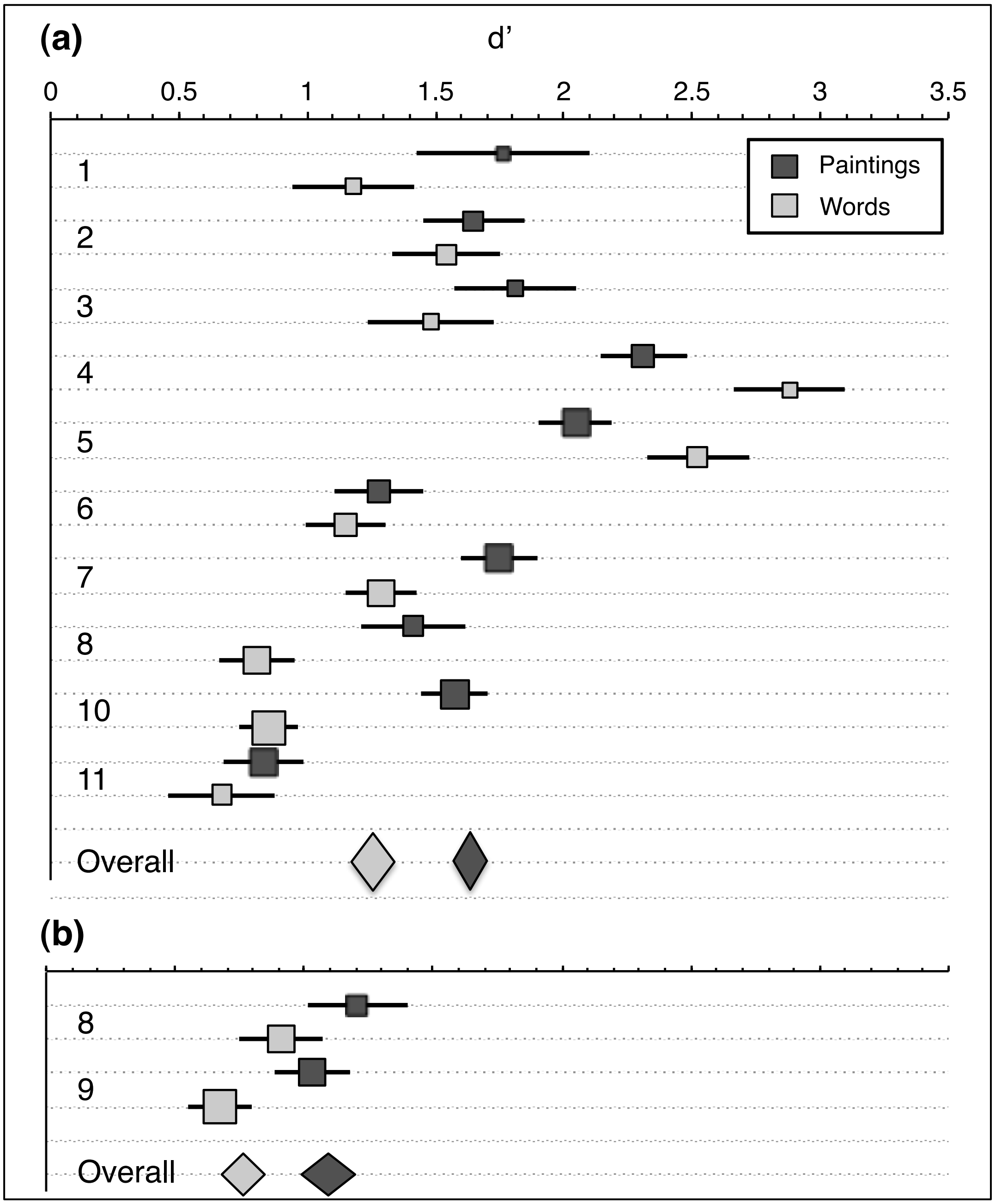

Figure 3. **Forest plot depicting mean sensitivity (d') values obtained for paintings (■) and words (■) for Experiments 1-9 and overall estimates of d' (shown as diamonds) from four meta-analyses.**

Item type was manipulated either within- (a) or between-subjects (b). Numbers on the left hand side refer to specific experiments (see Appendix A). Error bars are 95% confidence intervals (CIs), and square sizes are proportional to the weighting of each experiment's mean in its corresponding meta-analysis calculation based on sample size and variance. 95% CIs for overall estimates are represented by the left and right corners of each diamond. Plots constructed using ESCI (Cumming, 2001).

Figure 1 depicts response bias in the five paintings-only experiments conducted by Lindsay and Kantner (2011). Means for all experiments were significantly conservative (i.e., c differed significantly from zero in the positive direction) and yielded a meta-estimate of c of 0.35 (95% CI: 0.312, 0.386). Response bias data for all within-(Figure 2a) and between-subjects (Figure 2b) MBBE experiments are presented in Figure 2, with painting means in dark grey and concentrated on the right, and word means in light grey and mostly on the left. C was significantly conservative for paintings in every single study included in this analysis, as evidenced by the fact that none of the confidence intervals for paintings overlap with zero in either panel a or b. The within-subjects meta-analysis of c for paintings yielded a mean estimate of 0.319 (95% CI: 0.290, 0.348; Figure 2a), while the corresponding between-subjects result was 0.255, (95% CI: 0.176, 0.334; Figure 2b).

With respect to words, response bias was significantly liberal in all of the within-subjects experiments (Figure 1a), and the associated meta-estimate for c was -0.229 (95% CI: -0.265, -0.194). In the between-subjects experiments, however, response bias for words was approximately neutral in one case and conservative in the other, and the meta-estimate of c across these two experiments did not differ significantly from zero/neutrality (0.041; 95% CI: -0.037, 0.118; Figure 2b). Experiment 8, in which c for words was directionally conservative, was also the only case in which response bias did not differ significantly between words and paintings (although note that the confidence intervals in these plots, being based on variance in c for their corresponding mean only and not mean differences, are not themselves designed to answer this question; the relevant independent-samples t-test result, however, supports this interpretation, *p* = 0.06). The corresponding meta-estimates of c across the two between-subjects experiments, however, were still significantly different for the two item types (Figure 2b).

Figure 3 shows the parallel analyses for sensitivity (d’). As can be seen at a glance, the magnitudes of d’ and the pattern of differences (or lack thereof) between words and paintings differed a great deal across experiments, but on average, sensitivity came out to be significantly higher for paintings in the meta-analyses for both the within-subjects (Figure 3a) and between-subjects (3b) experiments. In the former case, d’ for paintings was estimated as 1.625 (95% CI: 1.572, 1.678) while mean d’ for words was 1.261 (95% CI: 1.209, 1.314). The corresponding estimates from the between-subjects analysis were 1.094 for paintings (95% CI: 0.980, 1.210) and 0.762 for words (95% CI: 0.666, 0.857).

Discussion

The overall picture painted by the above response bias and sensitivity meta-analyses was, of course, no surprise. Indeed, the response bias pattern clearly seen in Figures 1 and 2 is essentially the premise of this entire research line, with the overall meta-estimates only confirming what was already known: response bias for painting stimuli is consistently and significantly conservative in recognition memory studies using either only paintings (Figures 1 & 2b) or both paintings and words (Figure 2a), and in the latter case, is also markedly *more* conservative for paintings than words, which tend to yield a significantly liberal bias in experiments using this design. Although the magnitude of this difference in response bias appears to be reduced in the between-subjects context, and this finding may be interesting in itself, the more critical point for current purposes is that the tendency toward conservative responding to paintings persists regardless of whether words are included in the experiment or not.

In addition to the abovementioned persistence of conservative bias for paintings across within- and between-subjects designs, these forest plots also serve to emphasize the robustness of this phenomenon in the face of a number of other manipulations. Although the absolute magnitude of c for paintings varied a fair amount across experiments, it was still significantly conservative in every single experiment presented in the current paper. This means the effect has withstood various changes to the study task, the response scale used at test, and the stimulus pool, and has been apparent in every group of participants to participate in the study across a roughly eight-year span, exhibiting a degree of reliability that is quite remarkable.

This reliability is also important to consider in relation to d’ differences between words and paintings and the variability in this pattern across experiments. Although c and d’ are, as previously mentioned, ostensibly unrelated and statistically independent measures, this independence is in fact model-dependent. Specifically, c and d’ are only orthogonal statistics when none of the assumptions of the equal-variance signal detection (EVSD) model are violated, which is rarely if ever the true state of affairs in recognition memory experiments. The particular assumption that appears to be consistently violated in such experiments is that of equal variance itself, which refers to the idea that variance – which, in the context of recognition memory, can be thought of as the distribution of values corresponding to familiarity or strength of evidence – is equal for the “old” and “new” item distributions (Macmillan & Creelman, 2005). This tends not to be the case in recognition studies, in which the distribution of “evidence strength” for old items is almost always shown to be more variable than that for new items (e.g., Glanzer, Kim, Hilford, & Adams, 1999; Mickes, Wixted, & Wais, 2007; Ratcliff, Sheu, & Gronlund, 1992). In such a case, c and d’ are not technically pure measures of bias and sensitivity; for example, d’ will tend to overestimate and underestimate actual sensitivity under conservative and liberal criteria, respectively (e.g., Dougal & Rotello, 2007).

The interdependence of d’ and c when EVSD assumptions are violated is an important consideration in drawing conclusions about differences in these measures among groups or conditions, and the exclusive reliance on these measures in the current paper could be rightly criticized, particularly given the centrality of c in this line of research. One could conceivably argue, for example, that the reported effect of stimulus type on response bias might be a meaningless artefact of what is actually an effect on sensitivity or some other aspect of the old and new item distributions, an explanation potentially consistent with the overall tendency for d’ to be higher for paintings (Figure 3). However, there are several pieces of evidence that argue against this being the case. Probably the most convincing is discussed in a later section devoted to receiver operating characteristics (ROCs), but the results in the current section also provide reason to doubt such an explanation.

While d’ was indeed higher for paintings on average in the current studies, a closer look at the results of individual experiments in Figure 3 reveals that this pattern was not without exception; for example, d’ was markedly higher for words than paintings in Experiments 4 and 5, and other experiments yielded roughly equivalent d’ results for the two item types (e.g., Experiment 2). The pattern of bias differences, in contrast, was extraordinarily similar across studies (Figure 2), and if fluctuations in c and d’ were in fact related in some way, this relationship is certainly not obvious based on a comparison between the two forest plots. This is not to dismiss the validity of criticisms of reliance on c and d’ in recognition memory experiments – on the contrary, this is something to be addressed in future analyses – but it seems unlikely that the effect of central interest in the current paper is attributable to mere differences in the discriminability of various item types or some underlying distributional fluke, particularly in light of some of the results described in later sections.

Quartile Analyses

Calculations of an individual participant's hit and false alarm rates – and therefore sensitivity and response bias – are typically made by collapsing across all items on the test list, and this was indeed the case in the meta-analyses discussed above. However, the test lists in the experiments in the current paper were fairly long, comprising 192 items each. Given such a lengthy test phase, it seems unlikely that the subjective experience of making a recognition decision about the first item would be the same by the time the participant reaches the final item. Similarly, there are numerous reasons one might expect the processes underlying recognition memory decisions to change in some way over the course of the test, such as an increased potential for inter-item confusion as more items are introduced, the possibility of self-calibration and/or meta-memorial processes that might be adjusted or evolve as the test proceeds, and basic variables that might tend to change over time in any experiment, like increasing fatigue or wavering attention.

The decision to conduct order-based analyses with the MBBE data, specifically by dividing responses to test items into four sections according to the order of presentation, was the result of a suggestion from Jim Nairne. These analyses were exploratory in nature, with the goal being the rather broad one of seeing whether response bias might differ in some way over the course of the test. For example, participants might adopt a conservative criterion for paintings almost immediately and maintain this throughout the test, or it might be an evolving process; either result might provide clues regarding the origin of this bias and guide future research. As such, these analyses were not initiated with any specific hypotheses in mind. Although response bias was, of course, the focus of interest, comparable analyses are also presented for sensitivity (d') and hit and false alarm rates, and all were conducted for both paintings and words. The only expectation was that sensitivity might show a downward trend across quartiles, given the ever-increasing number of items participants are exposed to over the course of the test, the possibility of general fatigue, and the established tendency for performance to decrease over the course of recognition memory testing (Criss, Malmberg, & Shiffrin, 2011; Ratcliff & Murdock, 1976). There was no clear basis on which to hypothesize about response bias or how across-quartile patterns might or might not differ between item types.

Method

In all nine of the previously mentioned experiments, the test list comprised 192 items. For each experiment, participants' test phase responses were divided into 48-item quartiles, such that the first quartile included a participant's responses to the first through 48th items, the second quartile included responses to the 49th through 96th items, and so on. As previously mentioned, responses in these nine experiments were made on a 6-point confidence scale ranging from 1 ("definitely new") to 6 ("definitely old"), with intermediate responses representing "probably [new/old]" and "maybe [new/old]". However, confidence ratings were not of interest in this analysis, so responses 1 through 3 were simply coded as "new" and responses 4 through 6 as "old".

Hit and false alarm rates were calculated based on the responses and numbers of old/new words/paintings in each individual quartile; on average, each quartile would be expected to include 12 items in each category (i.e., 12 old paintings, 12 new paintings, etc.), but there was some variation around this due to the randomized nature of the test phase, sometimes yielding as few as 3 items in a given category. Due to the low number of items and responses on which these analyses were sometimes based, hit and false alarm rates of 1 and 0 were far more frequent than in the case of the whole test analyses. These ceiling and floor rates were replaced as usual according to Macmillan and Kaplan (1985), who suggested replacing rates of 0 using the formula 0.5/N, where N is the number of new (or old, in the rare case of a zero hit rate) items, and rates of 1 with the formula 1-(0.5/N). In the whole-test analyses, wherein N was always 48, these formulas always yielded estimates of approximately 0.01 and 0.99, therefore changing mean hit and false alarm rates very little; in the quartile analyses, however, these estimates had the potential to differ more substantially from 0 and 1 (e.g., a false alarm rate of 0 based on only 4 items would be replaced with 0.125), and this combined with the higher frequency of such replacements meant they would exert greater influence on mean hit and false alarm rates than usual. Other options were considered in light of this, but reassuringly, the overall pattern of results was largely similar regardless of whether rates of 0 and 1 were replaced with the above formula, values of 0.01 and 0.99, or excluded from analysis entirely. The exact magnitudes were of course nontrivially affected by these changes, but because the focus of interest was the overall pattern more so than the individual values, it was decided that the usual formula would be suitable provided appropriate caution in interpreting the results.

Based on these hit and false alarm rates, sensitivity and response bias calculations were conducted for the first set of 48 items, the second set of 48 items, and so on, yielding four points per measure for both words and paintings. Hit rate, false alarm rate, sensitivity, and response bias data were then averaged across subjects (but within experiments), and the means were plotted with their respective 95% within-subjects confidence intervals (Loftus & Masson, 1994). ANOVA results used in calculating the confidence intervals for the within-subjects experiments came from a series of 2 (item type: word or painting) × 4 (test quartile: first, second, third, or fourth) repeated measures ANOVAs. For between-subjects experiments, these ANOVAs used test quartile as the only dependent measure.

To supplement the quartile analyses for the within-subjects experiments, each participant's responses to the first and last items on the recognition test were extracted and overall proportions of "old" responses were calculated according to item type. These proportions were calculated by collapsing across all participants in the ten within-subjects experiments and are presented with 95% binomial CIs.

Results

The results of the quartile analyses are depicted in Figures 4 through 11. The results for multiple studies are shown in each figure, and individual plots, which include results for both paintings and words, are labeled by experiment number (see Appendix A). Figures 4, 5, and 10 show the hit and false alarm rates for the within-subjects Experiments 1, 2, 7, and 8; the subjective memorability experiments, 3-5; and the between-subjects experiments 8 and 9, respectively. The corresponding graphs for c, grouped in the same manner, are shown in Figures 6, 7, and 11 (top row). Figures 8, 9, and 11 (bottom row) illustrate the results for d'.

For the within-subjects experiments, the overall proportion of "old" responses given to the first item on the test list was 0.56 (95% CI: 0.54, 0.60) when said item was a word and 0.49 (95% CI: 0.46, 0.52) when said item was a painting. Corresponding proportions for the last item on the test list were 0.60 (95% CI: 0.57, 0.63) for words and 0.37 (95% CI: 0.33, 0.40) for paintings.

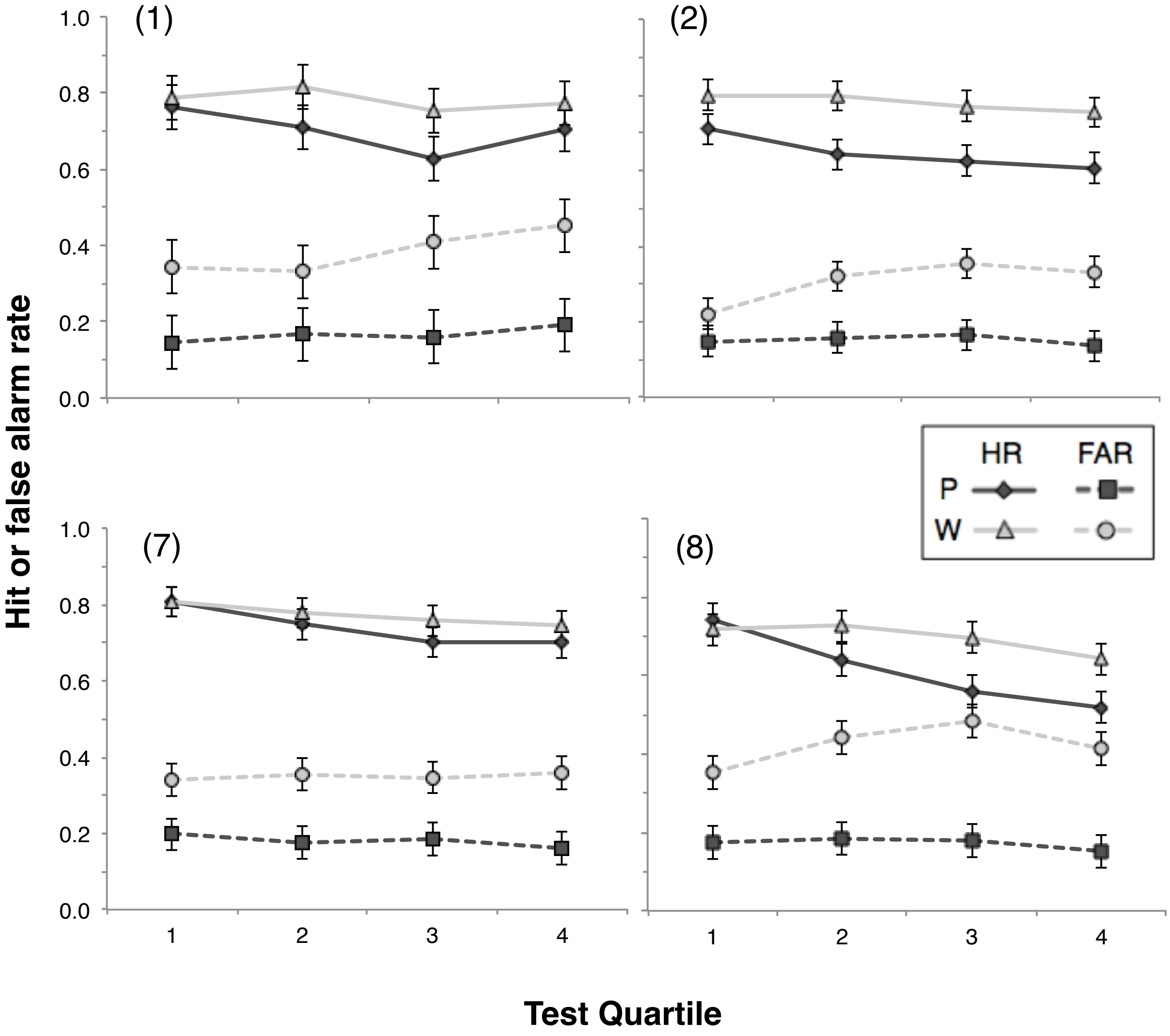

Figure 4. **Mean hit (HR) and false alarm rates (FAR) for paintings (P) and words (W) in each 48-item test quartile in Experiments 1, 2, 7, and 8, wherein item type was manipulated within-subjects.**

Error bars are 95% within-subjects confidence intervals (Loftus & Masson, 1994) calculated based on the results of a 2 (item type) x 4 (quartile) repeated-measures ANOVA.

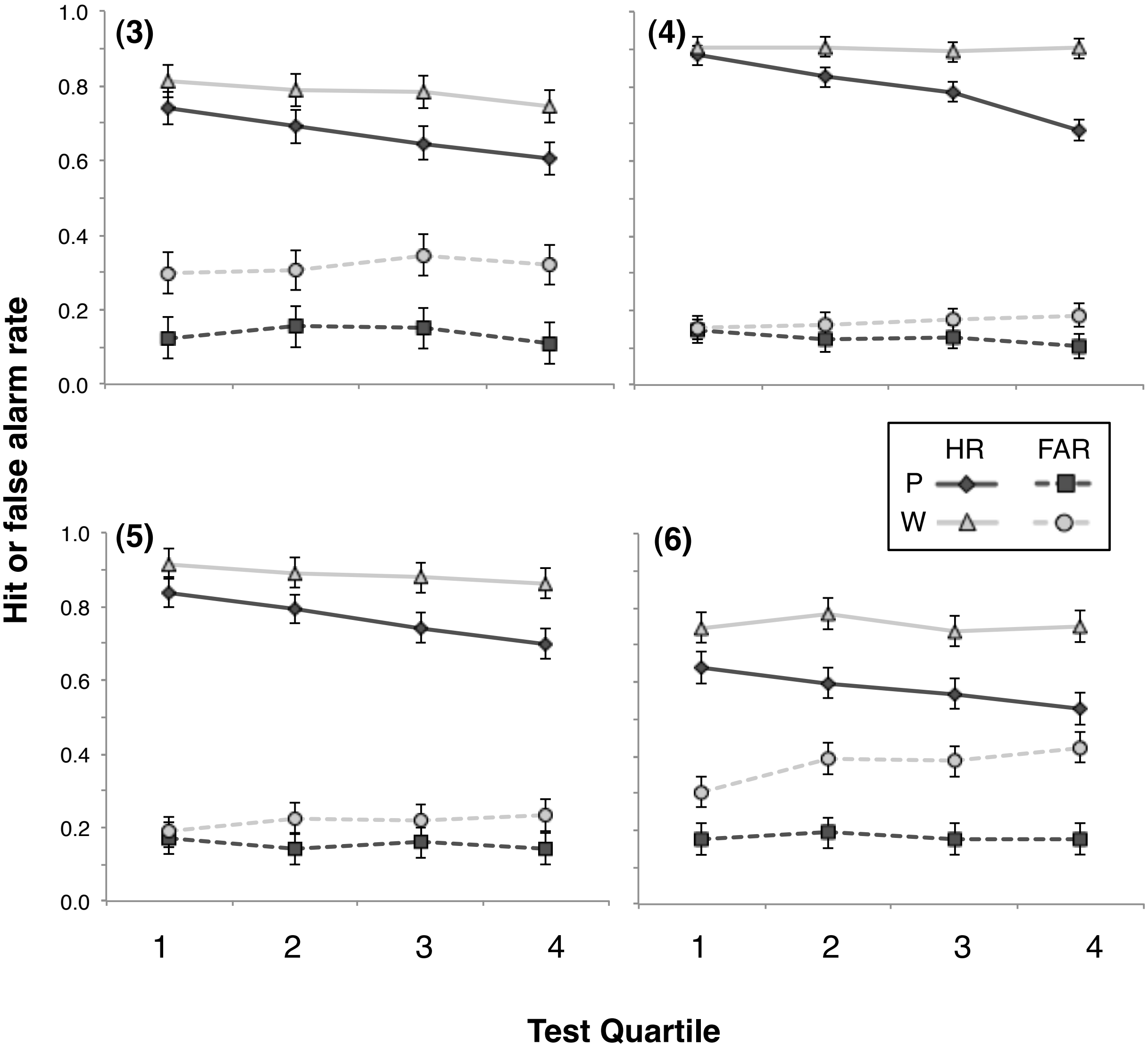

Figure 5. **Mean hit (HR) and false alarm rates (FAR) for paintings (P) and words (W) in each 48-item test quartile in the subjective memorability experiments (Experiments 3-6).** Error bars are 95% within-subjects confidence intervals (Loftus & Masson, 1994) calculated based on the results of a 2 (item type) x 4 (quartile) repeated-measures ANOVA.

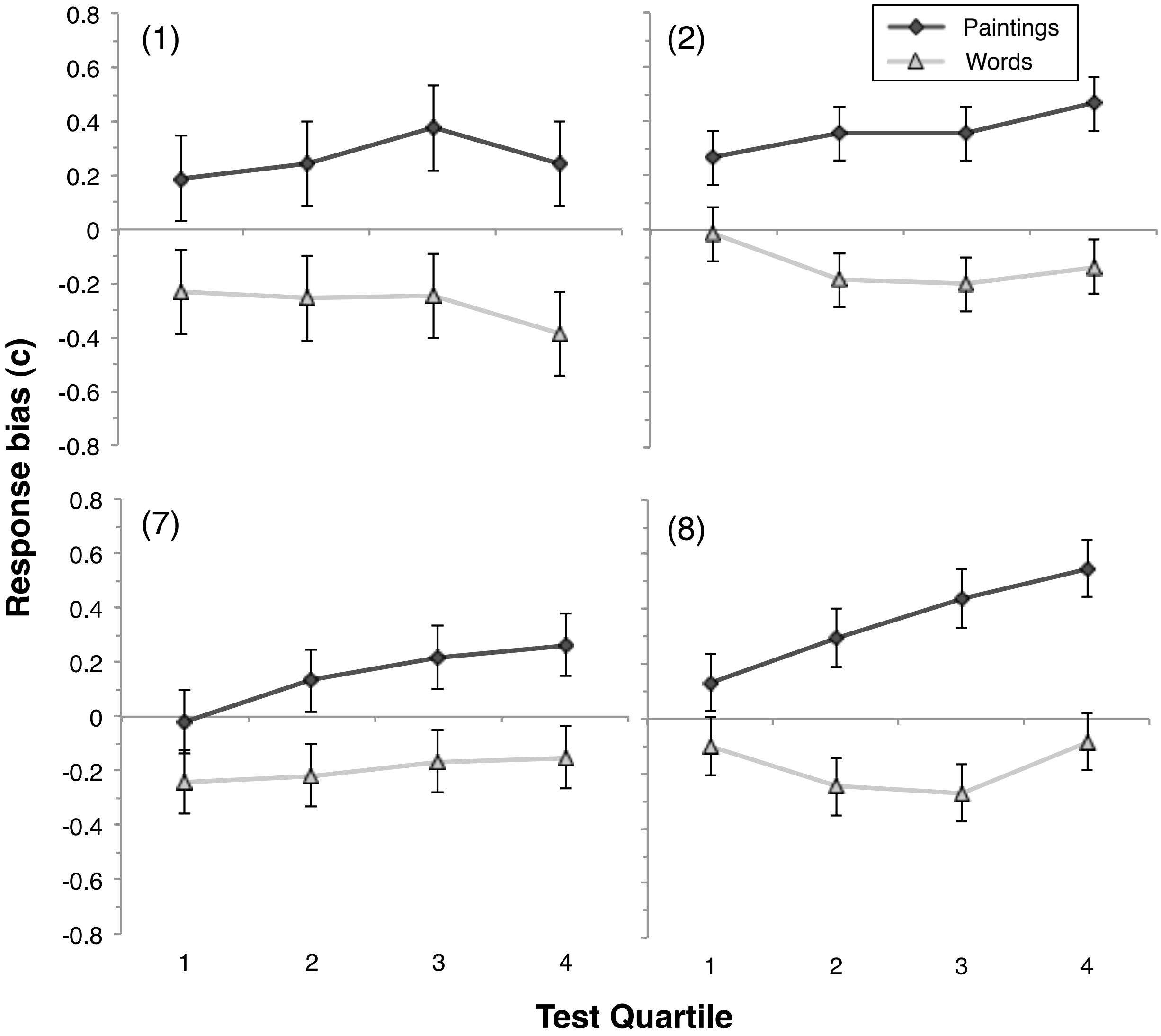

***Figure 6.* Mean response bias (c) for paintings and words in each 48-item test quartile in Experiments 1, 2, 7, and 8, wherein item type was manipulated within-subjects.**

Error bars are 95% within-subjects confidence intervals (Loftus & Masson, 1994) calculated based on the results of a 2 (item type) x 4 (quartile) repeated-measures ANOVA.

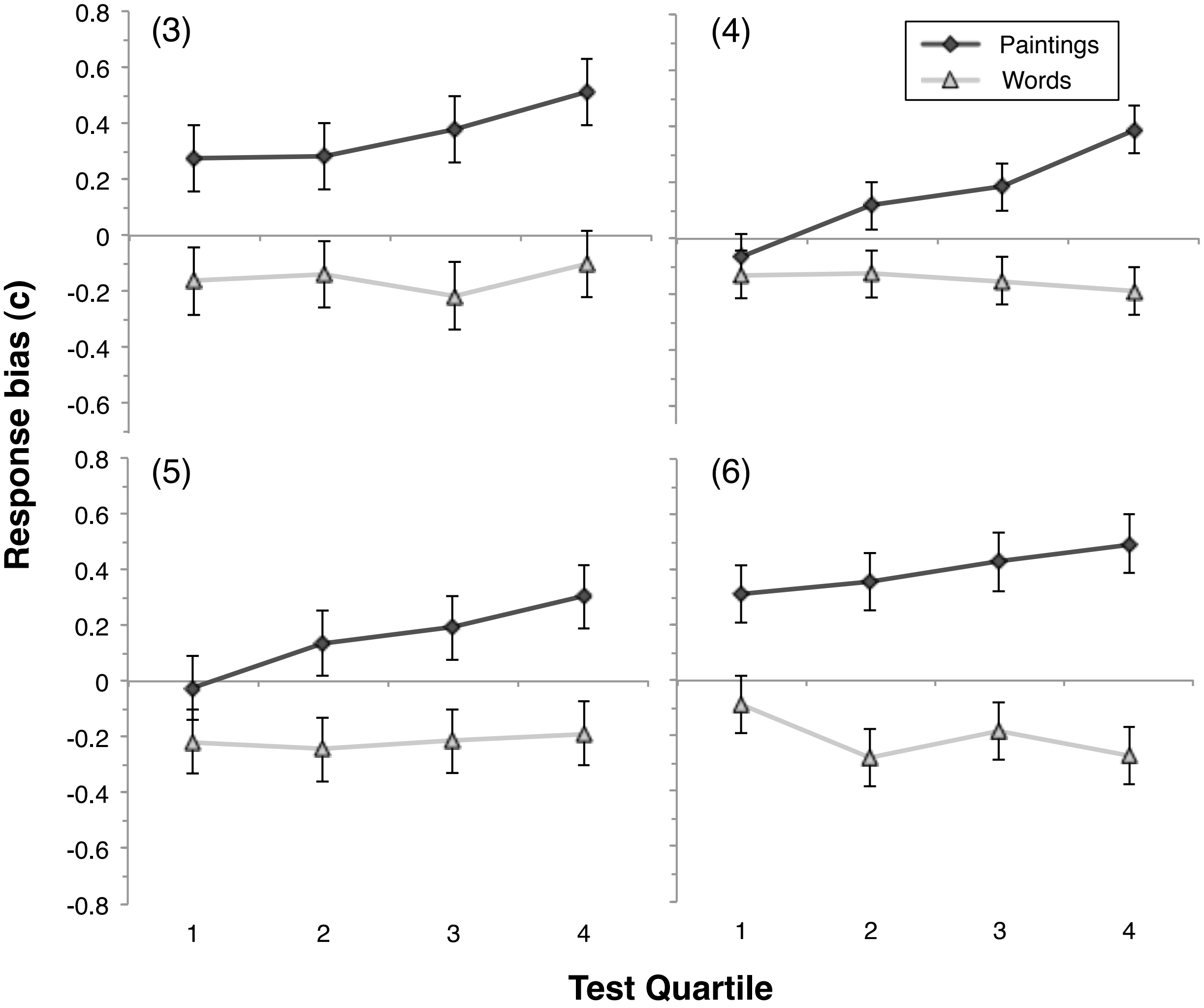

***Figure 7.* Mean response bias (c) for paintings and words in each 48-item test quartile in the subjective memorability experiments (Experiments 3-6).**

Error bars are 95% within-subjects confidence intervals (Loftus & Masson, 1994) calculated based on the results of a 2 (item type) x 4 (quartile) repeated-measures ANOVA.

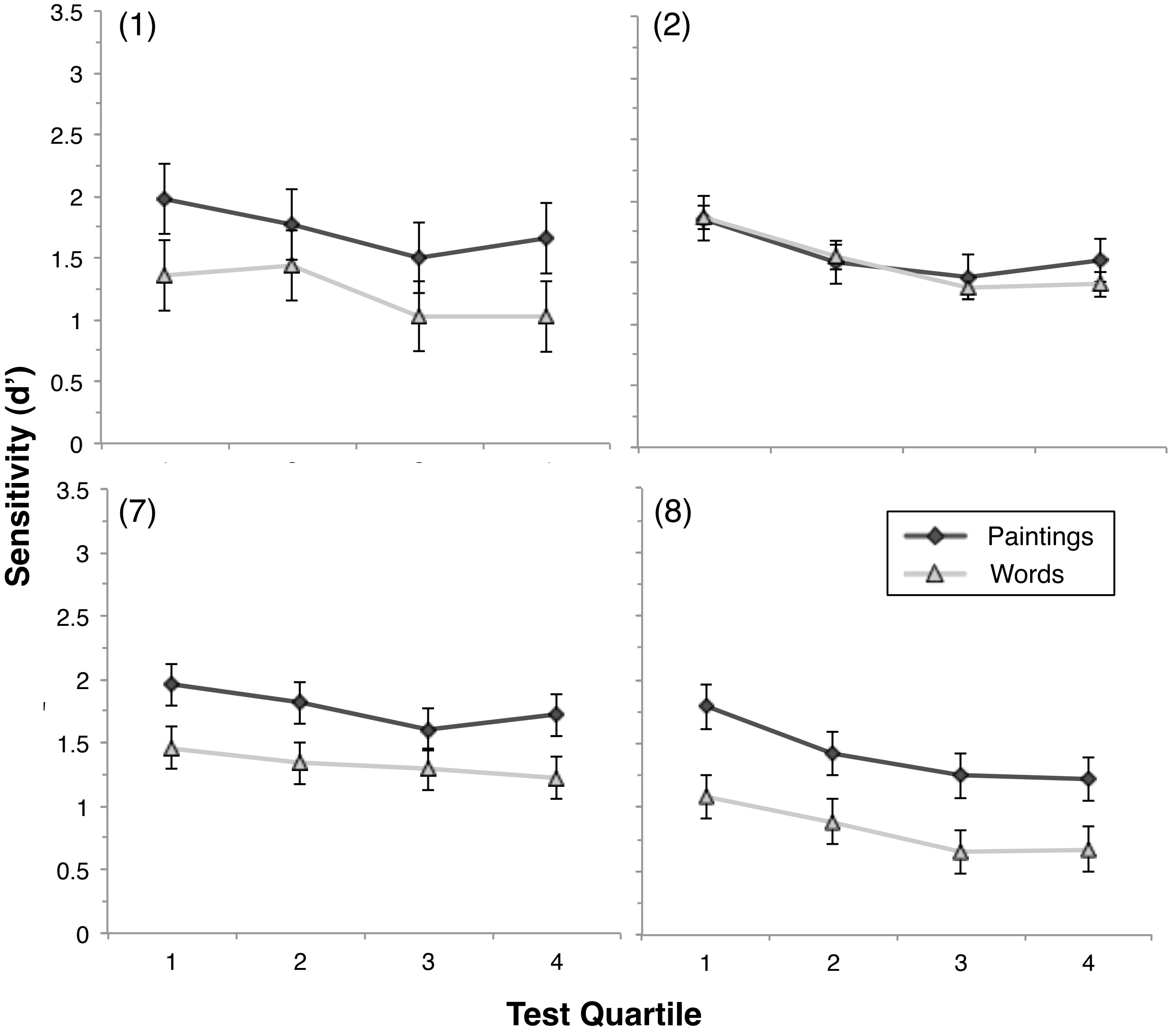

Figure 8. **Mean sensitivity (d') for paintings and words in each 48-item test quartile in Experiments 1, 2, 7, and 8, wherein item type was manipulated within-subjects.**

Error bars are 95% within-subjects confidence intervals (Loftus & Masson, 1994) calculated based on the results of a 2 (item type) x 4 (quartile) repeated-measures ANOVA.

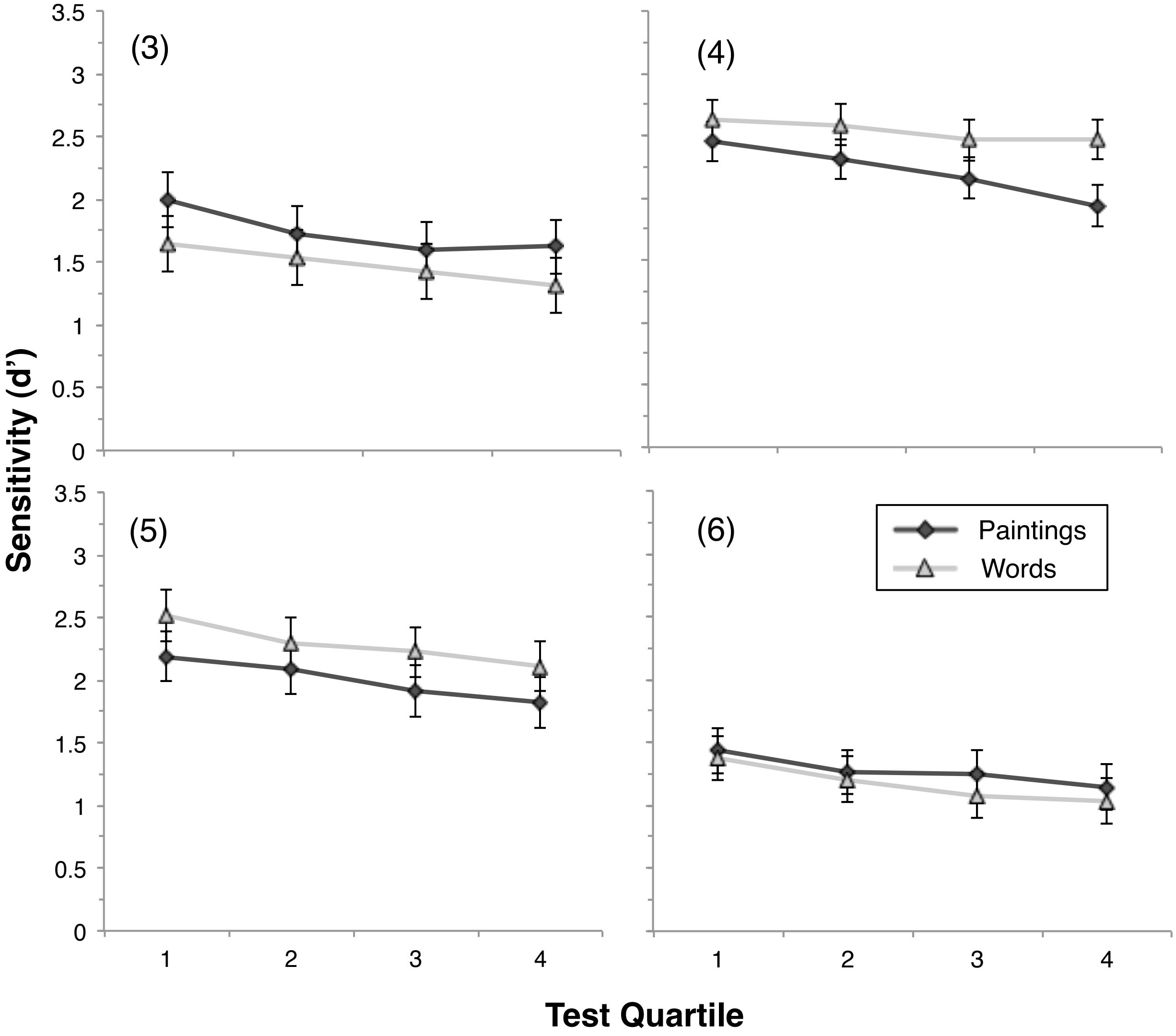

***Figure 9.* Mean sensitivity (d') for paintings and words in each 48-item test quartile in the subjective memorability experiments (Experiments 3-6).**

Error bars are 95% within-subjects confidence intervals (Loftus & Masson, 1994) calculated based on the results of a 2 (item type) x 4 (quartile) repeated-measures ANOVA.

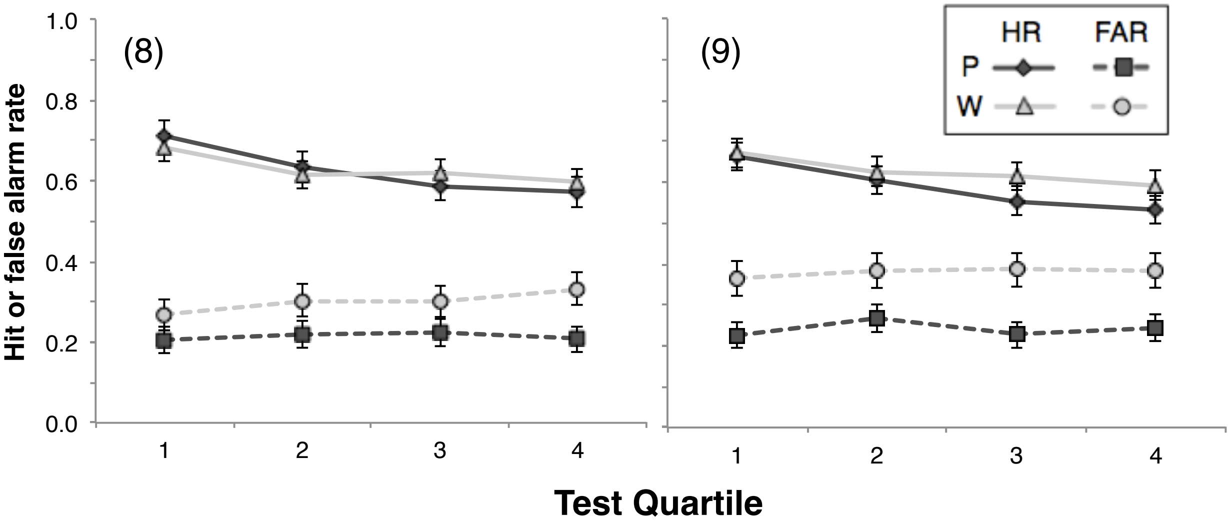

***Figure 10.* Mean hit (HR) and false alarm rates (FAR) for paintings (P) and words (W) in each 48-item test quartile in Experiments 8 and 9, wherein item type was manipulated between subjects.**

Error bars are 95% within-subjects confidence intervals (Loftus & Masson, 1994) calculated based on the results of individual repeated-measures ANOVAs.

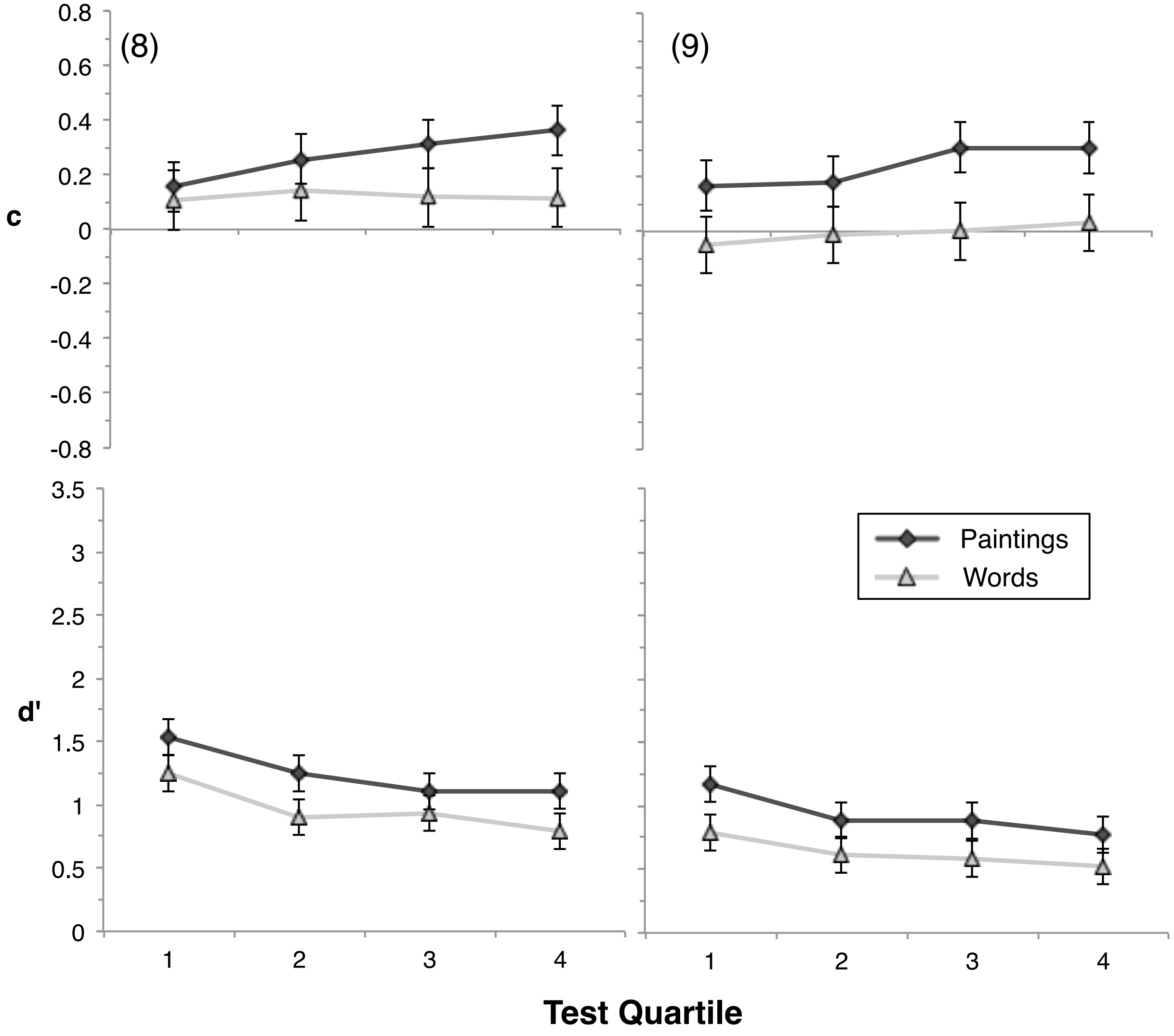

Figure 11. **Mean response bias (c; top row) and sensitivity (d'; bottom row) for paintings and words in each 48-item test quartile in Experiments 8 and 9, wherein item type was manipulated between subjects.**

Error bars are 95% within-subjects confidence intervals (Loftus & Masson, 1994) calculated based on the results of individual repeated measures ANOVAs.

Discussion

A multitude of different statistical comparisons could be conducted and presented for the quartile results depicted in Figures 4 through 11. However, to address every single significant difference or lack thereof would be tedious, difficult to interpret, and likely add little to overall understanding of the MBBE. Indeed, results from the repeated measures ANOVAs are amply cumbersome, let alone even smaller-grained interquartile comparisons. As such, the discussion will focus largely on the general patterns in these data. The reader interested in more specific statistical comparisons can, to some extent, approximate these answers by examining the relevant quartile graphs; error bars, as mentioned above, are 95% CIs (Loftus & Masson, 1994), so at the within-experiment level, one can infer that two means with error bars that do not overlap or overlap very little (at a rough approximation, by less than half) differ significantly at the .05 level. Hit and false alarm data were included largely for supplementary purposes and will not be uniquely discussed.

Some overall patterns were indeed apparent in the quartile graphs for both sensitivity and response bias. As previously mentioned, there was no strong, known basis on which to hypothesize about interquartile differences in response bias, so the observed patterns – or, for that matter, the existence of any patterns at all – were unanticipated and intriguing. Taking the response bias analyses as a whole (Figures 6, 7, and 11), there was a striking tendency for mean c for paintings to become more conservative (i.e., increase) across quartiles, with remarkably few data points deviating from this directional trend; arguably, only the result for the final quartile in Experiment 1 (Figure 6) strayed from this pattern to a notable extent, with c being lower than in the two preceding quartiles. There are certainly other cases where the difference between two adjacent quartiles would not achieve significance, but the first and last quartiles differed markedly in almost every single case. Remarkably, this pattern was even apparent in the data for the first and last individual test items. Among participants whose first test item was a painting, almost exactly half (49%) endorsed said painting as old, while paintings presented in the last test position were far more likely to be rejected (only 37% were called “old”). In contrast, words in the final test position were approximately as likely to be endorsed as “old” as words in the first test position (60% vs. 57%, with overlapping 95% CIs).

As there was no obvious basis upon which to predict changes in c across quartiles, possible explanations for the observed pattern are, necessarily, highly speculative. If there is some known widespread phenomenon whereby bias changes in a predictable way over the course of a recognition test, it was not uncovered in the literature, which was incredibly sparse in this regard. One study was found in which response bias became more liberal over the course of the test, specifically in a recognition memory study using three-digit numbers as stimuli (Donaldson & Murdock, 1968). With respect to the passage of time more generally, there seems to be a tendency for participants to respond more liberally with increasing delays between study and test (although the delays at which such effects emerge are typically on the order of days or weeks rather than minutes; e.g., Gehring, Toglia, & Kimble, 1976; Singer & Wixted, 2006). The results of the quartile analyses in the current experiments, therefore, were not only unanticipated given the relative dearth of research in this area, but if anything exhibit the opposite pattern of what one might expect based on the study-test delay findings.

The increase in conservatism across quartiles associated with paintings was more apparent in the results for the within-subjects experiments (Figures 6 and 7) than for the paintings-only groups in the two between-subjects experiments (Figures 11). Within the former group of experiments, the trend of increasing conservatism over time appeared particularly noticeable in experiments 4, 5 (Figure 7), 7, and the within-subjects group in experiment 8 (Figure 6). Furthermore, in the former three cases, c for paintings was statistically neutral in the first quartile. This was in clear contrast with the rest of the experiments – both within- and between-subjects – all of which were associated with conservative responding to paintings from the first quartile onward. Although the overall structure of Experiments 4, 5, and 7 was comparable to the others, these three studies do, in fact, share a feature that the other experiments do not: namely, experiments 4, 5, and 7 all included an extra task at study beyond the usual "try to remember this for the memory test later". Experiments 4 and 5 both included orienting tasks wherein participants were asked to judge the subjective pleasantness of words and paintings, and Experiment 7 was the reminding study, in which participants were instructed to press the spacebar each time a stimulus elicited a strong and spontaneous experience of being reminded of a previous item.

Although the study task was not the same in Experiments 4, 5, and 7, the fact that these three experiments were collectively unique both in having a study task at all and in yielding a neutral bias for paintings in the first quartile of the test suggests that something about this setup may be important in understanding the conditions that lead to the MBBE. One seemingly plausible scenario relates to item context. Although the focus of recognition tests is typically on the items themselves, the idea that various elements of the context in which learning occurs are also encoded and can exert appreciable influence over later memory performance has been a frequent topic of study in the memory literature, and seems at this point fairly uncontroversial. To name only a few examples, the importance of context has been illustrated in cases like the context-reinstatement effect dating back to Fisher and Craik (1977), in which memory performance is enhanced when the test context is similar to the study context; state-dependent learning effects, in which items studied in one state of consciousness, for example while intoxicated, will be better remembered in that same state (e.g., Eich, 1980); and effects of semantic context, such as in the paired-associates recognition tasks described by Tulving and Thomson (1973) wherein recognition performance was superior for target words that were accompanied at test by the same word they had previously been studied with.

Context effects have been explained in different ways according to different models and theories, but evaluating the validity of these explanations is beyond the scope of the current paper. For current purposes, it is sufficient to think of such effects in terms of the degree of match between the retrieval cue and the contents of memory, in line with accounts of context effects in recognition memory like that of Murnane and Phelps (1993, 1994, 1995). When a studied item is encoded, the resultant memory trace is thought to include not only features of the item itself but also certain contextual elements, which might include things like spatial location, idiosyncratic personal associations, or, in the case of several of the MBBE experiments, features of an associated task and the processes involved in performing it. To illustrate with a simple example, consider presentation of the word “river” in a recognition memory study list in which the task is to read the word aloud – the ensuing trace will include not only some representation of the word itself, but also representations of contextual elements such as the act of reading the word, hearing one’s own voice, the fact that the word was presented in black font, etc.

If “river” is later encountered on the test list, the probability of successfully recognizing it as “studied” will depend on the overlap between the cue used to probe memory and the existing representation. According to this interpretation, then, successful recognition will tend to be more likely in a condition wherein test words are also presented in black font and associated with a “read aloud” task than a condition in which words are presented in red and participants are only asked to complete the recognition task. This type of context effect on accuracy is not so relevant for current purposes, particularly in the case of the orienting task experiments, because these tasks are themselves designed to boost accuracy. What is more important is the general implication that the degree of similarity between study and test contexts can influence the final recognition decision.

With this general idea in mind, the scenario in the MBBE experiments may be conceptualized as follows. Studied items are associated with their respective contexts in memory, which in the case of Experiments 4, 5, and 7, might include details about the associated task. Encoded contextual details could be idiosyncratic, such as a representation of thinking "that painting looks like my childhood home, so I'll judge it as pleasant", or simply some representation of the task at hand (e.g., keeping the reminding task in mind). Test items, too, will be represented in memory in association with their contexts; although encoding at this stage may not be intentional, it will still occur to some extent, especially considering the active nature of the task. It seems plausible that encoded representations of studied items and their contexts might, on average, differ more substantially from encoded representations of test items when there is an additional task at study relative to experiments wherein the only task is to memorize the items. These differences could potentially influence the recognition decision process in a few ways. The speculative account presented below borrows elements from a few different models and frameworks of recognition memory, but was probably most influenced by global matching models (e.g., Clark & Gronlund, 1996; Humphreys, Pike, Bain, & Tehan, 1989), particularly with respect to terminology and some more general ideas about item representation.

Consider the beginning of the test phase in a study like Experiment 4, which, to reiterate, included a pleasantness orienting task at study. The contextual elements that are encoded along with individual test items might not, initially, tend to overlap very much with context representations from the study phase (or from other experiences). Disregarding baseline familiarity for the moment, if the first or second test item produces a high degree of match with a pre-existing representation, this can quite reliably be interpreted as indicating the item in question was in fact studied – if study and test context representations are highly dissimilar, familiarity is unlikely to be attributable to context. As the test proceeds, however, the likelihood that a probe representing the current item and its context will overlap with a pre-existing representation will tend to increase, because many aspects of context remain fairly consistent across the test and will therefore be shared among items. A high degree of match between the current probe and pre-existing contextual representations might then produce a spuriously high "strength of evidence" value even if the item itself has not been previously encountered, which may (but need not) be subjectively experienced as a sense of familiarity.

To return to the observed response bias patterns, then, the apparent tendency for participants to start the recognition test with little or no bias toward either response when items were studied in association with an additional task might be related to the low degree of context overlap at this stage. If an item early in the test list yields strong evidence of oldness, this will tend to be diagnostic of its presence on the study list, so there is no obvious reason for the individual or the memory system to interpret such evidence in a conservative manner. Later in the list, in contrast, this kind of evidence strength may be less diagnostic because more items associated with overlapping contexts have been added to memory. This might promote more cautious interpretation of memorial evidence and perhaps conservative responding, potentially explaining the tendency for c for paintings to increase over the course of the test. In experiments wherein studied items are not associated with comparably distinguishing context features – that is, in the majority of the MBBE experiments, which did not include any additional tasks at study – this conservatism might manifest immediately because of the relatively higher degree of context overlap between study and test items.

The speculative account presented thus far could potentially fit with the response bias trends observed for paintings in the quartile analyses, which were described above, but explaining the corresponding results for words presents an additional challenge. With respect to response bias, the data for words did not yield any discernible consistent cross-quartile pattern (Figures 6, 7, & 11). C tended to remain liberal throughout the test in the within-subjects studies and close to neutral in the between-subjects studies, but fluctuated apparently randomly within these ranges. These results parallel the overall findings in the current research line, specifically the observation that average response bias for words was liberal in the within-subjects experiments (Figure 2a) and closer to neutral in the between-subjects studies (2b). These findings have not been extensively addressed in the current paper, as the paintings data are of more central interest, but at this point it is useful to briefly mention that the current working explanation for liberal responding to words in the within-subjects context is that this is essentially a by-product of conservative responding to paintings, which is thought to be the dominant effect. The tendency to respond "new" for paintings may lead participants to feel they are not endorsing enough items, and perhaps they attempt to compensate for this by taking a more liberal approach with words. The observation that the response bias results for paintings appear quite similar regardless of whether the experiment also included words, whereas the reverse is not true, is largely consistent with such an explanation.

The quartile analysis results for words, then, were not particularly surprising in themselves, but any explanation that attempts to explain the paintings data will also need to take the words data into account. In the context of the speculative account presented above, if the tendency for c to increase over the course of the test for paintings is indeed attributable to something like increasingly overlapping context representations, there must be some reason words are not susceptible to the same effect. There may be a number of mnemonically important differences between words and paintings as stimuli, and some of these differences are discussed more thoroughly in the final chapter of the current paper. Any one of these differences might be relevant to understanding response bias differences between the two types of items, but one possibility that seems to fit particularly well with the kind of scenario hypothesized above relates to the nature of item representations.

There is mounting evidence to suggest that words and non-word stimuli are processed and/or treated quite differently by the memory system and that effects regularly observed in the recognition memory literature, having been established primarily in experiments with words, may not necessarily generalize to non-word stimuli (Osth, Dennis, & Kinnell, 2014). Dennis and Humphreys (2001) proposed the idea that representations of individual words in memory may be relatively sparse and unitary, rendering them less susceptible to item-based interference. This type of interference, a central tenet of global matching models (Clark & Gronlund, 1996; Humphreys, Pike, Bain, & Tehan, 1989), can essentially be thought of "noise" resulting from representations of previous items in the experiment that may complicate the recognition decision process or impair performance. Experiments conducted by Dennis and Humphreys (2001) as well as follow-ups by Kinnell and Dennis (2012) and Osth and colleagues (2014) yielded results consistent with the idea of words – at least familiar words – being represented in memory in a unitized fashion, while certain types of non-word stimuli, such as novel faces and fractal images, may be associated with more overlapping representations. If the words used in the MBBE experiments are indeed represented unitarily in memory and therefore relatively invulnerable to interference from other items in the experimental context, the context similarity considerations discussed above may not be as applicable. The main source of noise for words under such a model would be a different type of context noise, namely that resulting from the variety of contexts in which they have been encountered prior to the experiment (Dennis & Humphreys, 2002). Unlike noise from other test items and their experimental contexts, context noise from pre-experimental familiarity would not be expected to change over the course of the test, which could explain the absence of any cross-quartile response bias pattern for words.

In contrast with the response bias results, the sensitivity results were fairly uncomplicated, with the only standout pattern being the expected tendency for d' to decrease over the course of the test (Figures 8, 9, & 11). Although there were differences between paintings and words with respect to the actual magnitudes of d', these were as expected based on the overall analyses (Figure 3), and the extent and rates of decrease across quartiles were comparable for the two stimulus types. The stark contrast between d' and c in this regard lends further credence to the argument against d' differences as a possible explanation for the MBBE, which was presented in the discussion of the previous section regarding the meta-analyzed results for these two measures.

Receiver Operating Characteristics (ROCs)

The cross-experimental analyses presented thus far have all been based on hit and false alarm rates calculated according to the number of "old" responses participants give to old and new items. This amounts to treating responses in the recognition test as binary; however, as noted a few times in the current chapter, the experiments discussed thus far have in fact used a 6-point scale that allowed participants to respond "old" or "new" with varying levels of confidence. Disregarding these confidence judgments and treating responses of 4, 5, and 6 – corresponding to “maybe”, “probably”, and “definitely” studied – simply as “old” responses may be sufficient for most purposes, but the subjective experience associated with each of these responses presumably differs, and the processes underlying these differences may be important to understanding aspects of the MBBE and how recognition decisions are made more generally. For example, in calculating c and d’ in previous sections, responses of “3” and “4” were coded as categorically different, but both are low-confidence “maybe” responses. The only difference is that a response of “maybe studied” over “maybe not studied” implies the participant might have experienced some slight sense of oldness or familiarity that led them to lean toward this response (although it may well have been a guess). To group these responses separately is to treat a somewhat continuous scale, and potentially a similarly continuous underlying process or experience, as discrete. In doing so, one risks obscuring patterns that might not show up in the overall old/new data, a risk of particular concern in the case of an effect as stubbornly mysterious as the MBBE.

Receiver Operating Characteristics (ROCs) represent a means of both addressing this shortcoming of the previous analyses and visualizing the data in a manner that facilitates comparison with theoretical predictions and empirical data from the recognition memory literature. Generally speaking, ROC plots display the relationship between hit and false alarm rates across varying levels of the decision criterion, which is the SDT parameter response bias measures are meant to approximate (Egan, 1958; Macmillan & Creelman, 2005). In the case of confidence-weighted scales like the one used in the preceding MBBE experiments, the confidence levels can be thought of as criteria of varying stringency, and the number of points on the corresponding empirical ROC will be one fewer than the number of confidence levels (e.g., 5 in the case of the MBBE). These points can then be fitted to produce a function, aspects of which – such as overall shape and symmetry – can be interpreted in various ways in the context of existing models of recognition memory.

Some theories and models suggest specific predictions about such attributes of ROCs, and when ROCs from actual participants turn out as expected, this can be compelling evidence in support of the model in question (e.g., Yonelinas, 1994). Similarly, when empirical data are repeatedly inconsistent with predictions, this can cast doubt on the validity of a theory and shed light on its limitations (Yonelinas & Parks, 2007). Ideally, ROC analyses are conducted on a participant-by-participant basis such that the resulting functions include some measure of error and can be statistically compared, for example to model-predicted ROCs or, as would be particularly useful in the case of the MBBE given the frequent comparisons made between painting and word data, other empirical ROCs.

Such individual participant ROC analyses are ongoing, but the current section is a starting point, presenting ROCs (and their z-transformed counterparts) that were constructed by collapsing data across participants, but within experiments. This approach precludes certain types of analyses, so comparisons between the results for paintings and words should be interpreted somewhat cautiously, but can still be useful in understanding aspects of the data and speculating about possible explanations.

Method

ROCs were constructed for paintings and words in each MBBE experiment, with the calculations for an individual experiment being based on the total numbers of responses in the relevant categories across all participants in that experiment. There were essentially 24 response categories in the first stage of calculations, reflecting the numbers of responses made at each confidence level (6) to test stimuli of each status (2; old or new) divided by stimulus type (2; word or painting). These 24 totals were used to calculate proportions of responses made at each confidence level for all four possible item types (i.e., studied paintings, new paintings, studied words, and new words). Critically, these proportions were calculated in a cumulative manner, such that the first proportion reflected only high confidence "old" responses, in other words responses of "6"; the second proportion included both high and medium confidence "old" responses (i.e., responses of "5" or "6"), and so on. Each subsequent proportion included responses made at decreasing levels of "confidence in oldness", such that the final proportion included all responses to the item type in question, and was therefore always equal to 1.

These cumulative proportions served as the empirical data points in the plotted ROCs. As is standard practice, cumulative proportions for new items were plotted along the x-axis, while the corresponding y-axis values were based on studied items. These axes were labelled as "false alarm rate" and "hit rate", respectively, to reflect the idea that points on the plot can be thought of as the hit/false alarm rates obtained at varying criterion levels. To clarify, the leftmost point in the resulting scatterplot always represented the proportion of new items (x-axis) and old items (y-axis) that were given a high confidence "old" rating, i.e., a response of 6, and the rightmost point in all cases was (1,1), representing the proportion of responses to new and old items given at any confidence level (1-6) (although this point was not actually plotted).

ROC functions were plotted separately for paintings and words for each experiment, but both functions were combined into a single plot to simplify comparison. The proportions mentioned above were also z-transformed and converted to functions in the same manner, yielding z-ROC plots for each experiment as well. Proportions were also entered in the online software JROC (Eng, n.d.) to produce fitted functions for both the standard and z-transformed

ROCs, obtain estimates of the area under the ROC curve, and facilitate calculation of the parameters of the line of best fit for the z-ROCs (i.e., slope and y-intercept values).

Results

ROC and zROC plots are depicted in Figures 12 through 16. Individual plots within each figure include functions for both paintings and words and are labelled by experiment number (see Appendix A). Figures 12 and 13 show ROCs and zROCs, respectively, for the within-subjects experiments 1, 2, 7, and one group in experiment 8, and Figures 14 and 15 show these same plots for the subjective memorability studies (Experiments 3-6). Figure 16 depicts ROCs (left) and zROCs (right) for the two between-subjects experiments (8 & 9).

Individual points represent the cumulative proportions of hits and false alarms at each level of confidence described in the method section, while the smooth lines represent the fitted ROCs (or z-transformed equivalent) constructed using JROC (Eng, n.d.). Dotted diagonal lines in both ROC and zROC plots represent chance performance. zROC plots also include equations in the top left corner representing the lines of best fit.

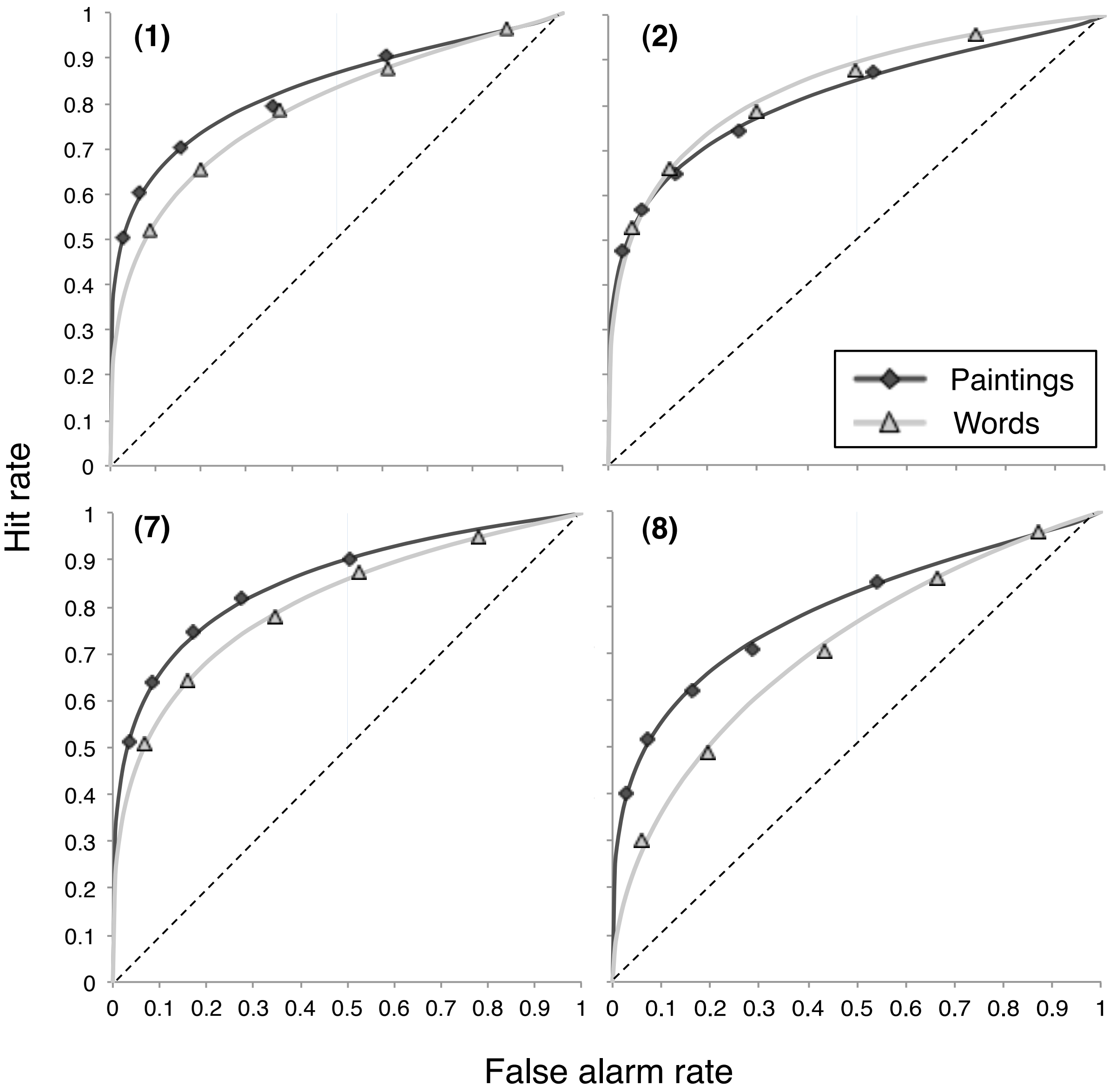

Figure 12. **Receiver operating characteristic (ROC) plots for paintings and words in within-subjects experiments 1, 2, 7, and 8.**

Individual points represent cumulative proportions of hits and false alarms at varying levels of confidence that an item is old, with confidence decreasing from left to right; for example, the second point from the left corresponds to the proportions of hits and false alarms when only the two highest confidence responses ("probably old" and "definitely old") are considered "old" responses. Proportions were calculated by collapsing across all participants. Smooth lines represent fitted ROCs constructed using JROC (Eng, n.d.) and dotted diagonal lines represent chance performance.

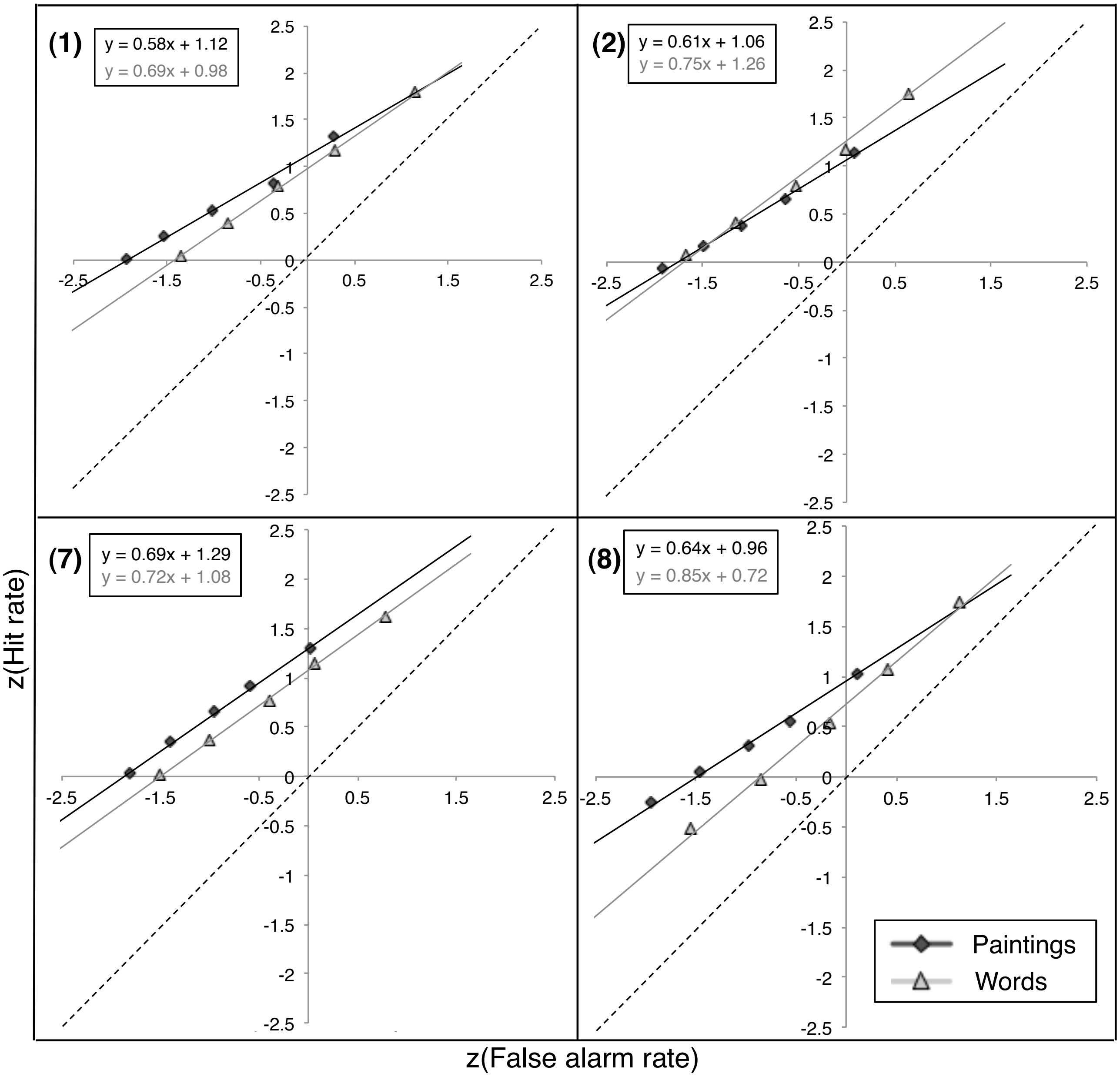

Figure 13. **Z-transformed receiver operating characteristic (ROC) plots for paintings and words in Experiments 1, 2, 7, and 8.**

Individual points represent z-transformed cumulative proportions of hits and false alarms at varying levels of confidence that an item is old, with confidence decreasing from left to right; for example, the second point from the left corresponds to the proportions of hits and false alarms when only the two highest confidence responses ("probably old" and "definitely old") are considered "old" responses. Proportions were calculated by collapsing across all participants. Lines represent z-transformed ROCs fitted using JROC (Eng, n.d.) and corresponding equations are presented in the upper left corner of each plot. Dotted diagonal lines represent chance performance.

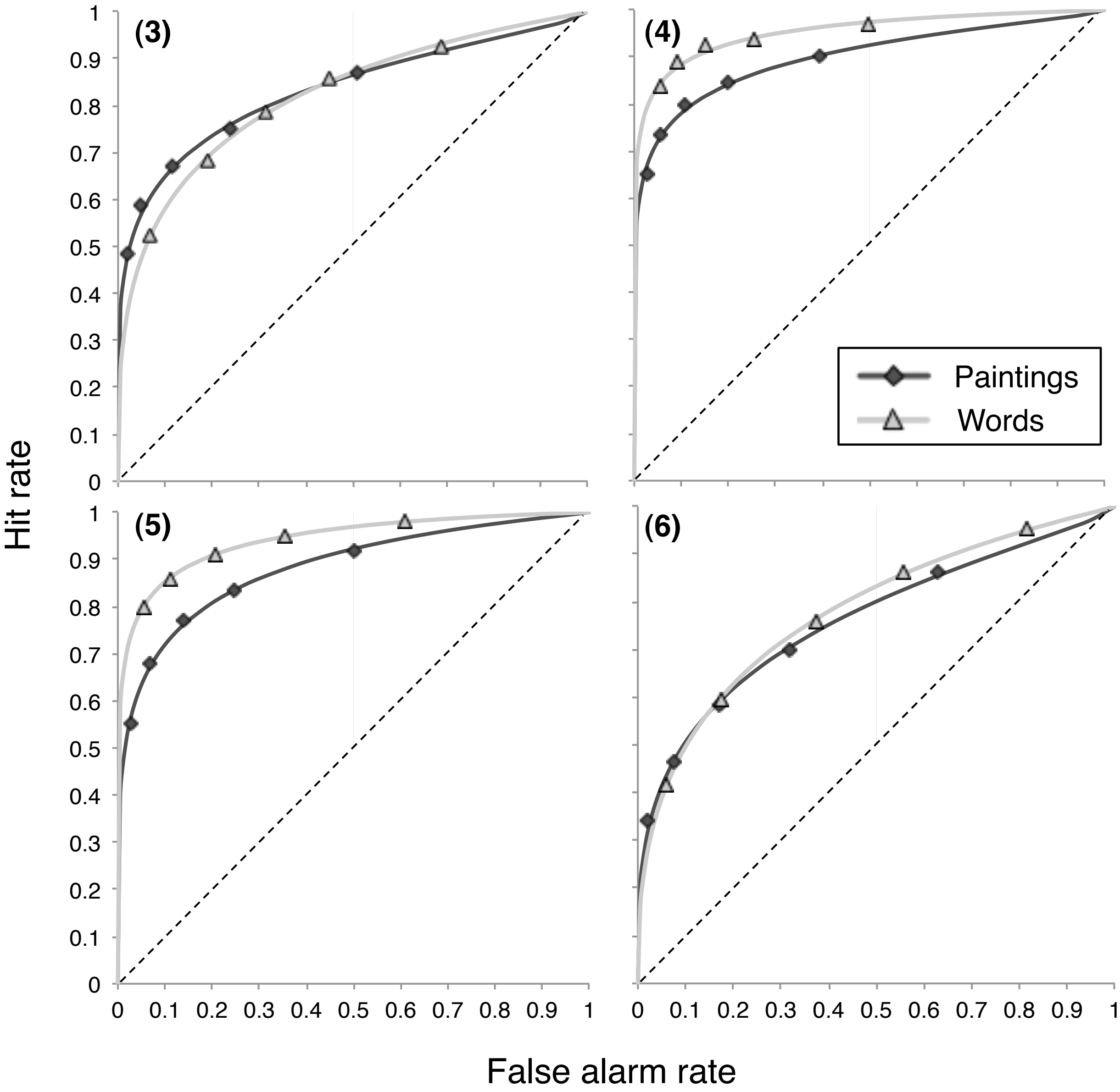

Figure 14. **Receiver operating characteristic (ROC) plots for paintings and words in the subjective memorability experiments (Experiments 3-6).**

Individual points represent cumulative proportions of hits and false alarms at varying levels of confidence that an item is old, with confidence decreasing from left to right; for example, the second point from the left corresponds to the proportions of hits and false alarms when only the two highest confidence responses ("probably old" and "definitely old") are considered "old" responses. Proportions were calculated by collapsing across all participants. Smooth lines represent fitted ROCs constructed using JROC (Eng, n.d.) and dotted diagonal lines represent chance performance.

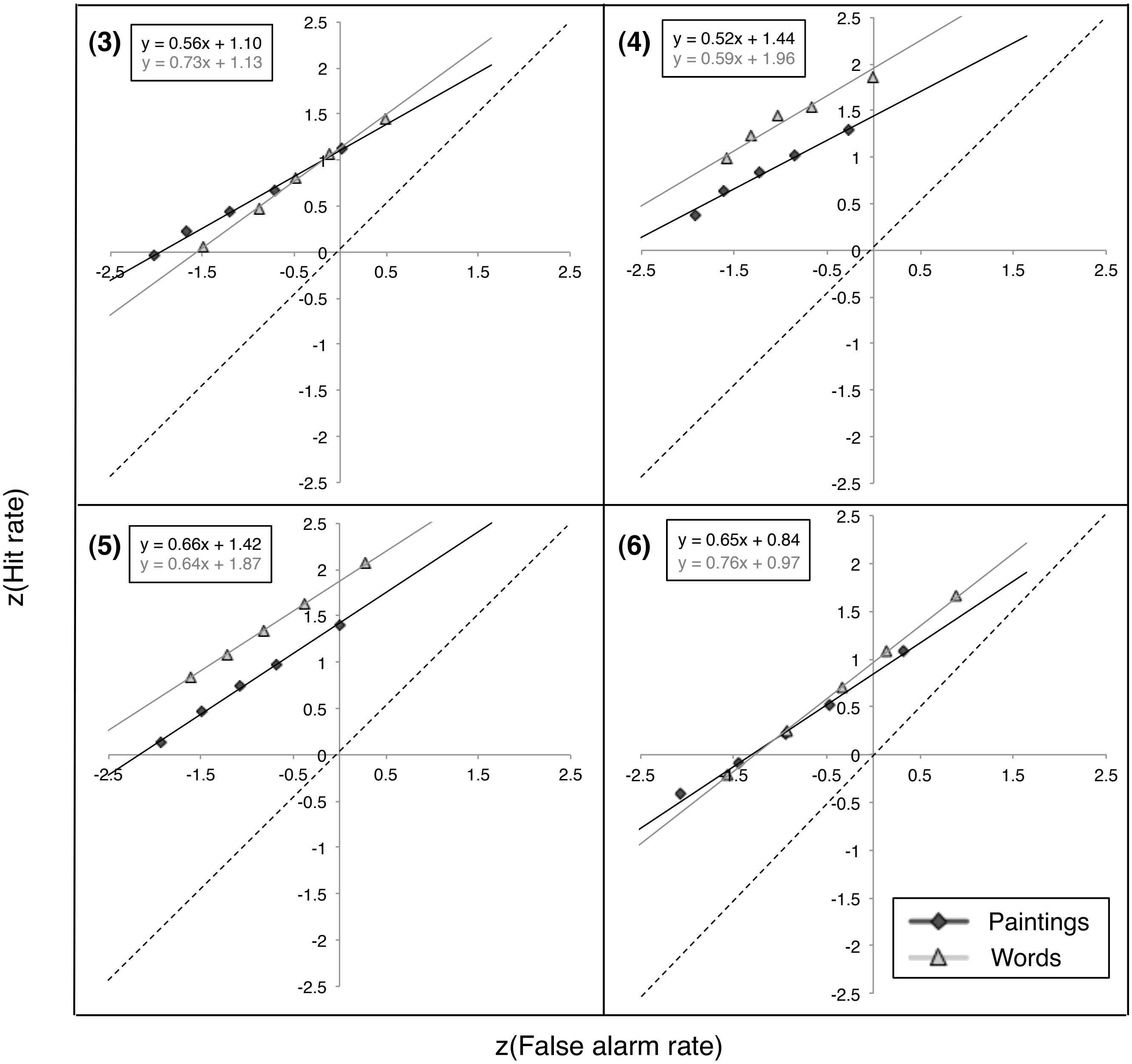

***Figure 15.* Z-transformed receiver operating characteristic (ROC) plots for paintings and words in the subjective memorability experiments (Experiments 3-6).**

Individual points represent z-transformed cumulative proportions of hits and false alarms at varying levels of confidence that an item is old, with confidence decreasing from left to right; for example, the second point from the left corresponds to the proportions of hits and false alarms when only the two highest confidence responses (“probably old” and “definitely old”) are considered “old” responses. Proportions were calculated by collapsing across all participants. Lines represent z-transformed ROCs fitted using JROC (Eng, n.d.) and corresponding equations are presented in the upper left corner of each plot. Dotted diagonal lines represent chance performance.

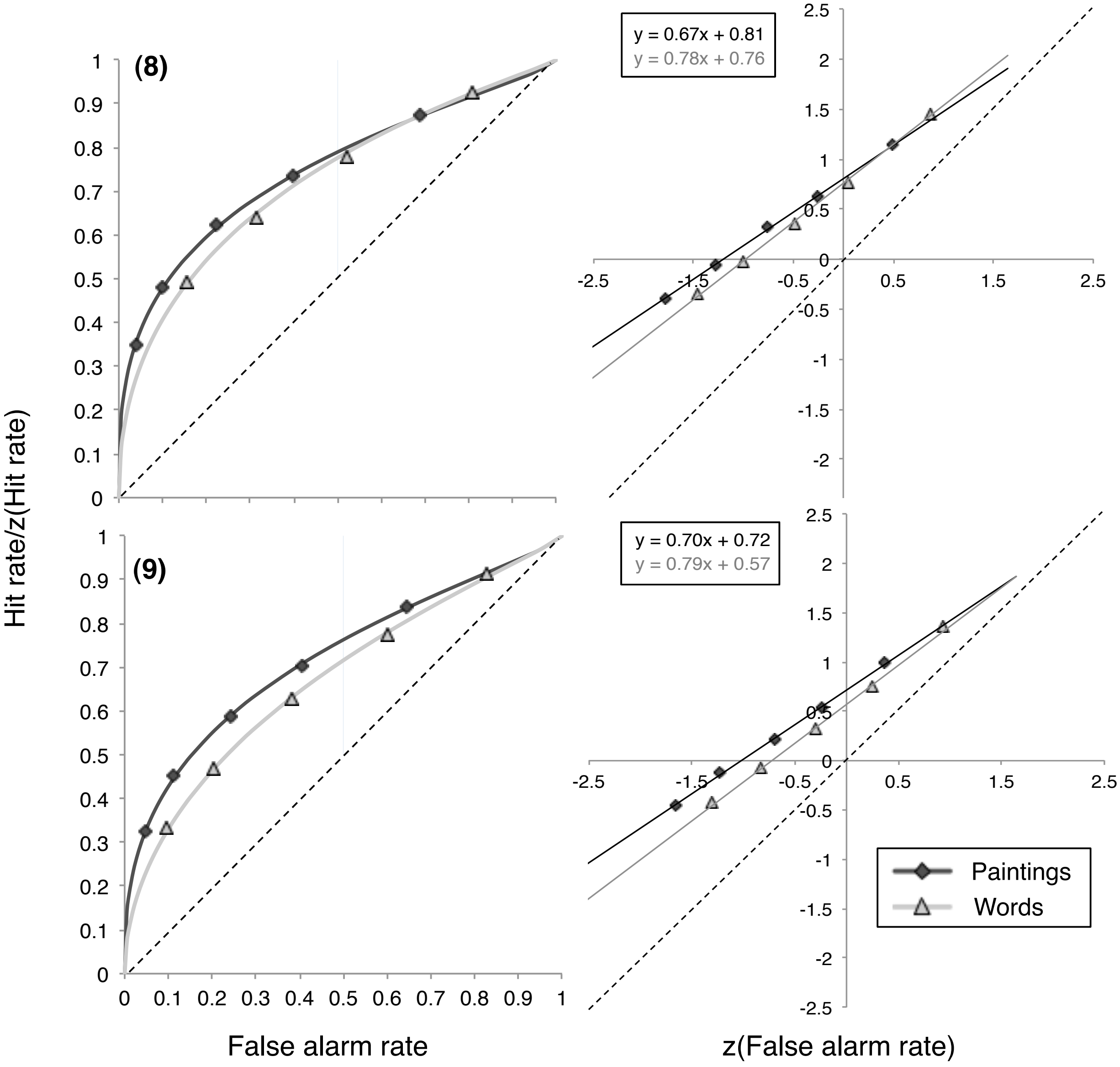

***Figure 16.* Receiver operating characteristic (ROC) (left) and corresponding z-transformed ROC plots (right) for paintings and words for two between-subjects experiments (8 & 9).**

Individual points represent cumulative proportions (or z-transformed equivalent) of hits and false alarms at varying levels of confidence that an item is old, with confidence decreasing from left to right; for example, the second point from the left corresponds to the proportions of hits and false alarms when only the two highest confidence responses ("probably old" and "definitely old") are considered "old" responses. Proportions were calculated by collapsing across all participants. Continuous lines are the result of fitting the abovementioned points using JROC (Eng, n.d.) and dotted diagonal lines represent chance performance. Equations correspond to the best-fitting lines shown in the zROCs.

Discussion

Similarly to the quartile analyses, the ROC and zROC results will largely be discussed in terms of trends that can be observed across several experiments as opposed to a detailed analysis of each plot. A few such trends are visually apparent. With respect to the ROCs (Figures 12, 14, & 16), the function for paintings curves more toward the upper left corner than the function for words in the majority of experiments; in other words, the area under the curve (AUC) tends to be greater for paintings. Greater AUC values indicate higher discriminability (e.g., Yonelinas & Parks, 2007), so this difference simply reflects the previously discussed tendency for discrimination/sensitivity to be superior for paintings in many experiments (e.g., Figure 2). Notable reversals of this trend are Experiments 4 and 5, wherein d’ was higher for words, manifesting as a greater AUC for words than paintings in these experiments (Figure 14).

Z- transformed ROCs can be thought of as a means of quantifying various aspects of the ROC’s shape, and when zROC functions are approximately linear, the parameters associated with those lines can be thought of as rough approximations of accuracy (y-intercept) and asymmetry of the ROC (slope; Yonelinas & Parks, 2007). The zROCs for the current experiments (Figures 13, 15, & 16) do not appear to deviate markedly from linearity, so such approximations are suitable in this case. Therefore, differences in discriminability between paintings and words (or the relative lack thereof, as, e.g., in Experiment 6) that manifested as AUC differences in the ROC plots are also represented in the y-intercept values in the zROCs (Figures 13, 15, & 16), with higher y-intercepts corresponding to higher sensitivity.

The slopes of the zROCs, which index the asymmetry of the ROC functions, were directionally lower for paintings than words in all cases except Experiment 5 (Figure 15). Interpreted in SDT terms, these slopes represent the ratio of the standard deviation of the underlying lure strength distribution to the standard deviation of the target strength distribution, with values below 1 indicating the latter distribution is more variable. That this was the case for all zROCs in the current paper is not unusual; recognition memory experiments consistently yield zROC slopes below 1, with most in the range of 0.5-0.9 (Glanzer, Kim, Hilford, & Adams, 1999; Yonelinas & Parks, 2007), and all slopes for paintings and words were indeed within this typical range. This indicates that target distributions were more variable than lure distributions for both stimulus types, but that this true was to an even greater extent for paintings, as signified by the lower slope values associated with these zROCs (Figures 13, 15, & 16). Prior to speculating as to what the zROC slope results might mean as far as underlying memory processes, especially in the context of the MBBE, it should be reiterated that these slope values were not statistically evaluated nor compared. That the observed tendency toward lower slope values for paintings emerged in 9 of ten comparisons seems promising, but is at this point only a directional trend; whether it is statistically meaningful remains to be seen.

The previously mentioned regularity of zROC slopes below 1 in recognition memory studies is, of course, inconsistent with EVSD assumptions, suggesting such a model cannot adequately capture the phenomenon of recognition. The question of what type of model might be better suited to this task has been, and continues to be, the topic of much heated debate in the literature. Probably the majority of this debate has occurred between proponents of two types of models in particular: Unequal Variance Signal Detection (UVSD) and Dual Process Signal Detection (DPSD) models (e.g., Wixted, 2007; Yonelinas & Parks, 2007). The characteristic assumption of UVSD is that of a single underlying evidence variable, often described in terms of a “degree of match” between some memory probe and pre-existing traces, upon which recognition decisions are based (Clark & Gronlund, 1996; Dennis & Humphreys, 2001). Values representing this match are normally distributed for both targets and lures, but both the mean and the variance are higher in the target distribution (Mickes, Wixted, & Wais, 2007); this latter assertion accounts for the prevalence of zROC functions with slopes below 1 according to UVSD.

DPSD models, in contrast, posit two underlying systems or processes: recollection, whereby recognition decisions are based on remembering certain details of a learning experience, and familiarity, which is more of a vague sense of previous encounter that is unaccompanied by specific details (Yonelinas, 1994; Yonelinas & Parks, 2007). DPSD theorists suggest that while responding on the basis of familiarity would be expected to produce zROC slopes of 1, it is unlikely that all decisions in a recognition task will be based only on familiarity, and slopes below 1 simply reflect some involvement of recollection (Yonelinas, 1994; Yonelinas & Parks, 2007).

These are not the only two possible explanations for zROC slopes less than one, and researchers who subscribe to other categories of models, like diffusion models, have reported data seemingly inconsistent with both DPSD and UVSD predictions (e.g., Starns, 2014). However, the DPSD-UVSD debate remains a central issue in the literature, and given its historical association with the issue of zROC slopes, the current paper will focus on these two models in speculating about the observed differences between paintings and words. To understand why zROC slopes associated with paintings are lower than those for words in a UVSD context necessitates some explanation as to why the ratio of variability in the underlying target and lure distributions might be higher in the former case. While the UVSD model itself makes no claims as to why the target distribution should be more variable than the lure distribution in the first place, one posited mechanism is encoding variability, the idea being that study will not necessarily increase the strength of all items by the same amount (Wixted, 2007). It seems quite plausible that in the current experiments, this might have been true a greater extent for paintings than for words, particularly when baseline familiarity levels are considered.

The words used as stimuli in all studies were fairly common nouns, and it is unlikely any of them would be unfamiliar to undergraduate student participants. The paintings in the stimulus pool, by contrast, were intentionally selected for their relative obscurity in hopes that they would be mostly novel to participants. Further, and perhaps more importantly, participants' previous experience with examining, processing, and even actively memorizing words and paintings probably differed a great deal. With this in mind, it makes sense that words might have been encoded in a more consistent manner than paintings; some participants may even have gone into the task with a pre-existing strategy for word memorization, while this seems less feasible for paintings.

Further support for such an account comes from the same three experiments that yielded anomalous results in the quartile analyses discussed above. zROC slope differences were least pronounced in the three experiments with an additional task at study: Experiments 4 and 5, which both used orienting tasks intentionally aimed at minimizing encoding strategy differences between words and paintings, and Experiment 7, which used the reminding task. These findings, then, seem to fit quite well with a UVSD account of zROC slope differences that suggests encoding variability as the mechanism responsible for higher variability of the target distribution. The target distribution tends to be even more variable for paintings than words because they are encoded in a less consistent manner, perhaps partly due to participants' relative inexperience with paintings, but a study task may minimize such inter-item encoding differences, thus making target distributions more comparable and manifesting as more similar slopes. This is, of course, purely speculative, and it is important to emphasize that because the zROC slope represents a ratio, the observed lower slopes for paintings as compared to words could theoretically be attributable to a less variable lure distribution for paintings instead of (or in combination with) a more variable target distribution as suggested above. Indeed, this also seems quite sensible in light of the previously mentioned differences in baseline familiarity: lure paintings might tend to be more similar with respect to familiarity than lure words, some of which may be more familiar than others due to differences in frequency and prior experience. This distinction cannot be made based on the existing data.

With respect to DPSD, the observed zROC differences would suggest that recognition decisions for paintings tend to be based on recollection more often than decisions for words. This, too, has intuitive appeal; the paintings used were perceptually complex, evocative, and vivid, all of which seem more conducive to a more elaborate recollective experience than common nouns presented in black font. Familiarity is often described as fairly rapid, automatic, and more perceptually based, while recollection is thought to be slower and more intentional (Atkinson & Juola, 1974; Mandler, 1980). Some theorists describe recollection as a resource-intensive backup process which is only initiated when the initial familiarity assessment is inconclusive (e.g., Atkinson & Juola, 1974) while others contend the two processes occur in parallel, with decisions being made more rapidly on average when familiarity “wins” the race (Mandler, 1980; Yonelinas & Jacoby, 1994). The idea of recognition decisions to paintings more often being recollection-based seems compatible with either explanation given their relatively low baseline familiarity. Under the former description, the familiarity assessment might often be inconclusive for test paintings, perhaps because they overlap little with pre-existing memory traces. This would more often necessitate initiation of more deliberate processes of recollection when paintings are viewed, in contrast with words, for which familiarity might tend to be sufficient. Similarly, if the two processes operate in parallel, familiarity might be slower to reach some threshold and therefore less likely to "win" the race and support a decision when paintings are viewed. This could conceivably be attributable to the sheer number of perceptual details in such stimuli, a relative dearth of comparable representations in memory, or – potentially relevant to the MBBE itself – a higher threshold for these stimuli than words. Recollection processes might then result in a decision before the familiarity threshold is reached.

Ultimately, zROC data are insufficient for distinguishing between UVSD and DPSD accounts, as the two make nearly identical predictions in this regard (Yonelinas & Parks, 2007). Other data presented thus far are similarly ill equipped for supporting a conclusive decision in this regard. However, the DPSD account of zROC slope differences between paintings and words does imply a specific prediction: namely, if responses to paintings are more often based on recollection, they should also be associated with longer response times than words. This being the case would not rule out the UVSD explanation, but if such a pattern were not observed, it would raise serious doubts regarding the applicability of the DPSD account in this context. Response time comparisons between paintings and words are presented in the following section.

Response Time Analyses

Method

Response times (RTs) associated with correct responses (i.e., hits and correct rejections) were ordered from least to greatest for each participant in each MBBE experiment. This was done separately for paintings and words, yielding four sets of RTs per participant. RTs exceeding 17 seconds were excluded as outliers based on examination of RT frequency distributions (not presented), amounting to less than 1% of all correct responses. Each set was then divided into four bins. A lower cut-off of 16 responses per category was set at this stage such that no means would be calculated based on fewer than four responses, and twelve participants were excluded on this basis overall. Bins were constructed as evenly as possible such that, for a given item/RT category, the number of responses in each of the four bins for a given participant never differed by more than one. To illustrate, if a participant correctly rejected 30 paintings, the ranked RTs were divided into two bins with eight RTs and two with seven RTs. Specific bin assignments in such cases were quasi-random such that no bin consistently comprised more or fewer RTs than the others. Mean RTs were then calculated individually in the four bins, yielding 16 such averages per participant (e.g., bin 1 word hits, bin 2 word hits, etc.). These participant means were then averaged within each experiment and plotted accordingly along with 95% confidence intervals (CIs were within-subjects when applicable, i.e., in all cases but Experiments 8 and 9).

Results

Response time graphs are plotted in Figures 17-21. Mean RT in seconds (s) is shown on the y-axis while bins are labelled on the x-axis, with bin 1 representing the fastest quartile of responses and bin 4 representing the slowest quartile. Each figure presents results from two experiments, which are labelled as usual (see Appendix A). Individual rows in each figure represent the results for a single experiment, with hit RTs on the left and correct rejections (CRs) on the right. Results are also ordered as usual: Figures 17 and 18 depict results from within-subjects Experiments 1-2, and 7-8; Figures 19 and 20 show results for the subjective memorability experiments (3-6), which were also within-subjects; and Figure 21 shows results for Experiments 8 and 9, wherein item type was manipulated between subjects.

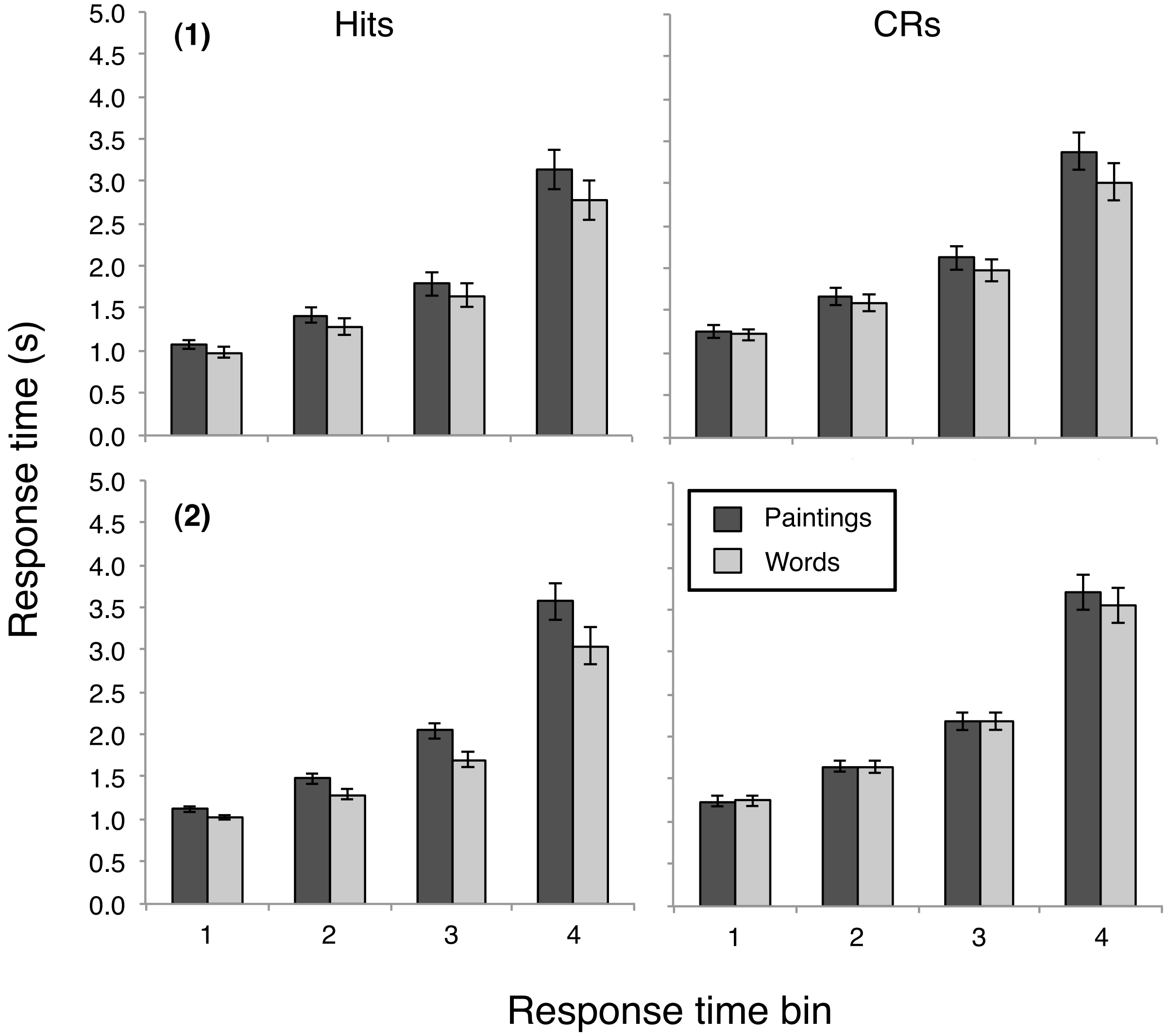

***Figure 17.* Mean response times (s) for correct responses (hits in left column; correct rejections [CRs] in right column) to paintings and words in MBBE Experiments 1 and 2.**

Means were calculated in each of four bins (x-axis), with bin 1 including the fastest quartile of responses and bin 4 including the slowest quartile. Error bars are 95% within-subjects confidence intervals (Masson & Loftus, 1994).

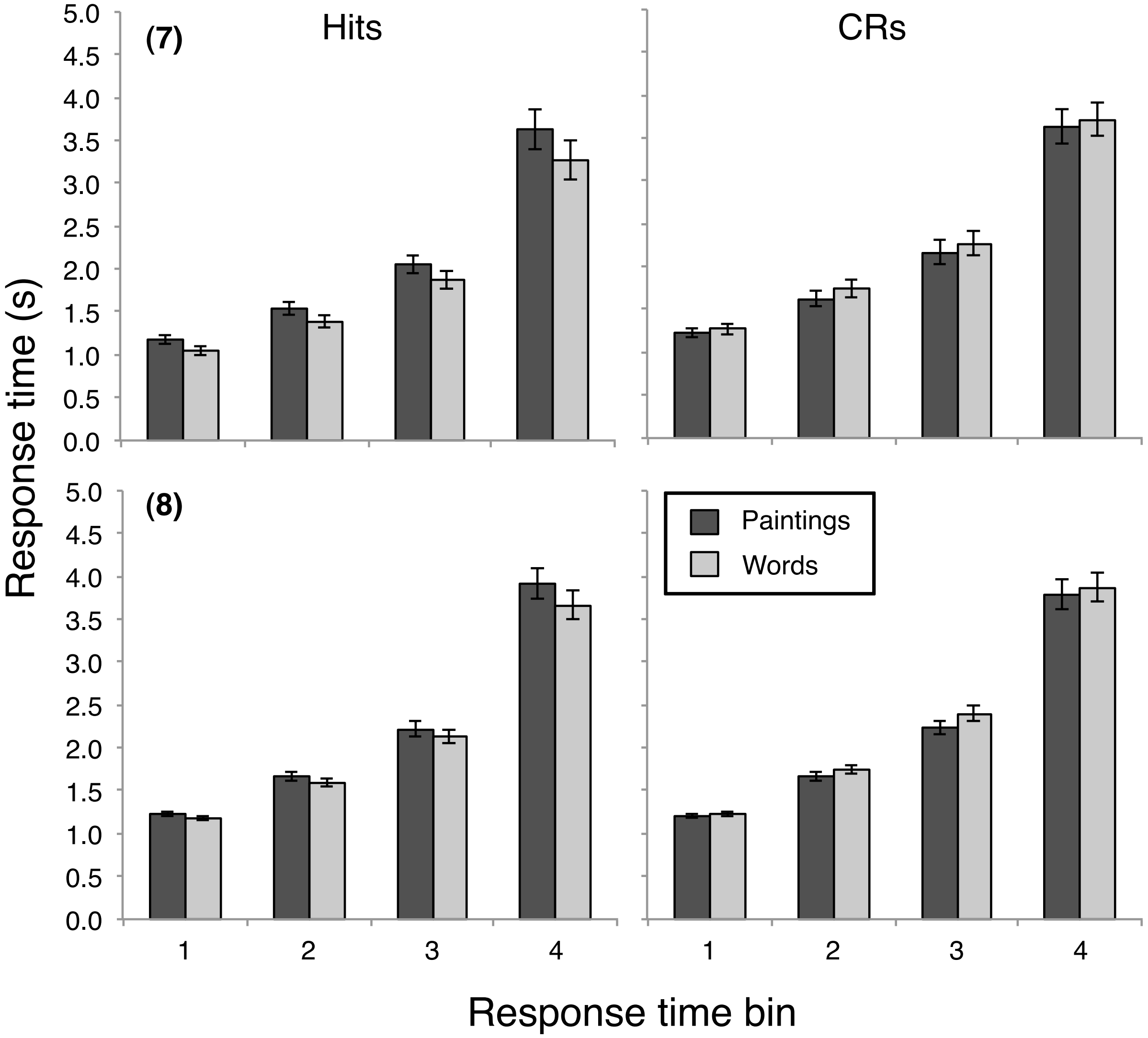

Figure 18. **Mean response times (s) for correct responses (hits in left column; correct rejections [CRs] in right column) to paintings and words in MBBE Experiments 7 and 8.**

Means were calculated in each of four bins (x-axis), with bin 1 including the fastest quartile of responses and bin 4 including the slowest quartile. Error bars are 95% within-subjects confidence intervals (Masson & Loftus, 1994).

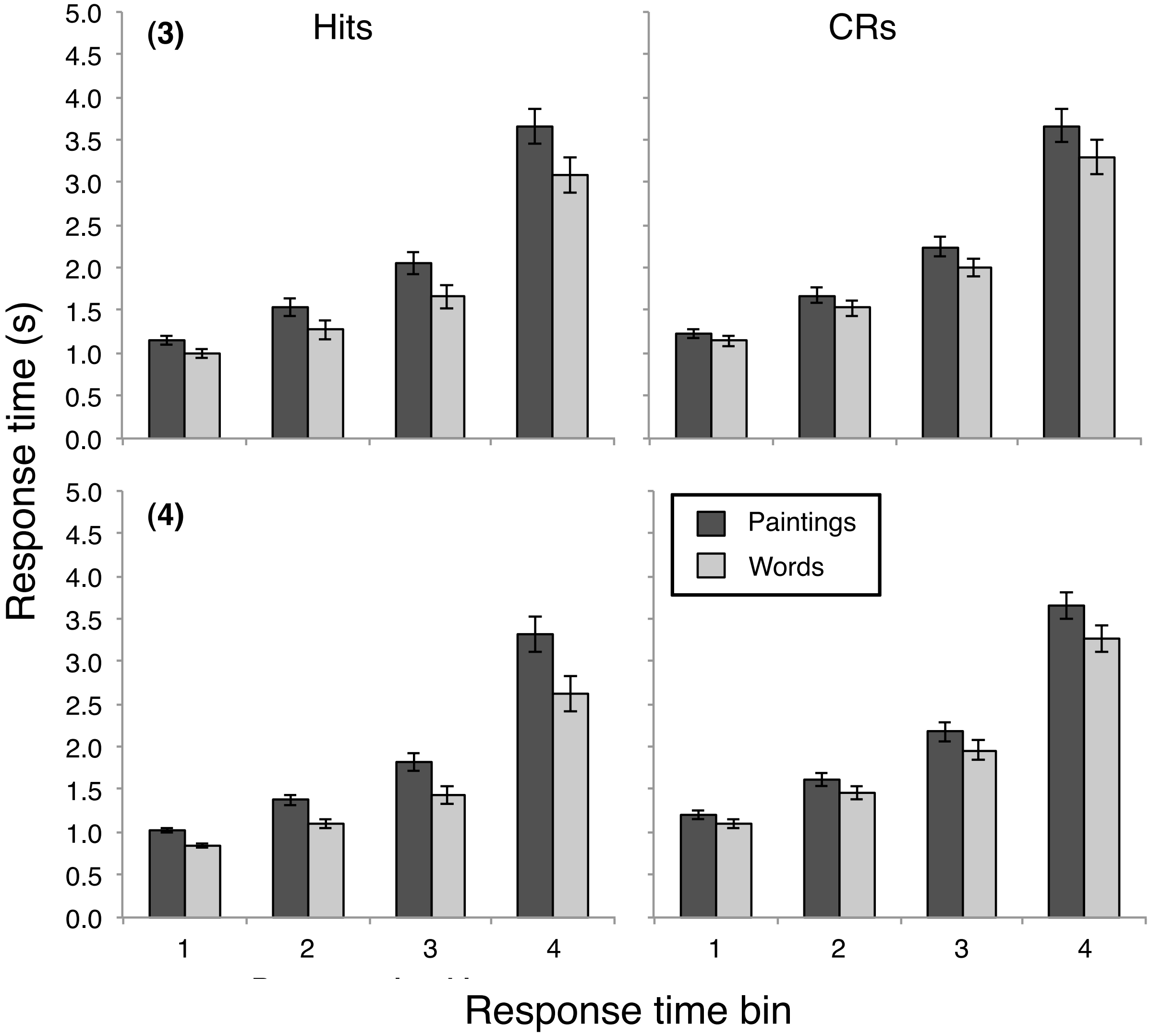

Figure 19. **Mean response times (s) for correct responses (hits in left column; correct rejections [CRs] in right column) to paintings and words in MBBE Experiments 3 and 4.**

Means were calculated in each of four bins (x-axis), with bin 1 including the fastest quartile of responses and bin 4 including the slowest quartile. Error bars are 95% within-subjects confidence intervals (Masson & Loftus, 1994).

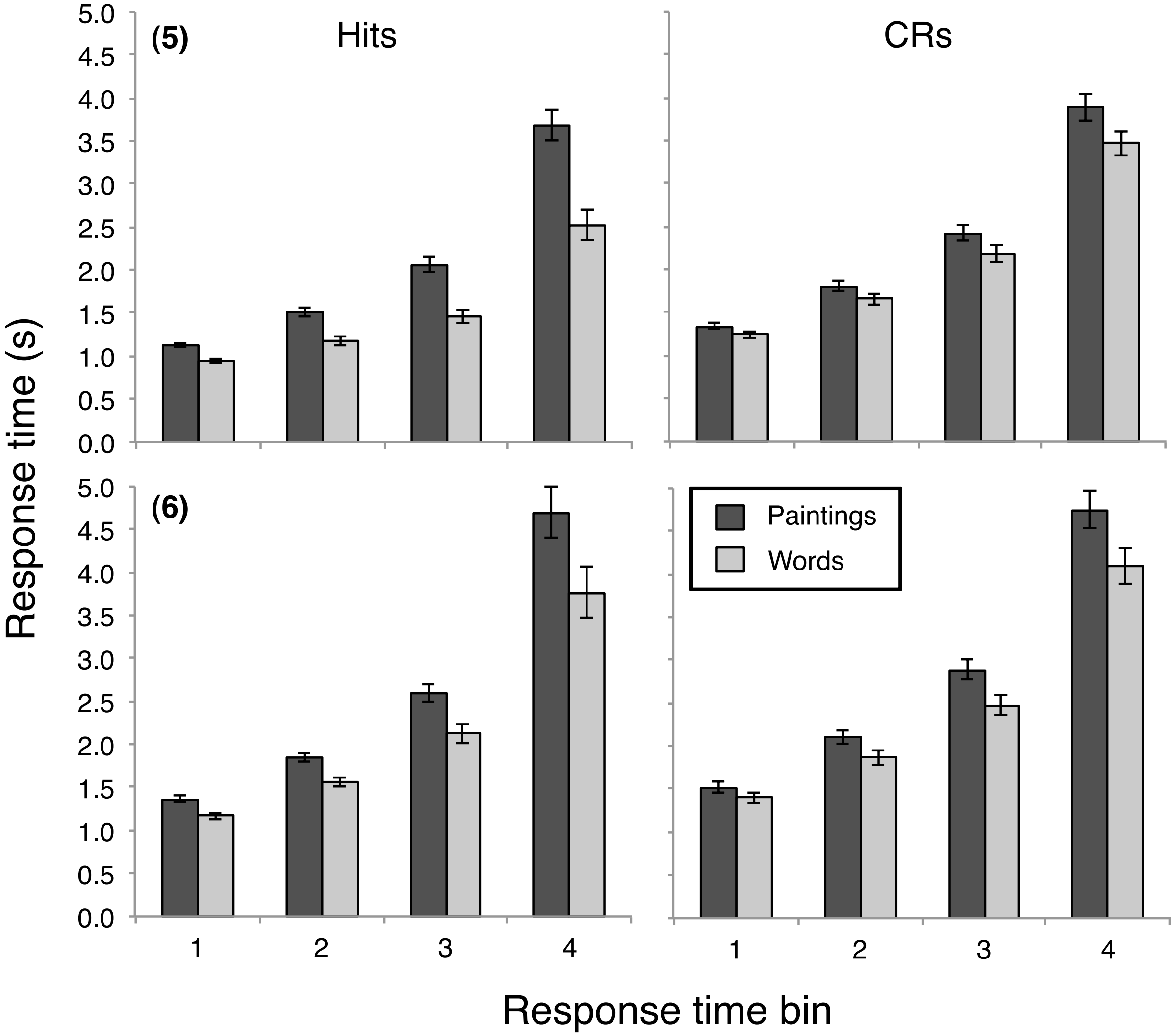

***Figure 20.* Mean response times (s) for correct responses (hits in left column; correct rejections [CRs] in right column) to paintings and words in MBBE Experiments 5 and 6.**

Means were calculated in each of four bins (x-axis), with bin 1 including the fastest quartile of responses and bin 4 including the slowest quartile. Error bars are 95% within-subjects confidence intervals (Masson & Loftus, 1994).

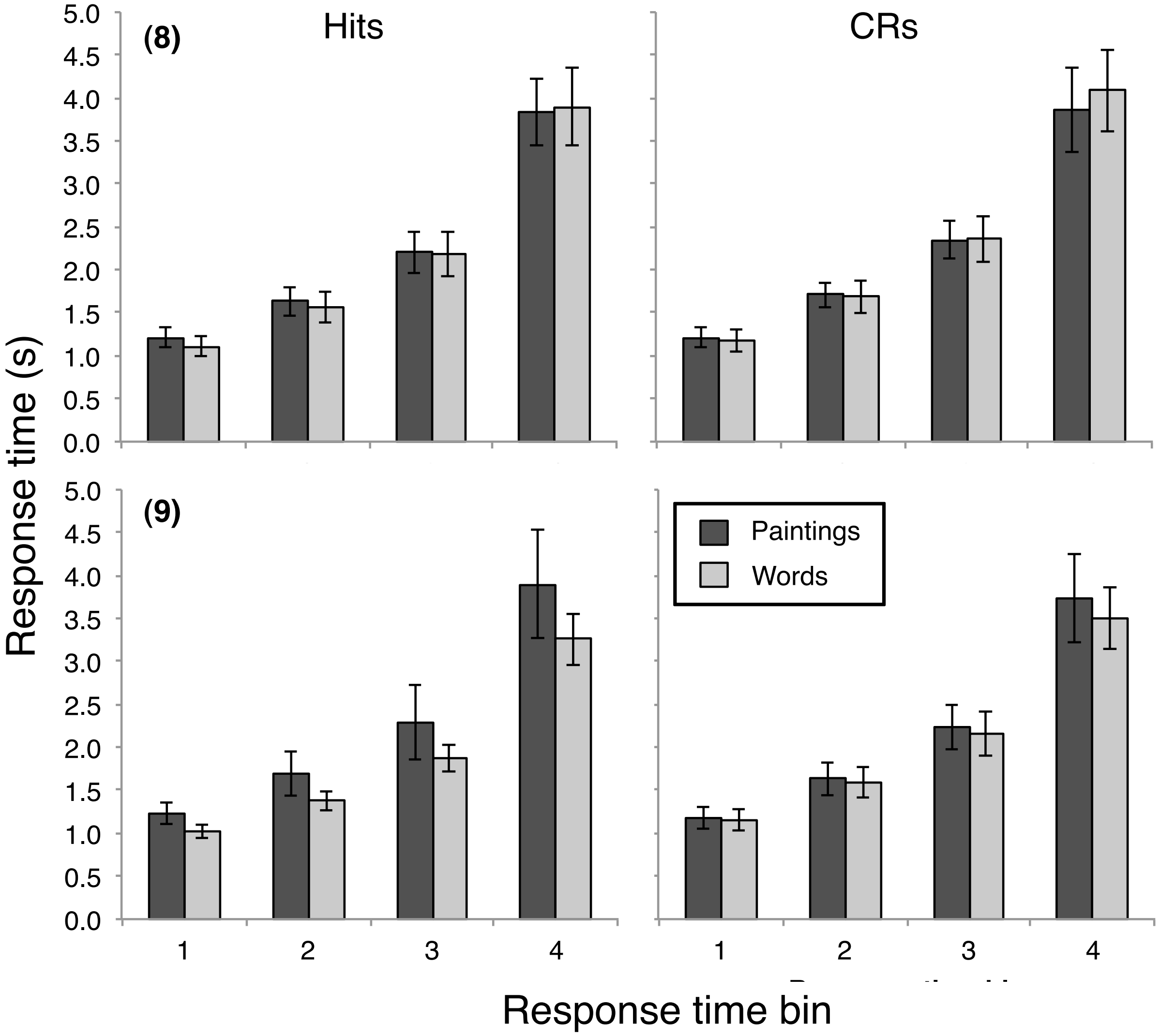

***Figure 21.* Mean response times (s) for correct responses (hits in left column; correct rejections [CRs] in right column) to paintings and words in MBBE Experiments 8 and 9.**

Means were calculated in each of four bins (x-axis), with bin 1 including the fastest quartile of responses and bin 4 including the slowest quartile. Error bars are 95% confidence intervals. Error bars are 95% within-subjects confidence intervals (Masson & Loftus, 1994).

Discussion

At the time RT analyses were undertaken, this was done so for exploratory purposes. The presented analyses are more recent, but initial construction of RT frequency distributions, calculations of means, etc. were all done with no hypothesis in mind. As such, a disclaimer should be made that the DPSD prediction introduced in the previous section – namely, that if the previously discussed zROC slope differences between paintings and words are explicable in terms of a DPSD account wherein judgments to paintings are more often recollection based, RTs to paintings should be longer – was post hoc not only in that the experiments had been conducted, but also with respect to this domain of analysis. With this in mind, the overall trend appears consistent with such a DPSD account: Mean RTs associated with correct responses were almost always directionally – and in many cases significantly – higher for paintings than words (Figures 17-21), although this was markedly less consistent in the between-subjects experiments (Figure 21). This trend appeared more reliable for hits than CRs (see, e.g., Experiments 7 & 8 in Figure 18), which also seems sensible in a DPSD context; recollection as it is typically conceptualized supports "old" decisions (but see, e.g., Rotello & Heit's [2000] discussion of recall-to-reject strategies).

With respect to differences between bins, there did not appear to be any overwhelming tendency for mean RT differences between words and paintings to be more pronounced for responses in a particular speed category. Although the differences were generally larger among the slowest responses in absolute terms, RTs in this bin were, of course, more variable. What is more important is the reliability of the observed RT differences, which 95% CIs show was comparable across bins (see Figure 19 for just two examples). A DPSD explanation positing more recollection-based judgments to paintings than words might predict less of an RT difference among the fastest responses, as these would more often reflect familiarity-based decisions. In this sense, the observed results would be incompatible with such an account. However, mean RTs in bin 1 were mostly in the neighbourhood of 1 s or more, which DPSD proponents suggest is ample time for recollection to occur (e.g., Yonelinas & Jacoby, 1994), so this does not necessarily rule out this explanation.

The relative absence of RT differences when item type was manipulated between subjects was an intriguing observation, especially in light of the earlier observation that the difference between stimulus types with respect to average response bias was also notably smaller in these than in within-subjects studies (Figure 2). The case of Experiment 8, in which RTs for both types of correct responses were remarkably similar for the words- and paintings-only groups in all bins (Figure 21), is probably the closest thing to evidence against a DPSD account of the previously discussed zROC findings thus far. zROC slopes for the two groups in this experiment exhibited the usual pattern, with results from the paintings-only group yielding a slope 0.11 lower than that for the words-only group (Figure 16). If this is attributable to the former group's greater reliance on recollection, one might expect this to manifest as correspondingly greater RTs. A within-subjects experiment exhibiting a similar pattern across the two analyses would be slightly more convincing, but the absence of the expected RT differences in a between-subjects context is far from sufficiently compelling to rule out the DPSD account considering the myriad potential factors that might contribute to individual differences in RT.

The discussion of these RT results has centred on implications for a potential DPSD account of memorial differences between paintings and words whereby recognition decisions associated with the former are more often made on the basis of recollection. There are, of course, a number of other potential explanations for RT differences. The observed patterns may have little or nothing to do with the recognition decision nor the associated underlying process or processes. Both the DPSD account and some of these alternative possibilities will be further discussed in Chapter 4, which describes an MBBE experiment wherein a 1-s response deadline was implemented at test.

Chapter 3: Perceived Confusability (Experiment 10)

With each subsequent experiment further cementing the consistency of the conservative bias for paintings and yet not revealing any apparent explanation, another idea was proposed. This hypothesis centred on the idea that if certain items are more conducive to study phase retrieval, also known as reminding (e.g., Hintzman, 2009), they may produce confusion when encountered on the test. Perhaps paintings, being arguably more visually striking than words, more often remind people of previously viewed paintings. If this feeling is experienced more frequently for paintings than words during the study phase, people may adjust their criterion accordingly when a painting is encountered on the test, uncertain – whether consciously or not – as to whether a painting which seems initially familiar was actually on the list or has merely called other paintings to mind.

An experiment was set up to test this idea by having participants press the spacebar each time an item on the study list reminded them of a previously presented item. The predictions were that on average, people would press the spacebar more often for paintings, and on an individual level, the magnitude of the difference between reported remindings on words and paintings would be correlated with the magnitude of the difference in conservative bias between words and paintings. Participants were again more conservative on paintings than words, and they did demonstrate a significant tendency to press the spacebar more often on paintings. However, contrary to expectations, these two tendencies were in no way correlated. Although paintings do seem more conducive to reminding than words, then, this tendency still does not adequately explain the response bias difference between the two item types.

One of the goals of the current experiment was to replicate the finding that people more often report a sense of reminding on paintings than words during the study phase. The main goal, however, was to explore a new potential explanation of the conservative response bias on paintings. All of the aforementioned experiments had participants make old/new judgments on a 6-point confidence scale, and calculations of hit and false alarm rates (and correspondingly c and d') were based on conceptualizing responses of 1, 2, and 3 as "old" and 4, 5, and 6 as "new". Such an approach, in addition to obscuring potentially interesting confidence results as discussed in the previous chapter, yields no information regarding the nature of the subjective experience of choosing one response over another. Going so far as to ask participants to rationalize all of their responses might disrupt the natural response process and cause its own problems, but the current experiment aimed to glean somewhat more qualitative responses while maintaining a relatively simple scale.

Specifically, the motivation for altering the response scale was the idea that people might tend to be more hesitant to endorse paintings as "old" because they are unsure as to whether a given painting was actually presented or is being confused with a similar painting. This could, of course, happen with words as well; after all, many of the nouns in the MBBE stimulus pool are semantically related. However, it seemed plausible that such confusion might be more likely with paintings, many of which depict similar scenes. A 4-point response scale was designed that included, in addition to standard "this was on the list" and "this wasn't on the list" options, two intermediate responses designed to capture this experience of having encountered a similar item and the confusion it sometimes elicits: "this, or an item very much like this, was on the list; not sure which" and "an item very much like this was on the list, but this item was not".

One group of participants completed a version of the new experiment including the previously discussed study-phase reminding component, while the rest completed a basic version where they did not respond at all during the study phase. The primary hypothesis was that on average, in both groups, people would select the intermediate responses – that is, those reflecting some memory for an item or items similar to the current one – more often on paintings than words. Secondly, in line with previous experiments, it was expected that participants would respond more conservatively to paintings than words despite the nonstandard response options. Finally, by including the reminding group, this experiment sought to replicate Experiment 7's findings: (1) people more often report a sense of reminding on paintings than words, and (2) the extent to which an individual reports more remindings on paintings does not predict the magnitude of the difference between paintings and words with respect to conservative bias (in other words, the MBBE).

Method

Participants.

Undergraduate psychology students at the University of Victoria completed the experiment for bonus course credit. Participants were drawn from a pool in which roughly 69% of individuals are female and the majority are between 18 and 25 years of age. Data were collected from 83 participants. Approximately half were quasi-randomly assigned to each version of the experiment (N = 39 for the basic version and N = 44 for the reminding version).

Materials.

The experiment was administered using E-Prime software (Schneider, Eschman, & Zuccolotto, 2002a & 2002b) on a Dell desktop computer running Windows XP. The words selected as stimuli were 192 4- to 8-letter medium- to high-frequency nouns taken from the MRC psycholinguistic database (http://www.psych.rl.ac.uk; Coltheart, 1981). The pictorial stimuli were 102 high-resolution digital scans of relatively obscure masterwork paintings by renowned artists that encompassed various styles and themes (e.g., abstract, portraits, landscapes, etc.).

This particular set of paintings has, in previous MBBE experiments, yielded sensitivity scores (d') more consistently similar to those obtained for words.

The study list comprised 96 items (48 words and 48 paintings) as well as three primacy and three recency buffers, and the test list included all studied items as well as 48 additional unstudied items of each type for a total of 192 items. Both study and test lists were randomly generated for each participant from the previously mentioned sets of words and paintings, with the obvious constraint that a given participant's test list always included all items from that participant's study list.

Procedure.

Words and paintings were presented one at a time in the centre of a white background, with words displayed in size 14 black font and paintings ranging in size from roughly 200×200 to 350×360 pixels. Study items were presented for 1 s each following display of a 250-ms fixation cross, with a 1-s interstimulus interval in the basic version of the experiment and a 2-s interval in the spontaneous reminding version.

In the basic version of the experiment, participants were instructed to attend to the items and try to remember them as well as possible for a later memory test. In addition to these instructions, participants in the spontaneous reminding condition were also asked to indicate each time a word or painting reminded them of a word or painting they had seen previously in the list by pressing the spacebar during the 2-s inter-stimulus interval. Instructions emphasized that this should only be done in cases in which the participant experienced a strong and spontaneous sense of being reminded.

Between the study and test phases, participants completed a distractor task, which entailed writing down as many countries as they could for 5 min. Test phase instructions were identical for both versions of the experiment. Participants were told they would again see a series of words and paintings – some of which would be items that had been on the study list and some of which would be new in the context of the experiment – and that they would be asked to select one of four response options for each item. The options were as follows: (1) this item was in the study list; (2) this or an item very much like this was in the study list, not sure which; (3) an item very much like this was in the study list, but this item was not; and (4) neither this nor an item very much like this was in the study list. Participants were encouraged not to purposely try to think of studied words and/or paintings that could be construed as being "like" the current item, but instead to select option (2) or (3) only when they spontaneously felt they had seen something a lot like the item in question. Upon selecting one of the four responses, participants were asked to judge how confident they were in their choice on a scale of 1 (not at all confident) to 3 (highly confident). All test responses were self-paced and the word or painting remained on the screen throughout both responses, with a one-second blank screen between test items.

Following completion of the test phase, participants were asked to report their academic major and were briefly interviewed regarding their thoughts and experiences, particularly with respect to making decisions on the test. Participants in the spontaneous reminding condition also provided estimates of how many times a word or painting in the study phase had reminded them of a previously presented word or painting, for a total of four subjective estimates.

Results

Four participants were excluded from analyses following an initial review of the data. One of these participants was highly knowledgeable about art and therefore familiar with all the paintings; one had completed a different experiment involving word memorization a few hours prior to the experimental session, and reported a lot of confusion on word trials as a result; one misunderstood the study phase instructions in the reminding experiment; and the last was excluded due to exhibiting roughly chance performance on words. Further analyses were therefore based on 41 participants in the reminding group and 38 participants in the basic group.

Sensitivity and response bias.

Sensitivity (d') and response bias (c) were calculated according to the usual formulas. Given that response 2 ("this or an item very much like this was in the study list, not sure which") was somewhat ambiguous and could be interpreted as an "old" or "new" judgment, separate hit and false alarm rates, and therefore separate d' and c values, were calculated to reflect both possibilities (i.e., in one case responses of 1 and 2 were counted as "old" responses and 3 and 4 as "new", and in the other only responses of "1" were counted as "old"). False alarm rates of 0 and hit rates of 1 were replaced according to Macmillan and Kaplan (1985); four values were replaced as such when 2 was counted as "old", and 12 were replaced when 2 was counted as "new". Paired sample t-tests showed that mean sensitivity was significantly higher for paintings than words in both the reminding and non-reminding groups, whether response 2 was conceptualized as an "old" or "new" response (Figure 22; $p < 0.001$ in all cases). Mean response bias was also significantly higher for paintings in all cases (Figure 22).

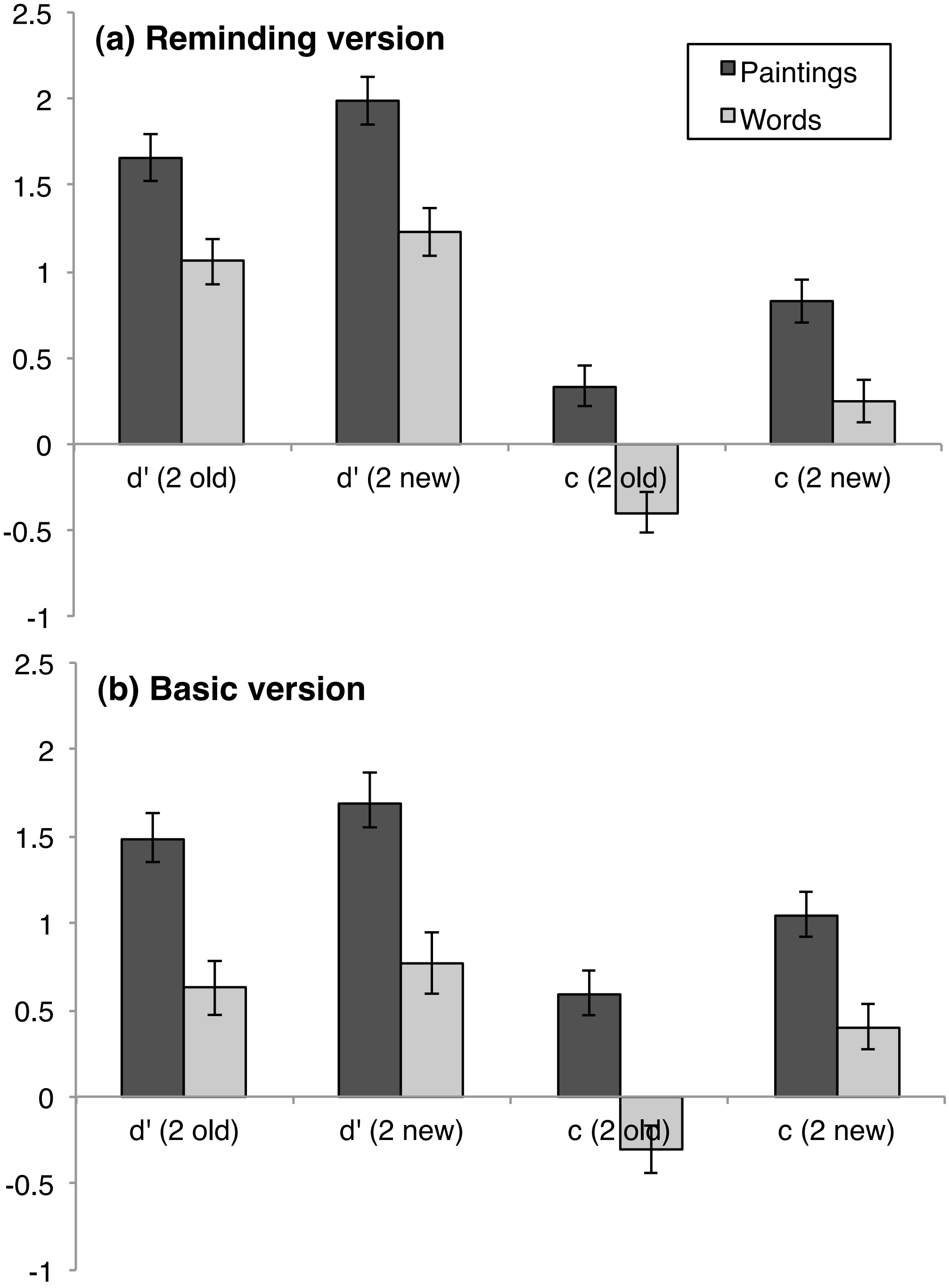

***Figure 22.* Mean sensitivity (d') and response bias (c) values for paintings and words in the reminding (N = 41; panel a) and non-reminding (N = 38; panel b) versions of Experiment 10.**

Values were calculated using hit and false alarm rates obtained by coding response 2 ("this or an item very much like this was in the study list, not sure which") as either an old (indicated as 2 old) or new (indicated as 2 new) response. Error bars indicate 95% within-subjects confidence intervals calculated according to Masson and Loftus (1994).

Quartile analyses.

Quartile analyses were conducted separately for the standard (a) and reminding (b) groups for hit rates and false alarm rates (Figure 23) and c and d’ (Figure 24). Responses of 2 were counted as “old” for the purpose of these analyses as this produced c and d’ values more comparable to those in previous studies and will, if anything, under- rather than overestimate conservatism on paintings. Refer to chapter 2 for more methodological detail about these analyses.

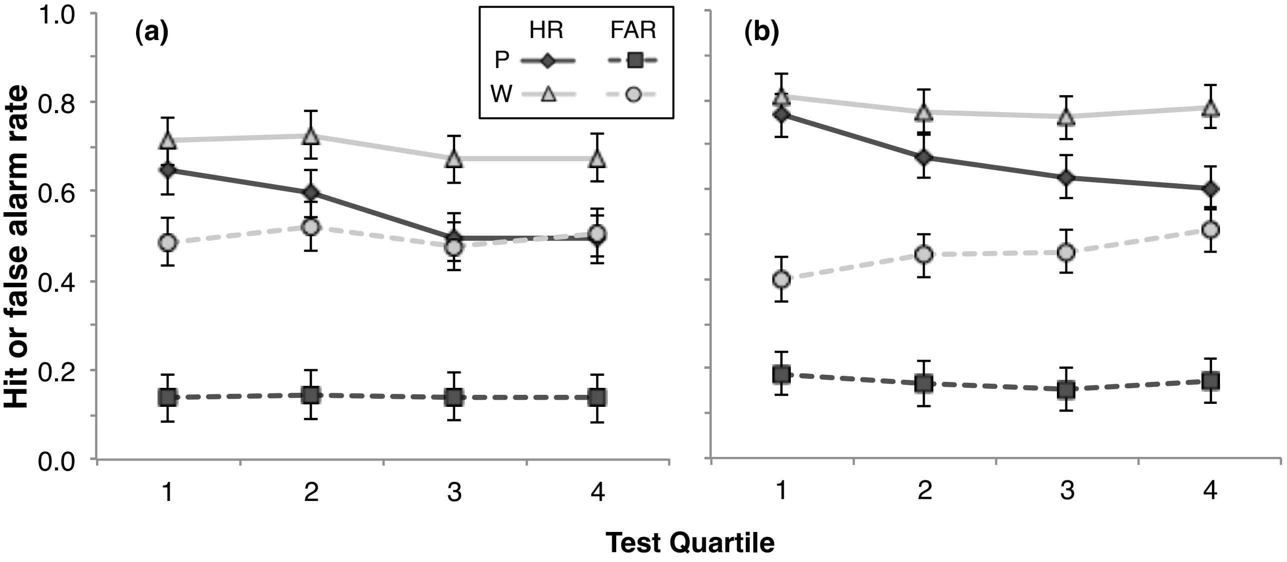

***Figure 23.* Mean hit (HR) and false alarm rates (FAR) for paintings (P) and words (W) in each 48-item test quartile in the standard (a) and reminding (b) conditions in Experiment 10.**

Error bars are 95% within-subjects confidence intervals (Loftus & Masson, 1994) calculated based on the results of 2 (item type) x 4 (quartile) repeated-measures ANOVAs.

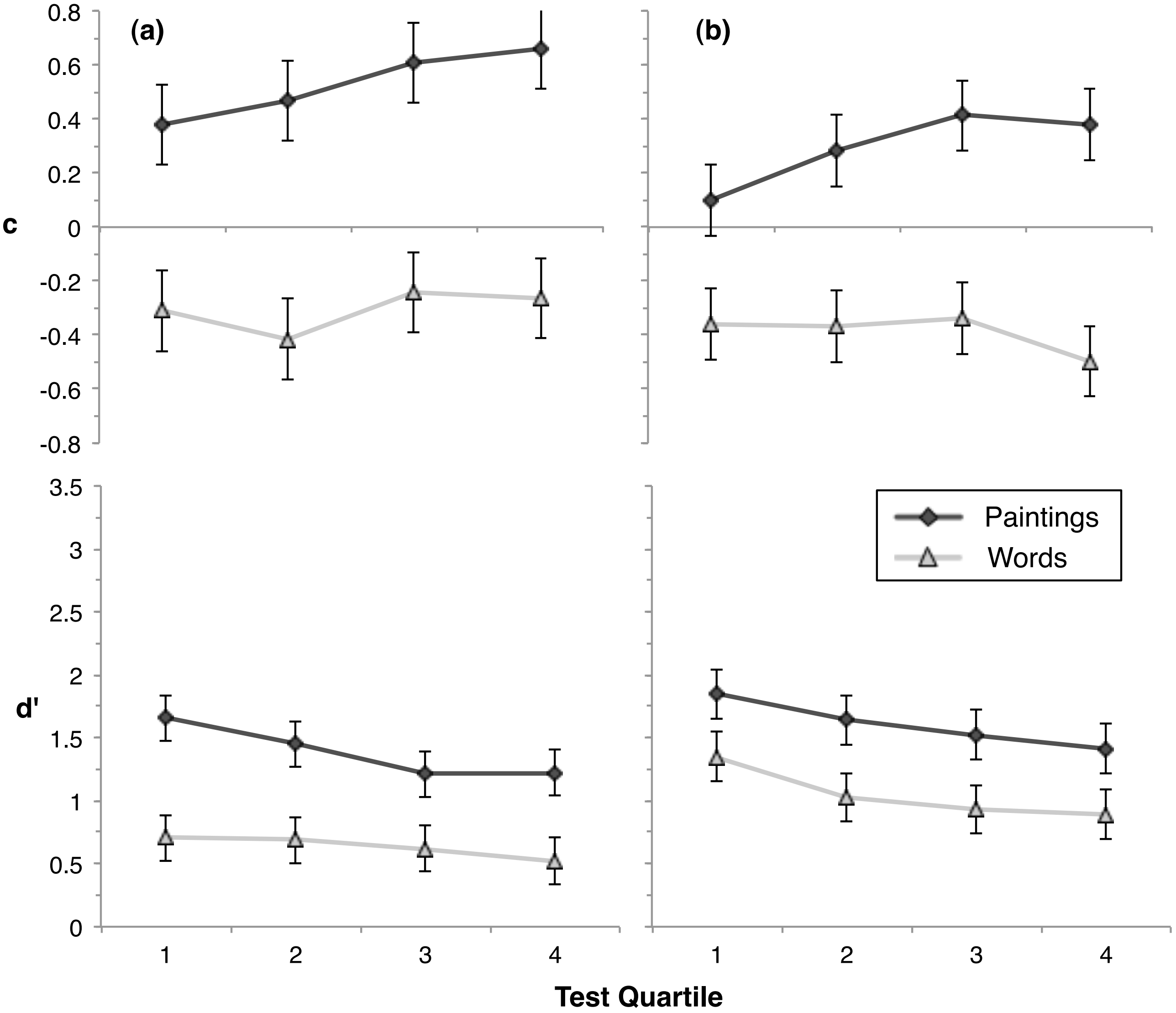

Figure 24. **Mean response bias (c; top row) and sensitivity (d'; bottom row) for paintings and words in each 48-item test quartile in the standard (a) and reminding (b) conditions of Experiment 10.**

Error bars are 95% within-subjects confidence intervals (Loftus & Masson, 1994) calculated based on the results of 2 (item type) x 4 (quartile) repeated-measures ANOVAs.

Frequency of responses.

Mean frequencies of responses 1-4 for paintings and words in both groups are presented as proportions in Figure 25. Figure 26 shows the overall mean frequency of each response for words and paintings for the basic (a) and reminding (b) groups, broken down by old (top row) and new (bottom) items.

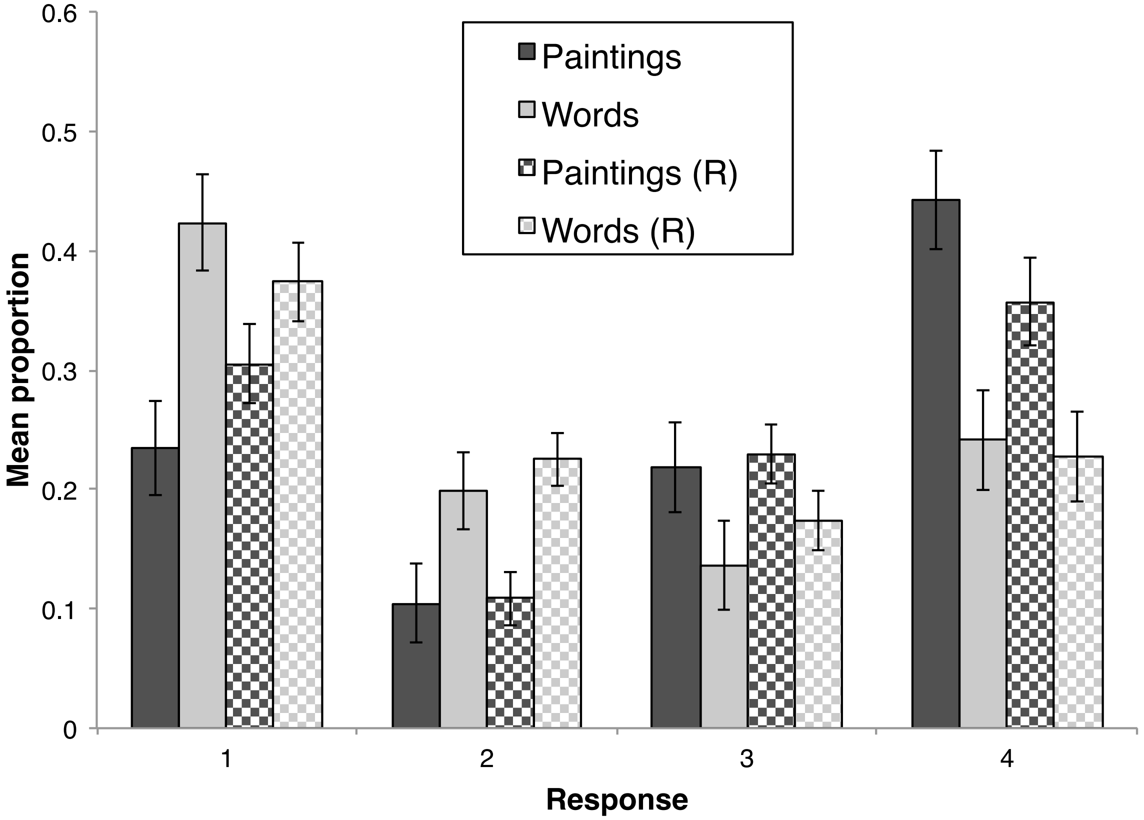

***Figure 25.* Mean proportion of test responses in each category for both paintings and words in both the standard and reminding (R) conditions of Experiment 10.**

Error bars are 95% within-subjects confidence intervals calculated according to Masson and Loftus (1994).

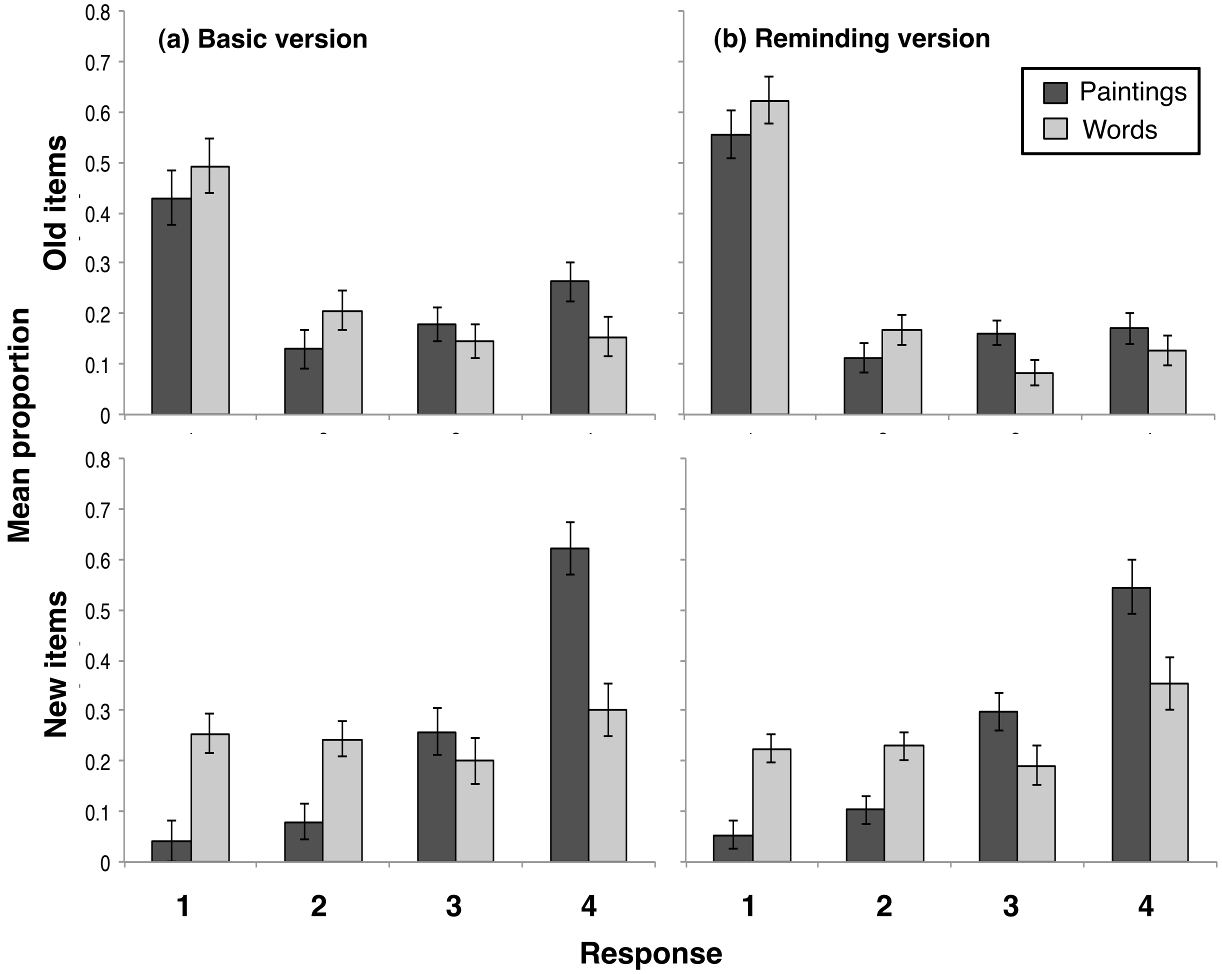

Figure 26. **Mean proportions of responses in each category (1-4) to old (top row) and new (bottom row) paintings and words in the basic (panel a) and reminding (panel b) conditions of Experiment 10.**

Error bars indicate 95% within-subjects confidence intervals calculated according to Masson and Loftus (1994).

Study phase retrieval.

Ten of the 41 participants who completed the reminding version of the experiment did not report any remindings for words or paintings. Of those remaining, 20 reported remindings more often on paintings, 10 reported remindings more often on words, and one reported remindings equally often for both stimulus types. A total of 467 remindings were reported across all participants, and approximately 61% of these were on paintings. On average, excluding the ten participants who did not hit the spacebar at all during the study phase, people reported 9.19 remindings on paintings (*SD* = 7.78) and 5.87 on words (*SD* = 5.64), and a paired sample t-test showed this difference was significant ($p < 0.01$). The results of a Pearson correlation between (1) the difference in number of remindings for paintings and words for each individual and (2) the difference in c between paintings and words for each individual were non-significant (r = -.025 for c values calculated counting 2 as “old”, r = -.011 if 2 is counted as “new”, $p > 0.05$ in both cases; Figure 27). Excluding the data for participants who did not report any remindings made little difference with respect to these correlations (r = -.051, r = -.045, $p > .05$ in both cases).

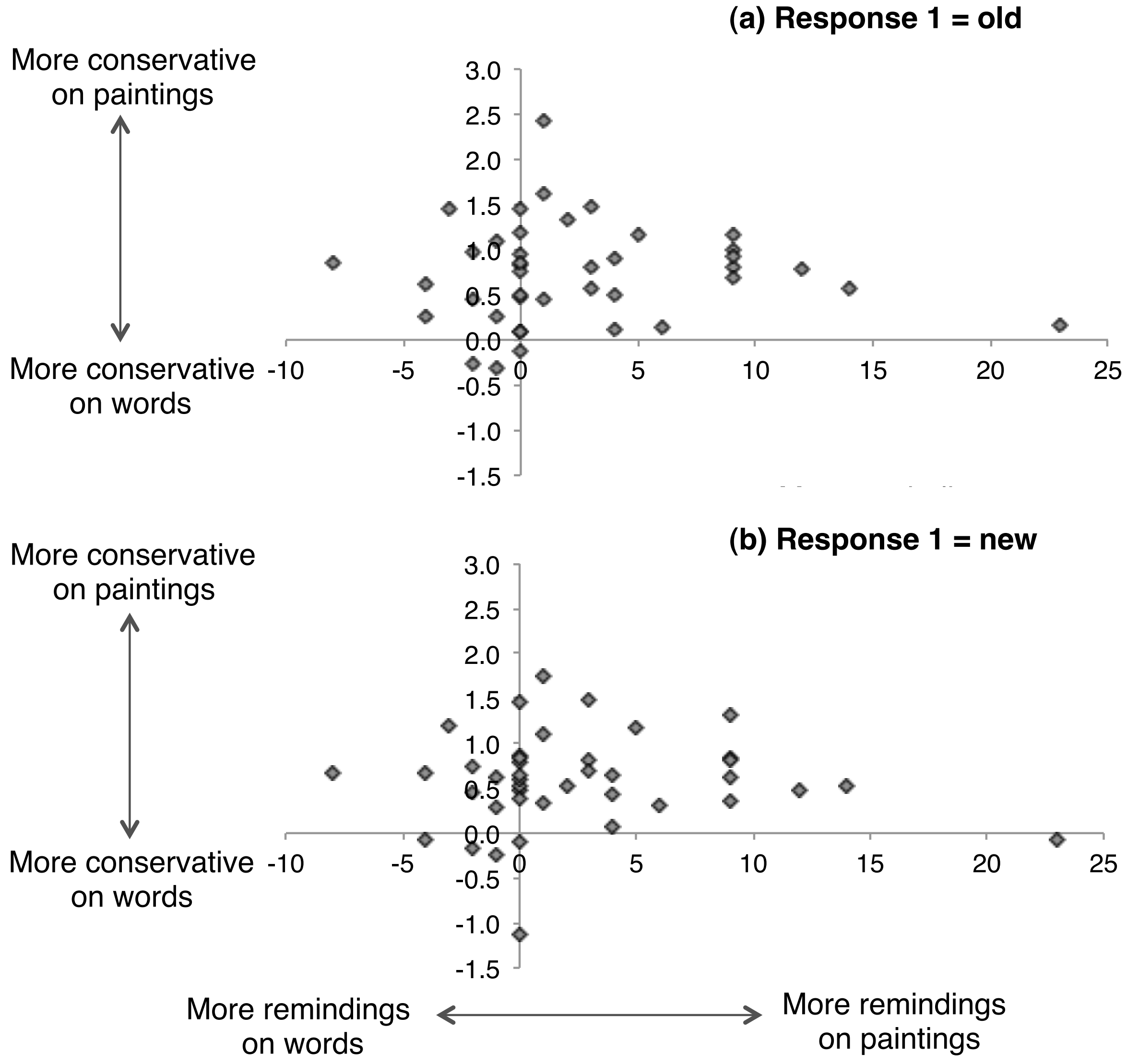

***Figure 27.* The difference in number of remindings reported on paintings and words by each participant (N = 41) plotted against the difference in each participant's conservative bias (c) for paintings and words in Experiment 10.**

Panel (a) shows this relationship when c values were calculated from hit/false alarm rates that counted responses of 2 as "old" responses (r = -.025; p > .05) and panel (b) shows the same relationship when responses of 2 were counted as "new" responses (r = -.011, p > .05).

Discussion

Tests of two previously hypothesized potential mechanisms for the MBBE – an inflated sense of the memorability of paintings, and higher rates of study phase retrieval on paintings – have yielded little insight into the causes underlying this effect. The study described in this paper tested a third hypothesis, namely that greater perceived similarity of paintings may produce confusion at test and thus a tendency to respond conservatively. This hypothesis was, on the whole, also not supported.

Sensitivity and response bias.

Response bias results were as expected, with the usual conservative bias for paintings (and the liberal bias for words that typically accompanies this in a within-subjects context) emerging regardless of where the old/new division was drawn on the response scale (Figure 22). The finding that d' was also significantly higher for paintings than words in both groups, on the other hand, was not necessarily expected with this stimulus set, although it has emerged in many previous studies (Figure 3). Although the picture superiority effect has been relatively well established, this has been mostly in the context of highly salient images, such as pictures of common objects (Shepard, 1967). The effect is not invariant and can be eradicated or even reversed by certain experimental manipulations, such as increasing the schematic similarity of the stimulus images (e.g., Nelson, Reed & Walling, 1978). The paintings used as stimuli were rich, complex, contained myriad features, and for the most part could not be easily captured by a brief phrase, in contrast with the types of images generally found to produce picture superiority. The reason for this tendency for d' to be higher for paintings, then, is not yet understood; however, as discussed more thoroughly in Chapter 2, it is not believed to be cause for concern in the context of the MBBE.

Quartile analyses.

The results of the quartile analyses will not be discussed extensively herein, but were included mainly for the purposes of comparison with the analyses described in the previous chapter. Directional trends were fairly comparable to those observed in previous experiments with respect to both c and d' despite the atypical response scale, although the false alarm pattern for words was somewhat anomalous (Figure 23). This is likely attributable to the abovementioned decision to base these analyses on hit and false alarm rates calculated when response 2 is considered an old response. The key question of interest here was whether the reminding group would show an across-quartile response bias pattern similar to that observed in experiments 4, 5, (Figure 7) and 7 (Figure 6) wherein c was initially conservative, thus further supporting the idea that the encoding task may influence how conservatism for paintings develops at test. As can be seen in Figure 24b, the 95% CI for the first quartile in the reminding group did overlap slightly with zero, but was still significantly conservative at the .05 level. However, as previously mentioned, quartile analysis calculations did take a "conservative" approach in that response 2 was counted as "old", producing lower estimates of c, so the unusual response scale may partly account for this discrepancy with the previous encoding task experiments. As a whole, the results for Experiment 10 were inconclusive in this regard.

Frequency of responses.

Figure 25 summarizes the mean frequencies of responses 1, 2, 3 and 4 for paintings and words in both experiments, and Figure 26 displays the same information divided according to whether the item in question was old or new. The frequencies of responses 1 and 4 followed the expected patterns with respect to both old and new items and words and paintings. When considered as a collective, the two intermediate responses were not chosen overwhelmingly more often for paintings than words, contrary to expectations. Response 2 was chosen significantly more often for words than paintings and response 3 showed the opposite pattern, but the more general expectation that the sum of these responses would be markedly higher for paintings was not supported.

With respect to response 2 being chosen more often for words than for paintings, it is possible that people tended to select option 2 as a “guess” response, as a number of participants mentioned guessing more often on words in the post-test interview. Because the wording of response 2 emphasizes an element of uncertainty, participants may have felt more comfortable choosing it than choosing one of the other responses with low confidence. Interview responses may yield some insight into the response 3 pattern as well; when participants were asked what criteria they used in deciding whether they had seen an item “like” the current item, many mentioned that for paintings, they tended to base such responses on recalling paintings with specific features – e.g., colours, certain types of clothing in portraits, number of people, furniture – similar to the current one, while for words such judgments tended to be based on semantic similarity.

Response 3 – “an item very much like this was on the study list, but this item was not” – is somewhat reminiscent of “recall-to-reject” processes that have been suggested by some authors (e.g., Clark, 1992) in that it implies memory for an item that is sufficiently vivid to identify it as being similar to the current item and yet recognize that it differs in some way. In line with the discussion of more standard recollection processes in the previous chapter, it is possible that paintings are more conducive to this kind of processing, perhaps because they offer a richer selection of visual features to attend to and later recall. This seems consistent with participants’ qualitative responses as well as with the finding of higher rates of correct rejections on paintings, but is of course merely speculative, especially since it is unclear how much insight participants have into their own decision-making processes in such contexts.

Future research could help shed light on the recall-to-reject idea as it applies to the MBBE. Rotello, Macmillan, and Van Tassel (2000) note that recall-to-reject processes can be investigated by comparing the ROC curves obtained by plotting (1) hit rates against false alarms to similar foils and (2) hit rates against false alarms to completely new, non-foil items. If such processes are operating, the upper x-intercept should be lower in the former case. Although the current experiment does not allow for such comparisons because similar foils were not incorporated in any sort of controlled manner, this may be an avenue worth exploring in the future.

Study phase retrieval.

The difference in number of remindings reported on paintings and words was not correlated with the difference in conservative bias (c) for paintings and words (Figure 27), regardless of whether response 2 was counted as “old” or “new”. This was consistent with the results of Experiment 7, which used the same remindings reporting procedure but employed a different response scale at test, confirming that differences between the two types of stimuli with respect to evoking study phase retrieval are insufficient to explain the conservative bias for paintings.

Although these expected differences in study phase retrieval may not explain the MBBE, they do seem to exist, as evidenced by the fact that remindings were reported significantly more often on paintings than words in both this experiment and the previous one. Importantly, however, there were substantial individual differences in this regard. While the mean number of total remindings reported by each participant was roughly 11, there was a standard deviation of

12, and values ranged from 0 reported remindings to a maximum of 45. The correlation between reported remindings on words and reported remindings on paintings was significant (*r* = .69 if participants who reported no remindings were included and *r* = .60 if they were excluded, *p* < .01 in either case), a finding which may not be particularly meaningful in the context of a combined paintings/words setup, but which could be interesting to explore using separate tests.

In summary, the current experiment further established the remarkable consistency of the MBBE. The finding that people do not seem to perceive more inter-item similarity and/or confusability for paintings than words marks the third hypothesis to have fallen short as far as explaining the origins of this effect, with the other two being inflated expectations of the memorability of paintings and a higher tendency for study phase retrieval on paintings. Despite this latter finding, the current study did replicate previous results in that participants reported more remindings on paintings than words, although this tendency was not predictive of differences in c. Future studies will continue to pursue the thus far elusive question of what causes people to respond conservatively to paintings.

Chapter 4: Response Deadline (Experiment 11)

As established in previous chapters, a series of recognition memory experiments conducted with paintings and words have revealed a strikingly consistent tendency for participants to respond conservatively to paintings, while words tend to produce liberal responding in experiments including paintings and roughly neutral responding when they are the only stimuli (e.g., Figure 2). Three hypotheses regarding the mechanisms underlying this bias have been tested thus far – one related to subjective memorability (see Chapters 1 & 2), a second inspired by Hintzman's (2010) ideas about reminding (Chapters 1-3), and most recently, the idea that the paintings used as stimuli may be perceived as highly similar and/or confusable, promoting a conservative approach at test (Chapter 3). None of these hypotheses were able to adequately account for the materials-based bias effect (MBBE), but a series of cross-experimental analyses (described in Chapter 2) revealed some interesting patterns and inter-stimulus differences aside from the response bias pattern itself that suggested a few potentially promising directions for future research.

One such pattern was in the response time (RT) data. Casual examination of data from a few experiments revealed a tendency for mean RT to be directionally higher for paintings than words, and more meticulous analyses confirmed that this was often the case, at least for correct responses, in experiments wherein item type was manipulated within-participants (Figures 17-20). There was some variability in this regard; for example, several experiments did not show this trend for correct rejections, such as those shown in Figure 18. Furthermore, in experiments wherein item type was manipulated between subjects such that a given participant studied and was tested on paintings or words only, this trend was not the norm, showing up only in certain bins for hits in Experiment 9 (Figure 21).

Although the RT trend was not without exception, and was less consistent for correct rejections than for hits, that it was evident in the vast majority of within-subjects experiments was compelling. The relative absence of such a pattern when item type was manipulated between subjects only made it all the more interesting, especially given the earlier observation that the difference between stimulus types with respect to average response bias – in other words, the MBBE itself – was also markedly smaller in these experiments than in the within-subjects studies (Figure 2). In Chapter 2, these RT results were discussed in the context of a potential DPSD account wherein recognition judgments to paintings are more often recollection based than those to words, an idea which had been touched upon in the previous section about receiver operating characteristics (ROCs). The overall picture with respect to the plausibility of this account can, thus far, be summed up as "inconclusive".

Better understanding the time course of recognition decisions to words and paintings might help elucidate the plausibility of the above account. Recollection is typically conceptualized as a relatively slow, effortful process in contrast with the more rapid, automatic assessment of familiarity (Atkinson & Juola, 1974; Jacoby, 1991; Mandler, 1980). The question of the relative speeds of recollection and familiarity is not entirely settled, but the prevailing view seems to be that – even though the two may unfold in parallel – recollection is the slower process on average, a contention that has been supported by numerous studies of associative-, source-, and pair-recognition (e.g., Dosher, 1984; Hintzman & Curran, 1994; Rotello & Heit, 2000) studies. More recently, using a variety of stimuli, Besson, Ceccaldi, Didic, and Barbeau (2012) reported a seemingly incompressible lower limit on visual recognition memory in general at 370 ms. Further, using a modified remember/know procedure, the authors found evidence that any responses under 420 ms were solely familiarity-based (Besson et al., 2012). Although some

DPSD theorists might not agree on the particular cut-off, it seems to be generally accepted that the most rapid responses tend to be familiarity based, and that speeded responding limits recollection (Jacoby, Jones & Dolan, 1998). This idea is one of several that inspired the response deadline experiment described in the current chapter.

Beyond the goal of further evaluating the DPSD account, implementing a response deadline may also help clarify the nature of the MBBE itself. As Rotello and Macmillan (2007) pointed out, SDT itself makes no claims about the intentional versus unconscious nature of criterion setting. Arguably, given the myriad variables described in Chapter 1 that have been linked to effects on response bias, it seems unlikely to be an either/or situation. Previous hypotheses regarding the cause of the MBBE have not taken a stance on this issue. Although measurement of the previously proposed mechanisms (subjective memorability, reminding, and confusability) relied on subjective judgments and therefore some degree of conscious insight into these processes, no claims were made as to whether these proposed mechanisms had to be available to conscious awareness to influence the criterion, nor whether criterion setting and/or shifting (in the context of the MBBE or otherwise) is automatic or intentional. Further investigation of this latter point might help in narrowing down the variables likely to be involved in this effect.

Event-related potential (ERP) studies of recognition memory have shed some light on the time course of the intentional and automatic processes that support recognition memory decisions. One common observation in such studies is a general “old/new effect” whereby studied items elicit more positive activation at test than new items. This general effect can be divided into multiple subcomponents. For current purposes, three subcomponents are relevant: an early component (occurring 300-500 ms following stimulus presentation) that has been linked to automatic and incidental memory processes, a later component (occurring roughly 500 ms post-stimulus) that has been linked to processes of controlled retrieval, and an additional late sustained component (occurring between 500 and 1500 ms post-stimulus) that has been associated with retrieval processes (e.g., Allan, Wilding, & Rugg, 1998). Windmann, Urbach, and Kutas (2002) investigated the time course of response bias with this timeline in mind. The authors were interested in the automatic versus strategic nature of response bias, and noted that some factors known to affect it, particularly item characteristics such as associative interrelatedness (Miller & Wolford, 1999), cast serious doubt on the conceptualization of bias as under strategic control. (Windmann, Urbach, & Kutas, 2002).

The methodology employed by Windmann and colleagues will not be discussed in detail, as the current experiment was not ERP-based. Key findings, however, were the observation of ERPs sensitive to bias at frontal recording sites that were particularly notable between 300 and 500 ms following stimulus presentation; the fact that these effects were largest for what the authors called the subjective comparison, which compared ERPs associated with items participants called “old” to those associated with items they called “new”; and the fact that within the subjective comparison, the ERPs for a group exhibiting low response bias showed a clear old/new effect at frontal sites that was essentially absent for a high bias group (Windmann et al., 2002). The authors suggested that the 300-500 ms effects might represent criterion setting by the prefrontal cortex, which would be generally consistent with anatomical findings (Miller, Handy, Cutler, Inati, & Wolford, 2001). Importantly, however, the fact that the apparent effect of bias was maximal so soon after stimulus presentation suggests that participants in this experiment set criteria relatively automatically, raising problems for theories that propose a more strategic, intentional role for the decision criterion (Windmann et al., 2002).

Although even a response deadline at the high end of this 300-500 ms window proved infeasible in pilot tests of the current MBBE experiment, it seemed plausible that forcing more rapid responses might still shed some light on the level of intentional processing required for the MBBE to unfold. Previous experiments have allowed participants unlimited time to respond to each test items, and as the previously discussed RT results revealed, they tended to take more time responding to paintings than words. With this in mind, it seemed possible that time, and the greater degree of intentional processing it allows for, might be critical to the emergence of the observed effects. Forcing participants to respond quickly should force responding on a more automatic, “gist” basis (possibly, but not necessarily, familiarity), thus limiting the availability and usefulness of more intentional processes. This idea was tested by incorporating a 1-s response deadline in a standard MBBE experiment. The primary hypothesis was that under such a deadline, the usual pattern of conservative bias for paintings and liberal bias for words would be eliminated or greatly reduced, and that c for both item types would be approximately neutral. It was also anticipated that sensitivity (d’) would be lower than in prior experiments, although this was not of central interest.

Method

Participants

Participants were undergraduate psychology students at the University of Victoria from the same pool as Experiment 10 (see Chapter 3). Of the 47 total participants, nine only took part in an initial pilot phase of the study designed to get feedback about the response deadline and the means of entering test responses; data from these participants were only analyzed to the point of determining the number of missed responses, and are not included in the analyses below. 38 students participated in the final version of the experiment.

Materials and procedure.

Materials and the overall structure of the current experiment were essentially identical to those described in Experiment 10, with the only differences being related to the test phase. In the current experiment, items randomly selected for the first 12 positions in the test list were assigned to a practice phase to allow participants to get used to the deadline and procedure. At the beginning of the test phase, participants were told that their task would be to decide whether each item had been in the study list or not by pressing the “z” key for studied items and the “/” key for novel items. The experimenter then informed participants of the response deadline, explaining that there would be a very brief 1-s response window for each item and that this deadline would be preceded by a series of three beeps, the third of which signalled 1 s. Participants were encouraged to try their best to respond to items by the third beep as often as possible, asked to keep an index finger near each of the response keys throughout the experiment to facilitate rapid responding, and told that a brief message indicating a missed response would be displayed each time they failed to respond quickly enough.

The experimenter told participants there would be a short series of practice items to allow them to get used to the deadline and the mapping of key press responses. Once the participant understood the instructions, they were told to press the spacebar to initiate the practice phase whenever they were ready. The experiment paused briefly following the 12 practice items and a message was displayed indicating the end of the practice phase. When this message appeared, the experimenter double-checked that the participant understood the procedure and ensured that they were comfortable with the volume of the response deadline signal. The participant then pressed the spacebar again to proceed to the remainder of the test.

At the end of the test phase, a few questions appeared related to participants' perceptions of the experiment and their own performance. Participants were then asked to report their academic major and debriefed regarding the purpose of the experiment.

Results

Any participants with d' values lower than 0.1 for words and/or paintings were excluded from analysis. Although a d' value of 0.1 represents a low level of discrimination relative to that typically observed in this line of research, it was expected that the response deadline would make the task markedly more difficult; accordingly, a fairly lenient cut-off was selected. Nine participants were excluded on this basis, and data for one additional participant were lost due to a technical error. Two participants were outliers with respect to the number of responses missed and were also excluded. The analyses below are based on the remaining 27 participants.

Sensitivity and response bias.

Sensitivity (d') and response bias (c) results are shown in Figure 28. Mean c for paintings was 0.23 (*SD* = 0.41), which was significantly higher than c for words (*M* = -.20, *SD* = .39) according to a paired samples t-test (*t*(26) = 3.84, *p* = .001). This corresponded to both significantly lower hit (*M* = .57, *SD* = .15) and false alarm rates (*M* = .28, *SD* = .15) for paintings than words (HR: *M* = .68, *SD* = .15; FAR: *M* = .46, *SD* = .15), *t*(26) = 2.73 and 4.54, *p*s = .011 and <.001, respectively. Although d' was directionally higher for paintings (*M* = 0.83, *SD* = 0.40) than words (*M* = 0.67, *SD* = 0.52), this difference was not significant (*t(*26) = 1.75, *p* = .09).

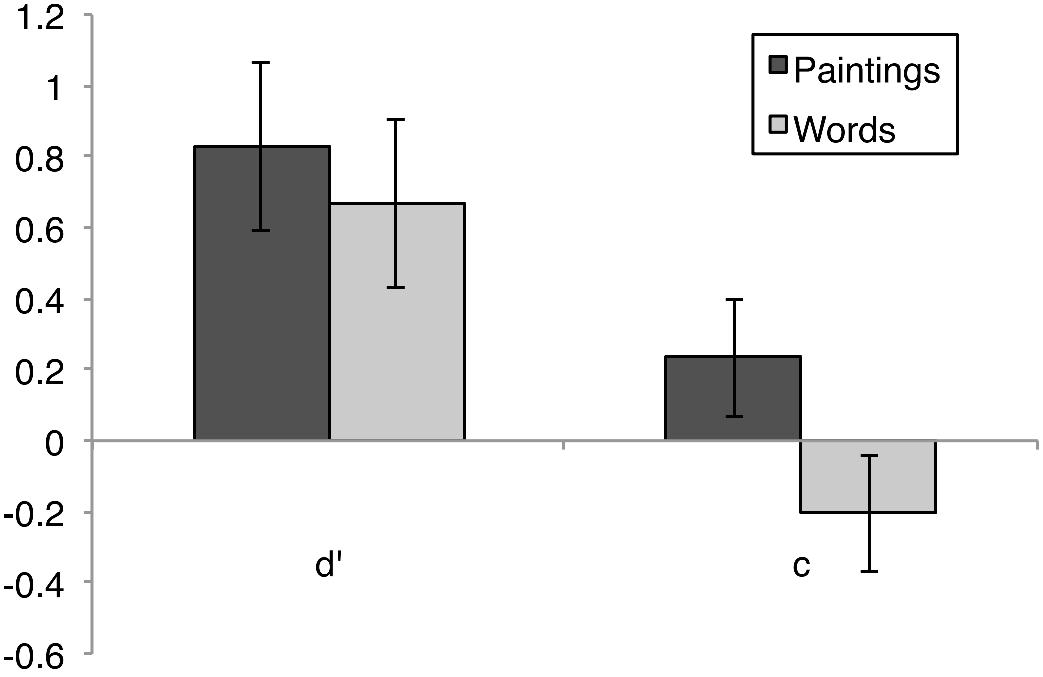

***Figure 28.* Mean sensitivity (d') and response bias (c) for paintings and words in Experiment 11.**

Error bars are 95% within-subjects confidence intervals calculated according to Masson and Loftus (1994).

Response time.

Response time data for words and paintings are shown in Figure 29 according to response type. Error bars are 95% within-subjects confidence intervals (Masson & Loftus, 1994).

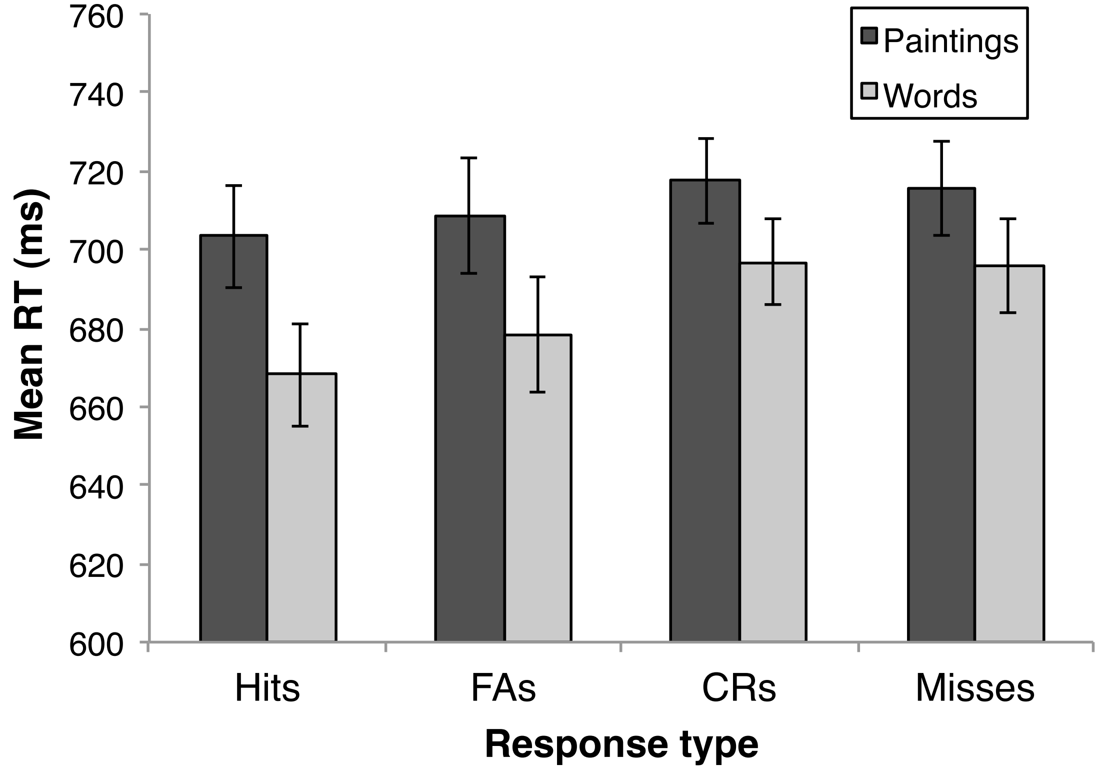

Figure 29. **Mean response times (ms) for all four types of responses to paintings and words in Experiment 11.**

Error bars are 95% within-subjects confidence intervals calculated according to Masson and Loftus (1994).

Quartile analysis.

Quartile analyses (see Chapter 2 for methodological details; the only difference in the current experiment is that quartiles only comprised 45 items each due to the inclusion of a practice phase) are shown in Figure 30 for hits and false alarms (panel a), response bias (c; panel b), and sensitivity (d'; panel c). Error bars are, again, 95% within-subjects confidence intervals.

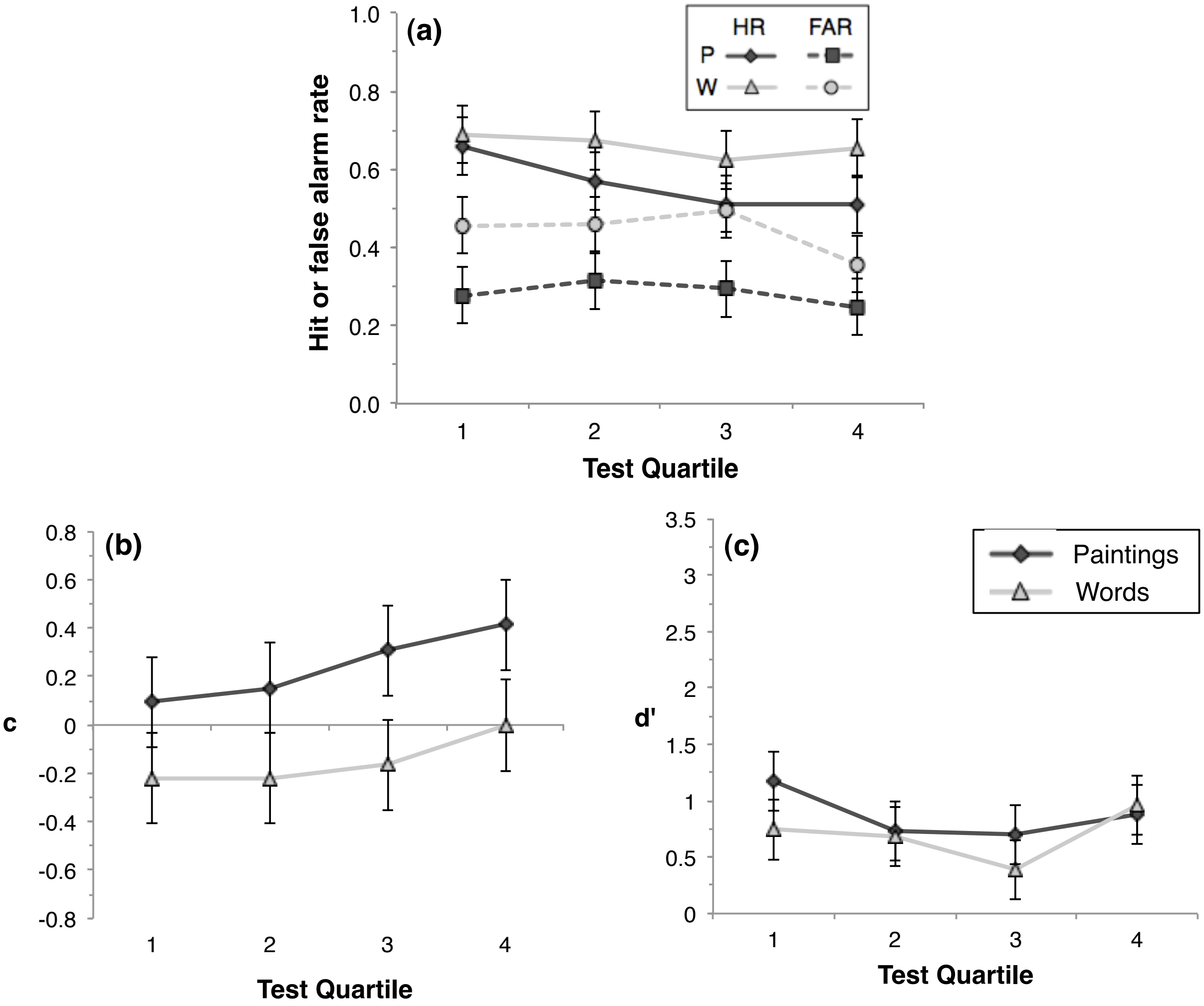

***Figure 30.* Mean hit and false alarm rates (panel a), response bias (c; panel b) and sensitivity (d'; panel c) for paintings and words in each 45-item test quartile in Experiment 11.**

Error bars are 95% within-subjects confidence intervals (Loftus & Masson, 1994) calculated based on the results of a 2 (item type) x 4 (quartile) repeated-measures ANOVA.

Discussion

The main hypothesis of the current experiment – namely, that implementing a response deadline would eliminate or noticeably reduce the typically observed response bias differences between words and paintings – was not supported. Participants still exhibited a significantly conservative response bias for paintings and a significantly liberal bias for words (Figure 28), and the absolute values of each, as well as the difference between them, were of comparable magnitude to those observed in a series of previous studies (see Figure 2, which includes results for this experiment at the bottom). Sensitivity was, as expected, lower than what is typically seen in these experiments (Figure 3), but did not differ significantly between words and paintings. This may, however, be partly a floor effect, as d' was low overall. Regardless, the remainder of the discussion will not focus on these findings.

The previously discussed tendency for c for paintings to increase across test quartiles was also observed in the current experiment, although in this case, it was also observed for words (Figure 30). High variability due to the relatively small sample size in this experiment rendered the current quartile results somewhat difficult to interpret statistically based on observation of the CIs alone. To aid interpretation, paired samples t-tests comparing c in quartile 1 to that in quartile 4 were conducted for paintings and words; the former difference was significant ($t(26) = 3.47$, $p = .002$) while the latter did not achieve significance at the .05 level ($t(26) = 1.91$, p = .067). Thus, although c for paintings did not differ significantly between all adjacent quartiles, there was still a significant increase in conservativeness from the first to the fourth quartile. Response bias results, then, were fairly analogous to those obtained in previous studies, both in terms of the overall picture and the quartile results.

An additional unexpected finding pertains to response time. Despite the limited window within which participants were able to respond, a significant difference in RTs still emerged for all response types whereby responses to paintings were, on average, slower than those to words (Figure 29). In numerical terms, these differences ranged from 20 ms (misses) to 35 ms (hits), but the consistency of this effect was fairly remarkable considering the low sample size. With respect to the DPSD account touched upon earlier in the current chapter as well as in Chapter 2, although still far from conclusive, these RT differences are indeed compatible with the idea that judgments to paintings may more often be made on the basis of recollection, while responses to words are more often familiarity based. Had these differences been observed under a more stringent response deadline, this would cast doubt on such an account, but the observed mean RTs are in a range in which most DPSD theorists suggest recollection-based decisions are possible. Precise estimates vary, but to give one example, Besson and colleagues (2012) suggested, by combining ERP findings and the 110 ms required for a decision and initiation of a motor response to occur (Kalaska & Crammond, 1992; VanRullen & Thorpe, 2001), that familiarity based responses can occur as early as 410 ms, and recollection as early as 610 ms.

While these data cannot rule out the DPSD explanation of RT and zROC slope differences (see Chapter 2), they do not necessarily constitute strong evidence in its favour. Although current data do not allow for anything but speculation, the true reason for RT differences between the two item types – both in this experiment and others – may be somewhat more mundane; namely, it seems plausible that these differences merely reflect differences in visual processing time. Painting stimuli both occupy more space on the display and are more visually complex in a number of dimensions (colour, number of features, etc.) than words. Early studies of visual processing speed, for example, found that response latency tended to increase with increasing number of elements in a visual display (Atkinson, Holmgren, & Juola, 1969), and such an explanation of the observed RT differences between words and paintings is arguably more parsimonious than attributing it to reliance on different underlying memory processes. The relative consistency of RT differences across different response types in the current experiment also seems to favour such an explanation over one proposing recollection as the mechanism underlying increased RTs for paintings, as one would not expect recollection to support incorrect responses to the same extent as correct responses. Future experiments could potentially investigate this issue further by attempting to make words and paintings more comparable in terms of visual complexity and size; for example, words could perhaps be presented in larger and more elaborate fonts or against a richly patterned background. If RT differences were to persist in such an experiment, this might speak against a mere perceptual complexity explanation.

Although the current experiment did not yield the expected results, this in itself says something about the MBBE. A 1-s response deadline is far from sufficient to conclude that the processes underlying the effect are necessarily automatic, but nonetheless, the mechanism seems to operate fairly quickly following stimulus presentation, suggesting extensive intentional processes are not required. It certainly could be automatic, perhaps resulting from some perceptual characteristic of paintings along the lines of attributes like symmetry, size, and contour length, all of which Brodeur and colleagues (2011) found to influence bias. However, the role of later, more intentional processing – for example, retrieval or meta-memory processes – in establishing the conservative criterion for paintings in the context of the MBBE still cannot be ruled out. Similarly, results from the current experiment are not sufficient to rule out a DPSD account of some of the observed differences between words and paintings (or, for that matter, the UVSD account also described in Chapter 2). The final chapter will summarize the findings thus far, review some proposed accounts, and suggest a few possibilities for further research.

Chapter 5: General Discussion and Concluding Remarks

The previous four chapters have summarized the bulk of the current state of knowledge regarding the materials-based bias effect or MBBE. To reiterate, this effect refers primarily to the remarkably consistent inclination toward conservative responding to paintings in recognition memory tests, whether paintings are the only stimuli (as in the experiments by Lindsay & Kantner, 2011, & Experiments 8 & 9 in the current paper) or are randomly intermixed with words at study and test (Experiments 1-7; one group in Experiment 8; & Experiments 10 & 11). The corresponding response bias patterns for words, namely liberal responding in a within-subjects context and neutral responding when they are the only stimuli, is considered a secondary component of the effect. The MBBE has proven robust to numerous changes at encoding and test – including orienting tasks, atypical test scales, and a response deadline – and the same pattern has emerged in cases wherein sensitivity was higher for paintings than words, lower for paintings than words, or roughly equivalent for the two stimulus types (e.g., Figures 2 & 3). Despite this impressive consistency, the underlying mechanism is still poorly understood. Three hypothesized mechanisms have been tested thus far, one of which was described in detail in Chapter 3, and although the associated experiments have yielded important information, none of these mechanisms has been able to account for the MBBE.

The analyses described in Chapter 2 revealed a few additional differences between paintings and words that may be important to understanding the overall bias differences between them. In brief, response bias typically becomes more conservative over the course of the test for paintings, while no such pattern has been observed for words; zROC slopes tend to be shallower for paintings than words; and response times, at least those associated with correct judgments, are often longer for paintings. Implementing a response deadline failed to eliminate both the first and last of these effects (confidence judgments were not obtained in Experiment 11, and would be difficult to obtain in such a case). These results are all seemingly compatible with either a UVSD or DPSD explanation of the MBBE and other differences between paintings and words. Neither can be conclusively ruled out at this point, but both will be discussed critically below.

DPSD Interpretation of the MBBE

In the context of the current paper, DPSD was first discussed in the context of the zROC results in Chapter 2. As mentioned previously, DPSD models explain zROC slopes below 1 as reflecting the involvement of a second process, namely recollection; accordingly, under DPSD assumptions, the tendency for these slopes to be lower for paintings than words would be attributed to increased involvement of or reliance on recollection in making decisions about these items. RT results presented herein would also be consistent with this idea, given that recollection is thought of as the slower process.

Of course, the question of ultimate interest is that of how DPSD would explain the MBBE itself. There may be a number of possibilities, and the most plausible might depend to some extent on the specific DPSD instantiation to which one subscribes. The classic model assumes recollection to be an all-or-nothing threshold process, while familiarity is continuous and operates in line with SDT (Yonelinas, 1999). However, it has since been proposed that recollection is also a continuous process that can be conceptualized in terms of SDT, and evidence supporting this alternative is mounting (Slotnick, 2010; Wixted, 2007). However, although the zROC explanation emphasizes the possibility of more recollection based responding to paintings, this does not mean recollection itself is the crucial process underlying the MBBE.

Under either of the above DPSD models, conservative responding to paintings could conceivably be attributable to a high familiarity criterion alone. Assuming this is the case for the

MBBE would in effect just force things back a step, necessitating an explanation as to why this is the case. If the critical difference between stimuli in terms of producing the MBBE lies at the familiarity stage, this would seem to favour an explanation based on perceptual differences, given that familiarity is conceptualized as a fast, automatic process. Perhaps the relative complexity and feature density of paintings play a role; for example, because these stimuli boast such a vast array of features, many of which might spuriously match contents of memory, the underlying system might demand a high number of matches to support a familiarity-based decision. Alternatively, the familiarity criterion might not differ between words and paintings, but the low baseline familiarity of the latter has the effect of this criterion less often being reached. This could produce both fewer false alarms on the basis of spurious familiarity and fewer true hits. None of the existing data seem incompatible with these explanations; that said, however, the critical aspects also do not necessitate a dual process assumption. At this juncture, there is still no compelling reason to favour this model over other possibilities.

A few MBBE experiments not reported in detail herein have attempted to evaluate the DPSD account somewhat more directly via the remember/know paradigm, which asks participants to report for each “old” response whether they made the decision because they “remember” specific details about the item (intended to reflect recollection) or simply “know” it was studied without such accompanying detail (thus corresponding to familiarity; Tulving, 1985; Yonelinas, 2002). Preliminary data suggest participants do more often select the “remember” option (and variations thereof for “new” responses) for paintings than words, which unquestionably says something interesting about the subjective experience of making decisions about these two item types. However, the interpretation of remember/know results in terms of qualitatively distinct underlying processes has been oft criticized (e.g., Donaldson, Mackenzie, &

Underhill, 1996; Rotello & Zeng, 2008). Results of source memory experiments in particular seem incompatible with the idea of the 1:1 correspondence between these subjective judgments and the proposed underlying systems. If “know” decisions are familiarity based, they should be relatively devoid of source detail; in contrast, findings like above-chance source accuracy for “know” judgments (Conway & Dewhurst, 1995), roughly equal proportions of “remember” and “know” responses in studies reporting extremely high source accuracy (Hicks, Marsh, & Ritschel, 2002), and above-chance recollection of encoding details like orientation and colour for “know” responses (Eldridge, Engel, Zeineh, Bookheimer, & Knowlton, 2005) dot the literature. Additionally, according to Wixted (2007), studies directly comparing DPSD and UVSD models consistently favour the latter. As a whole, then, although there is no basis for flat-out rejection of a DPSD interpretation of the MBBE, there may be reason for scepticism, and if positing two processes is not necessary to understanding the effect, perhaps it is better to consider simpler explanations.

UVSD & Noise-based Interpretations of the MBBE

UVSD models were also previously discussed with respect to zROC slope differences. Such models explain slopes below 1 in terms of higher variability in the target distribution relative to the lure distribution, and therefore the lower slopes for paintings than words under this assumption could reflect an even more variable distribution of targets and/or a less variable lure distribution in the former case. The idea that high variability at encoding might produce a highly variable target distribution for paintings was also discussed, the idea being that lack of previous experience with most of the paintings and/or a pre-existing encoding strategy for these items might lead the corresponding memory representations to vary widely with respect to strength. This in itself, of course, would not explain the MBBE. However, the familiarity-based mechanisms proposed in the DPSD account could just as easily apply here. Additionally, some of the ideas put forth in the Chapter 2 section on quartile analyses seem potentially compatible with such an account. For example, it was mentioned that some researchers have proposed that words are represented in a fairly unitary fashion in memory, perhaps as a result of extensive experience, while non-word stimuli tend to have more distributed, overlapping representations (Kinnell & Dennis, 2012; Osth et al., 2014). If this is true for paintings, it seems like it would fit with the UVSD idea of target distribution being more variable in this case than for words. These authors also suggested that such representations leave non-word stimuli more vulnerable to noise from other items in the experiment (Kinnell & Dennis, 2012; Osth et al., 2014), and conceivably this could also extend to the contexts in which items are presented, thus hearkening back to the quartile analysis discussion.

One potential issue with the above possibility is that if paintings are more susceptible to item noise than words due to overlapping representations, one might expect a correspondingly lower d’ for paintings. That this was not consistently the case (Figure 3), however, is not necessarily catastrophic for this explanation, as there is another types of noise to which the type of words used in MBBE experiments are likely to be highly vulnerable: context noise, which comes from previous experience with a stimulus in various contexts. Reder, Angstadt, Cary, Erickson, and Ayers’s (2002) explanation for why rare words do not produce the word frequency mirror effect emphasizes the nature of this kind of balancing act between item and context noise. The word frequency mirror effect refers to the common observation of increased hits and decreased false alarms for low frequency relative to high frequency words, but words of extremely low frequency do not show this effect. Reder and colleagues (2002) suggested that low frequency words outperform high frequency words because the latter are highly susceptible to context noise as a result of previous experience and associations, while for extremely rare words this advantage is counteracted by the overlapping nature of their representations and the increased susceptibility to item noise this causes.

The encoding variability idea and the concept of paintings having more overlapping representations seem like potentially related or at least not discordant ideas. If paintings are inconsistently encoded, this might lead to more variable and overlapping representations, leaving them more vulnerable to interference from other paintings in the experimental context. As proposed in Chapter 2, the inclusion of an additional task at study might make study representations more clearly distinguishable from test representations. Study tasks could also facilitate more consistent encoding and perhaps relatively less overlap among representations as a result. At the beginning of the test list that follows, then, vulnerability to item noise might be fairly low for paintings, because representations thus far overlap relatively little. However, as the test proceeds and more paintings are viewed – and probably encoded inconsistently, given study is not intentional at this stage – this vulnerability would increase. In the case of experiments lacking a study task, this same increase in noise susceptibility would occur, but it would also exist to a greater extent at the beginning of the test.

With respect to conservative responding, one could speculate that setting a strict criterion represents an attempt by the memory system to address the problem of representation overlap potentially producing a spuriously high match between a novel probe and some pre-existing representation. Evidence comes from research into list length and strength effects may speak to this somewhat, or at least suggest a potential future direction. Types of items associated with representations more susceptible to item noise should be vulnerable to both of these effects – decreasing performance with increasing list length or the inclusion of items of varying strengths

– and although failure to observe the latter effect (e.g., Hirshman, 1995) in many studies cast doubt on the initial wave of global matching models, more recent studies have produced it with non-word stimuli (Kinnell & Dennis, 2012; Osth et al., 2014). The effect itself is not so important for current purposes, but both these more recent studies and past investigations have also reported substantial criterion shifts accompanying list strength manipulations whereby responding tends to be significantly more conservative in lists of mixed strength (Dennis, 2012; Hirshman, 1995; Osth et al., 2014). It is unclear whether this conservatism is directly related to the increased noise in such lists, but future MBBE experiments could potentially manipulate list strength for paintings and words via repetition of some subset of items or variations in study time. If the criterion shift from pure strength lists to mixed strength lists, which tends to be conservative, is even more substantially so for paintings than for words, this could lend credence to such an explanation.

One caveat with respect to the above proposed explanation is that it focuses on an interpretation of zROC slope differences based on differences in target distribution variability and largely ignores the possibility of differences between words and paintings in variability of the lure distributions, which is also a reasonable possibility. Furthermore, data in the current paper are not sufficient to conclusively distinguish between DPSD and UVSD interpretations of the MBBE, or to rule either one out. Neither type of model appears to be glaringly inconsistent with any of the presented results. Future efforts should perhaps aim to design experiments or analyze data in such a way as to produce results which DPSD and UVSD would make discrepant predictions about. Diffusion model analyses of response time data seem to be one promising avenue in this regard, with a recent endeavour in this domain reporting evidence in favour of UVSD (Starns, Ratcliff, & McKoon, 2012). An MBBE experiment incorporating a source memory task might also provide further evidence for or against the DPSD account of this phenomenon. For the time being, although neither model can be rejected, it does not appear necessary to appeal to a second process to account for the existing MBBE data; therefore, a UVSD interpretation seems the most parsimonious option.

Concluding Remarks

A subset of results from the experiments described above was selected for inclusion in a book chapter entitled "Recognition Memory Response Bias is Conservative for Paintings and We Don't Know Why" (Lindsay, Kantner, & Fallow, 2015). This title was a succinct, effective summary of the state of the MBBE research line at the time, and is one that holds stubbornly – and on occasion, somewhat maddeningly – true to this day. But to deem a line of research unsuccessful because it has yet to yield a satisfactory answer to a central question is to risk selling it short; discovering what the answers are *not* can in itself generate valuable insights, and unexpected results often raise new questions and spark offshoots of inquiry.

The series of studies described in the current paper is no exception. The mechanism underlying conservative response bias in recognition memory for paintings remains enigmatic, but individual experiments have chipped away at its fortress by demonstrating a wide array of manipulations to which the effect is impervious and ruling out several potential hypotheses as to its origin. As efforts to elucidate the MBBE continue, the existing data and the variety of ways in which they have been analyzed will serve to inform decisions about whether a given direction is likely to be fruitful. The results of these experiments have also raised new questions regarding an assortment of memory phenomena, such as reminding/study phase retrieval, how the available options on a recognition test and the wording thereof might shape decisional processes, and individual differences in mnemonic strategy. Regardless of whether the cause of the MBBE is determined in the next experiment, the next ten experiments, or remains elusive for years to come, the pursuit of answers in this domain has the potential to spur new lines of inquiry and further understanding of the processes involved in recognition memory, ultimately adding its own splash of paint to the busy canvas that is our current understanding of brain and behaviour.

Bibliography

Allan, K., Wilding, E.L., & Rugg, M.D., (1998). Electrophysiological evidence for dissociable processes contributing to recollection. *Acta Psychologica, 98,* 231-252.

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

Atkinson, R. C., Holmgren, J., & Juola, J. F. (1969). Processing time as influenced by the number of elements in a visual display. *Perception & Psychophysics*, *6*(6), 321-326.

Atkinson, R. C., & Juola, J. F. (1974). *Search and decision processes in recognition memory*. WH Freeman.

Besson, G., Ceccaldi, M., Didic, M., & Barbeau, E. J. (2012). The speed of visual recognition memory. *Visual Cognition*, *20*(10), 1131–1152.

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

Balota, D. A., Burgess, G. C., Cortese, M. J., & Adams, D. R. (2002). The word-frequency mirror effect in young, old, and early-stage Alzheimer's disease: Evidence for two processes in episodic recognition performance. *Journal of Memory and Language, 46,* 199–226.

Brodeur, M.B., Chauret, M., Dion-Lessard, G., & Lepage, M. (2011). Symmetry brings an impression of familiarity but does not improve recognition memory. *Acta Psychological, 137,* 359-370.

Brown, J. (1976). An analysis of recognition and recall and of problems in their comparison. In J. Brown (Ed.), *Recall and recognition* (pp. 1-35). New York: Wiley.

Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, word frequency and negative recognition. *Quarterly Journal of Experimental Psychology*, *29*(3), 461–473.

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

Clark, S. E. (1992). Word frequency effects in associative and item recognition. *Memory & Cognition, 20,* 231-243.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review*,*3*(1), 37-60.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology, 33A*, 497-505.

Conway, M. A., & Dewhurst, S. A. (1995). Remembering, familiarity, and source monitoring. *The Quarterly Journal of Experimental Psychology*, *48*(1), 125-140.

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language*, *64*(4), 316–326. doi:10.1016/j.jml.2011.02.003

Cumming, G. (2001-2011). *ESCI, Exploratory Software for Confidence Intervals*. Computer software, available from: http://www.latrobe.edu.au/psy/research/projects/esci

Deason, R. G., Hussey, E. P., Ally, B. A., & Budson, A. E. (2012). Changes in response bias with different study-test delays: Evidence from young adults, older adults, and patients with Alzheimer's disease. *Neuropsychology*, *26*(1), 119–126.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological review*, *108*(2), 452-478.

Donaldson, W., Mackenzie, T. M., & Underhill, C. F. (1996). A comparison of recollective memory and source monitoring. *Psychonomic Bulletin & Review*,*3*(4), 486-490.

Donaldson, W., & Murdock, B. B. (1968). Criterion change in continuous recognition memory. *Journal of Experimental Psychology*, *76*(3), 325–330.

Dosher, B. A. (1984). Discriminating preexperimental (semantic) from learned (episodic) associations: A speed-accuracy study. *Cognitive psychology*, *16*(4), 519-555.

Dougal, S., & Rotello, C. M. (2007). "Remembering" emotional words is based on response bias, not recollection. *Psychonomic Bulletin & Review*, *14*(3), 423–429.

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

Eich, J. E. (1980). The cue-dependent nature of state-dependent retrieval.*Memory & Cognition*, *8*(2), 157-173.

Eldridge, L. L., Engel, S. A., Zeineh, M. M., Bookheimer, S. Y., & Knowlton, B. J. (2005). A dissociation of encoding and retrieval processes in the human hippocampus. *The Journal of Neuroscience*, *25*(13), 3280-3286.

Eng, J. (n.d.). ROC analysis: web-based calculator for ROC curves. Retrieved *16/06/2014*, from http://www.jrocfit.org.

Feenan, K., & Snodgrass, J. G. (1990). The effect of context on discrimination and bias in recognition memory for pictures and words. *Memory & Cognition*, *18*(5), 515–527.

Fisher, R. P., & Craik, F. I. (1977). Interaction between encoding and retrieval operations in cued recall. *Journal of Experimental Psychology: Human Learning and Memory*, *3*(6), 701-711.

Gehring, R. E., Toglia, M. P., & Kimble, G. A. (1976). Recognition memory for words and pictures at short and long retention intervals. *Memory & Cognition*, *4*(3), 256–260.

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 500-513.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). New York: Wiley.

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*, 554–563.

Hicks, J. L., Marsh, R. L., & Ritschel, L. (2002). The role of recollection and partial information in source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 503-508.

Hintzman, D. L. (2009). How does repetition affect memory? Evidence from judgments of recency. *Memory & Cognition*, *38*(1), 102–115. doi:10.3758/MC.38.1.102

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language*, *33*, 1–18.

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.

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), 1–7.

Humphreys, M. S., Pike, R., Bain, J. D., & Tehan, G. (1989). Global matching: A comparison of the SAM, Minerva II, Matrix, and TODAM models. *Journal of Mathematical Psychology*, *33*(1), 36-67.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541.

Jacoby, L. L., Jones, T. C., & Dolan, P. O. (1998). Two effects of repetition: Support for a dual-process model of know judgments and exclusion errors. *Psychonomic bulletin & review*, *5*(4), 705-709.

Kalaska, J., & Crammond, D. (1992). Cerebral cortical mechanisms of reaching movements. *Science*, *255*(5051), 1517-1523.

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

Kinnell, A., & Dennis, S. (2012). The role of stimulus type in list length effects in recognition memory. *Memory & Cognition*, *40*(3), 311–325. doi:10.3758/s13421-011-0164-2

Lewis, S., & Clarke, M. (2001). Forest plots: trying to see the wood and the trees. *Bmj*, *322*(7300), 1479-1480.

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* (pp.). Houndmills, UK: Palgrave Macmillan.

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*(4), 476–490.

Luce, R. D. (1959). *Individual choice behavior.* New York: Wiley.

Luce, R. D. (1963). A threshold theory for simple detection experiments. *Psychological Review, 70*, 61-79.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: a user's guide*. Mahwah, N.J.: Lawrence Erlbaum Associates.

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*(1), 185 – 199.

Mandler, G. (1980). Recognizing: the judgment of previous occurrence. *Psychological Review*, *87*(3), 252-271.

Marquié, J.C., & Baracat, B. (2000). Effects of age, educational level and gender on response bias in a recognition task. *Journal of Gerontology: Psychological Sciences, 55B*(5), 266-272.

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review*, *14*(5), 858–865.

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

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

Murnane, K., & Phelps, M. P. (1993). A global activation approach to the effect of changes in environmental context on recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(4), 882-894.

Murnane, K., & Phelps, M. P. (1994). When does a different environmental context make a difference in recognition? A global activation model. *Memory & Cognition*, *22*(5), 584-590.

Murnane, K., & Phelps, M. P. (1995). Effects of changes in relative cue strength on context-dependent recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 158-172.

Nelson, D.L., Reed, V.S., & Walling, J. R. (1976). Pictorial superiority effect. *Journal of Experimental Psychology: Human Learning and Memory*, *2*, 523-528.

Osth, A. F., Dennis, S., & Kinnell, A. (2014). Stimulus type and the list strength paradigm. *The Quarterly Journal of Experimental Psychology*, 1–16. doi:10.1080/17470218.2013.872824

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59-108.

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review*, *83*(3), 190-214.

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*(3), 518-535.

Reder, L. M., Angstadt, P., Cary, M., Erickson, M. A., & Ayers, M. S. (2002). A reexamination of stimulus-frequency effects in recognition: two mirrors for low-and high-frequency pseudowords. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(1), 138-152.

Rotello, C. M., & Heit, E. (2000). Associative recognition: A case of recall-to-reject processing. *Memory & Cognition*, *28*(6), 907–922.

Rotello, C. M., & Macmillan, N. A. (2007). Response Bias in Recognition Memory. In *Psychology of Learning and Motivation* (Vol. 48, pp. 61–94). Elsevier.

Rotello, C. M., Macmillan, N. A., & Van Tassel, G. (2000). Recall-to-reject in recognition: Evidence from ROC curves. *Journal of Memory and Language, 43,* 67-88.

Rotello, C. M., & Zeng, M. (2008). Analysis of RT distributions in the remember—know paradigm. *Psychonomic Bulletin & Review*, *15*(4), 825–832. doi:10.3758/PBR.15.4.825

Schneider, W., Eschman, A., & Zuccolotto, A. (2002a). E-Prime User's Guide. Pittsburgh: Psychology Software Tools, Inc.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002b). E-Prime Reference Guide. Pittsburgh: Psychology Software Tools, Inc.

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, *6*(1), 156–163.

Singer, M., & Wixted, J. T. (2006). Effect of delay on recognition decisions: Evidence for a criterion shift. *Memory & Cognition*, *34*(1), 125–137.

Slotnick, S. D. (2010). "Remember" source memory ROCs indicate recollection is a continuous process. *Memory*, *18*(1), 27–39. doi:10.1080/09658210903390061

Starns, J. J. (2014). Using response time modeling to distinguish memory and decision processes in recognition and source tasks. *Memory & Cognition*. *42*(8), 1357-1372.doi:10.3758/s13421-014-0432-z

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of zROC slopes with response time data and the diffusion model. *Cognitive Psychology*, *64*(1-2), 1–34. doi:10.1016/j.cogpsych.2011.10.002

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne*, *26*(1), 1-12.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological review*, *80*(5), 352-373.

Vanrullen, R., & Thorpe, S. J. (2001). The time course of visual processing: from early perception to decision-making. *Journal of cognitive neuroscience*,*13*(4), 454-461.

Windmann, S., & Chmielewski, A. (2008). Emotion-induced modulation of recognition memory decisions in a Go/NoGo task: Response bias or memory bias? *Cognition & Emotion*, *22*(5), 761–776. doi:10.1080/02699930701507899

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*(4), 681–690.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152–176. doi:10.1037/0033-295X.114.1.152

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1341-1354.

Yonelinas, A. P. (2002). The Nature of Recollection and Familiarity: A Review of 30 Years of Research. *Journal of Memory and Language*, *46*(3), 441–517. doi:10.1006/jmla.2002.2864

Yonelinas, A. P., & Jacoby, L. L. (1994). Dissociations of processes in recognition memory: effects of interference and of response speed. *Canadian Journal of Experimental Psychology*, *48*(4), 516-535.

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, *133*(5), 800–832. doi:10.1037/0033-2909.133.5.800

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

Appendix A: List of relevant experiments

Research Line/Experiment #		**N**
Paintings only (Lindsay & Kantner, 2011)		**233**
1		52
2		27
3		20
4		21
5		113
Materials-based Bias Effect (MBBE)		**727**
Within-subjects		
1		21
2		**54**
	a. Feedback[1]	26
	b. Control	28
Subjective memorability		
3		38
4 [2]		51
5 [2]		84
6		48
Reminding		
7 [3]		59
Within vs. between		
8		**116**
	a. Within	50
Between-subjects		
	b. Between	**66**
	Words	33
	Paintings	33
9		**80**
	Words	40
	Paintings	40
Within-subjects		
Confusability		
10 [3,4]		**83**
	a. Basic	39
	b. Remind	44
Response deadline		
11 [4]		27