Memory & Cognition (2022) 50:378–396
https://doi.org/10.3758/s13421-021-01227-5

Test position effects on hit and false alarm rates in recognition memory for paintings and words

Kaitlyn M. Fallow[1] · D. Stephen Lindsay[1]

Accepted: 30 July 2021 / Published online: 23 September 2021

Abstract
When old/new recognition memory is tested with equal numbers of studied and nonstudied items and no rewards or instructions that favour one response over the other, there is no obvious reason for response bias. In line with this, Canadian undergraduates have shown, on average, a neutral response bias when we tested them on recognition of common English words. By contrast, most subjects we have tested on recognition of richly detailed images have shown a conservative bias: they more often erred by missing a studied image than by judging a nonstudied image as studied. Here, in an effort to better understand these materials-based bias effects (MBBEs), we examined changes in hit and false alarm (FA) rates (and in sensitivity and bias) from the first to fourth quartile of a recognition memory test in eight experiments in which undergraduates studied words and/or images of paintings. Response bias for images tended to increase across quartiles, whereas bias for words showed no consistent pattern across quartiles. This pattern could be described as an increase in the MBBE over the course of the test, but the underlying patterns for hits and FAs are not easily reconciled with this interpretation. Hit rates decreased over the course of the test for both materials types, with that decline tending to be steeper for images than words. For words, FA rates tended to increase across quartiles, whereas for paintings FA rates did not increase across quartiles. We discuss implications of these findings for theoretical accounts of the MBBE.

Keywords Recognition memory · Response bias

Some types of stimuli tend to be better remembered than others. One example is the picture superiority effect, the general tendency for recall (e.g., Erdelyi et al., 1989; Paivio et al., 1968) and recognition (e.g., Defeyter et al., 2009; Fawcett et al., 2012; Gehring et al., 1976) to be better for pictures than words (for a review, see Madigan, 1983). Lindsay and Kantner (2011) stumbled across evidence that recognition memory response bias can also be affected by stimulus type (see also Lindsay et al., 2015). In numerous studies, most undergraduates tested on old/new recognition memory for scans of paintings showed a conservative response bias (i.e., when they erred it was more often by calling a studied painting "new").

Scientific interest in recognition memory response bias and the related signal detection construct of the decision criterion has grown dramatically in the last 20 years (Aminoff et al., 2012; Bowen et al., 2020; Cox & Shiffrin, 2012; Frithsen et al., 2018; Han & Dobbins, 2008; Heit et al., 2003; Hilford et al., 2019; Kent et al., 2018; Koop et al., 2019; Megla et al., 2021; Miller et al., 2001; Rhodes & Jacoby, 2007; Rotello et al., 2006). Criterion shifts have been put forth as a potential explanation for a number of mysterious effects in the recognition literature, such as strength-based mirror effects (Hirshman, 1995; Hockley & Niewiadomski, 2007) and the revelation effect (Aßfalg et al., 2017; Verde & Rotello, 2004). Response bias differences may also partly account for discrepancies across studies in how certain variables, such as emotional valence, affect memory performance (Dougal & Rotello, 2007; Grider & Malmberg, 2008). Understanding response bias and its mechanisms is crucial to developing a full picture of how recognition memory decisions are made.

✉ Kaitlyn M. Fallow
kmfallow@uvic.ca

[1] Department of Psychology, University of Victoria, P.O. Box 1700, STN CSC, Victoria, British Columbia V8W 2Y2, Canada

Variables associated with response bias differences

Researchers can induce more conservative or liberal biases on a recognition test by instructing subjects to be more or less lenient in endorsing items as old (Azimian-Faridani & Wilding, 2006;

Postma, 1999) or by giving subjects a larger reward for one type of correct response (Curran et al., 2007; Healy & Kubovy, 1978; Van Zandt, 2000). Another technique is to provide information—whether accurate or misleading—about the proportion of old items on the test (Criss, 2009; Rotello et al., 2006; Strack & Forster, 1995; Van Zandt, 2000). In addition to explicit incentives or instructions to favour a particular response, response bias can also vary as a function of certain stimulus features and elements of the experiment design (Hockley, 2011). Conditions of higher overall similarity between targets and distractors (Benjamin & Bawa, 2004; Brown et al., 2007), changes to stimulus context between study and test (Feenan & Snodgrass, 1990; Goh, 2005; Macken, 2002), and greater stimulus distinctiveness (Dobbins & Kroll, 2005; Lukavský & Děchtěrenko, 2017) have all been associated with more conservative responding, whereas more liberal biases have been observed with longer delays between study and test (Deason et al., 2012; Gehring et al., 1976; Singer & Wixted, 2006) and when test cues are degraded or obscured relative to studied items (Kent et al., 2018; Vokey & Hockley, 2012).

Some variables that affect response bias make intuitive sense, such as responding more conservatively when correct rejections are more highly rewarded or there are reasons to assume that studied stimuli will be easily remembered. But response bias effects are often inconsistent, context-sensitive, or otherwise more complex than such generalizations suggest. Researchers have long noted that response bias and shifts therein tend to be suboptimal, sometimes strikingly so. Participants rarely adjust responding as much as they should in response to payoff, probability, and difficulty manipulations (Aminoff et al., 2012; Benjamin & Bawa, 2004; Healy & Kubovy, 1978; Ratcliff et al., 1992; Verde & Rotello, 2007), and similar response patterns have been observed across mixed old/new recognition tests and those comprising exclusively old or new items (J. C. Cox & Dobbins, 2011; Ley & Long, 1987; Wallace, 1978; Wallace et al., 1978). Further, some manipulations that reliably affect response bias when applied to separate groups exert inconsistent or null effects in within-subjects designs (Hockley, 2011; Singer, 2009).

There is also mounting evidence for consistent individual differences in both overall tendency toward responding liberally or conservatively (Kantner & Lindsay, 2012, 2014) and in the extent of strategic response bias shifts (Aminoff et al., 2012; Frithsen et al., 2018; for a review, see Miller & Kantner, 2020). Efforts to explore systematic sources of individual differences in response bias on the basis of variables such as age and education have produced mixed results. Studies looking at age differences, for example, have variously found no difference in bias between younger and older adults (Deason et al., 2012), a more conservative bias in older than younger adults (Criss et al., 2014), and a tendency for bias to become more conservative with age among only the most highly educated subsample (Marquié & Baracat, 2000).

Within-test variation in response bias

Recognition memory response bias, much like sensitivity and accuracy, seems to depend on a variety of subject-level, experimental, and stimulus-based variables and the interactions among them. As alluded to above, there has been substantial research interest in how response bias can change within a single recognition test. Much of this has been in the context of debates regarding the nature and prevalence of strategic within-list bias shifts, but some have investigated less controllable sources of trial-by-trial response variability, such as sequential dependencies (Dopkins et al., 2010; Marken & Sandusky, 1974) and random noise in the decision process (Benjamin, 2013; Benjamin et al., 2009). Unlike sensitivity and other accuracy measures, which typically decline over the course of a recognition test (although the sources of this decline remain open to debate; see e.g., Malmberg et al., 2012), we are not aware of any consistent effects of test position on response bias. Examples can be found of bias becoming increasingly liberal (Berch & Evans, 1973; Donaldson & Murdock, 1968) or conservative (Osth et al., 2018; Potter et al., 2002; Ratcliff, 1978) over the course of a single recognition test, and of more nuanced patterns such as an initial liberal shift followed by stabilization (Criss et al., 2011). The relationship between test position and response bias may be sensitive to some of the same variables that affect overall response bias, but that question has received little attention relative to test position effects on recognition accuracy. We explored position-based effects on bias and sensitivity in the context of a broader effect of stimulus materials on bias.

A materials-based bias effect in recognition memory

We have observed a materials-based response bias effect that is consistently obtained in both within- and between-subjects designs, is robust to at least some procedural differences, and holds across variations in overall performance. Here, we focus on how this effect and its constituent response rates vary over the course of a recognition test. Our results (a) demonstrate the importance of examining raw response rates in addition to assumption-laden aggregate measures of bias and sensitivity, (b) illustrate some of the challenges of inference in recognition memory, and (c) point to stimulus materials as one potentially informative variable in future work on test position effects on various measures.

Our lab's interest in materials-based differences in response bias originated with a series of studies conducted by Lindsay and Kantner (2011). They were interested in the effects of accuracy feedback on recognition memory for complex, unfamiliar stimuli (namely, poetry excerpts, Korean melodies, and digital images of obscure masterwork paintings). Results did not suggest feedback had any consistent beneficial effect on sensitivity (d'), but Lindsay and Kantner noted that mean response bias (c) was significantly conservative in most cases, even though there had been a 1:1 old/new ratio and no incentive or encouragement to err toward the "new" response. This trend was especially pronounced in five experiments that had used paintings as stimuli, in which bias was significantly conservative in 11 of 12 groups.[1] Ten follow-up studies comparing response bias for paintings and words, in which stimulus materials were variously manipulated within or between subjects, found response bias for paintings was significantly conservative—and significantly *more* conservative than bias for words—in all cases (Lindsay et al., 2015). For words, average bias was neutral in between-subjects designs and liberal when materials were manipulated within subjects.

The above pattern of materials-based differences in response bias (which we refer to as the materials-based bias effect or MBBE) held despite substantial differences across the 10 experiments with respect to relative mean sensitivity for the two types of materials (such differences were created by adding orienting tasks at study and/or varying the composition of the stimulus sets, e.g., by excluding some of the most memorable painting stimuli). Sensitivity in most experiments showed a picture superiority effect, but this effect was reversed in two studies with a pleasantness judgment orienting task, and in three studies sensitivity was roughly equal for paintings and words.

To date, the results described above have been reported only in brief summary form in two chapters and in conference posters/papers. The current manuscript highlights the results of new follow-up analyses (suggested by Jim Nairne, personal communication, 2013) exploring changes in recognition memory judgments to studied and non-studied words and paintings as a function of test position. These analyses yielded surprising and informative patterns suggesting that overall-test-level materials-based differences in response bias are only part of the story.

This is not a typical *Memory & Cognition* paper. We do not report a series of experiments, but something akin to a mega-analysis in which data from multiple experiments were analyzed in a new way. We did not go into these analyses with a specific hypothesis, nor did we emerge with a clear sense of the implications of our results for memory theory. Despite the remarkable consistency of the MBBE across experiments that differed in methods, stimuli, and overall performance, its underlying mechanism has proven elusive.

Part of the challenge of studying response bias effects is that it is not always straightforward to discern exactly *what* they are, let alone why they occur. For example, effects on the well-known signal detection theory (SDT)-based bias measure c are often interpreted as definitive evidence the manipulation in question exerts some influence on decision-making processes, when in fact c is sensitive to bias in any constituent process(es) of the task at hand (e.g., see Witt et al., 2015, for a compelling demonstration of differences in c arising from perceptual factors). Efforts to understand response bias differences may be doomed to fail if all hypotheses take for granted that the mechanism involves the decision criterion.

The above is just one example of how our thinking evolved throughout the course of developing this paper. Numerous questions regarding measurement and inference in recognition memory are far from settled, and we changed our minds several times regarding how best to present and interpret our results. Ultimately, we decided to present the results as fully as possible (between the paper and supplemental materials) without offering much in the way of explanation, although we do offer some caveat-laden speculation in the discussion. Some readers may, understandably, find this unsatisfying. We share concerns that it can be counterproductive to get bogged down in the details of individual effects at the expense of big picture memory theory (e.g., as articulated by Hintzman, 2011). But we think there is value in the mystery we present here, and hope the discussion will clarify why we chose this relatively noncommittal path. Follow-up work may shed further light on what underlies these effects and their broader theoretical relevance, or perhaps these results will catch the eye of someone with different analytic or theoretical expertise than we have and spur new insights. The results point toward several avenues that may prove fruitful in future research investigating the basic mechanisms underlying these materials-based bias effects, with the ultimate goal of understanding the broader implications of such differences for general theories of human recognition memory.

Method

We analyzed data from 8 of the 10 experiments briefly summarized in Lindsay et al. (2015).[2] In seven of these experiments, subjects studied and were tested on a mix of paintings and words (i.e., stimulus materials were manipulated within subjects). One of these seven experiments also included a

[1] Each experiment included a feedback and control group, and one experiment also attempted to manipulate motivation between subjects, but response bias did not differ as a function of these variables.

[2] Data from the other two studies were originally included in the primary analyses described here (and produced similar results to those reported). However, both of these experiments used an unusual procedure during the study phase (participants were asked to indicate each time an item reminded them of a previously presented item), and one used an atypical test scale (in addition to standard "definitely studied" and "definitely not studied" options, there were two options for indicating uncertainty as to whether the current item, or just one very similar to it, had been seen at study). For the sake of simplicity, we have restricted our analyses here to studies that used more standard recognition memory procedures. Data for all 10 studies can be accessed (osf.io/3qfk5).

between-subjects condition (i.e., materials were paintings for some subjects and words for others), and the eighth experiment served as a replication of this between-subjects condition. We report the method for all eight experiments together. More details regarding the methods of each experiment, including the wording of instructions and experiment-specific manipulations, are available (osf.io/3qfk5/).

Participants

Participants were 499 undergraduate students at the University of Victoria who completed the experiments for optional bonus course credit between 2009 and 2012. Demographic data were not collected in most experiments, but the pool from which subjects were drawn is composed largely of 18-to-25-year-olds (78%) who identified as female (70%) and Caucasian (74%; numbers are as of 2014). The sample sizes for each experiment are shown in Table 1. Sample sizes were not planned according to current best practices, but were instead determined by prevailing norms at the time (e.g., typical sample sizes in the literature) and practical/time constraints on data collection.

Materials

All experiments were administered on desktop PCs using E-Prime software (Schneider et al., 2002a, 2002b). Painting stimuli were 234 digital scans of masterwork paintings by renowned artists (some extremely famous, such as Rembrandt and van Gogh, and others somewhat less so, such as Mary Cassatt and Gustave Caillebotte). The paintings were in various styles and depicting a wide range of subjects and themes (e.g., portraits, landscapes, still lifes). The images used differed somewhat among experiments, but all were selected from a larger collection originally assembled by Jeffrey P. Toth. We excluded extremely well-known paintings (e.g., Van Gogh's *Starry Night*, Munch's *The Scream*). The bitmap images ranged in size from 130 to 500 pixels in width and 270–500 pixels in height and were displayed in E-Prime at a maximum size of 75–95% of monitor width and height, with stretch settings set to prevent distortion. Word stimuli were 288 three-to-eight-letter medium-to-high-frequency English nouns[3] obtained from the MRC psycholinguistic database (bit.ly/mrc1981; Coltheart, 1981).

In each experiment the study list comprised 96 critical items bookended by 3–6 primacy and recency buffers. The test list included all 96 studied critical items plus 96 nonstudied items for a total of 192 trials. Study and test lists were randomly generated anew for each participant, such that individual words and paintings varied across subjects with respect to old/new status and study and/or test position. In within-subject versions of the experiment (i.e., Experiments 1–7a), half of the items at study and test were paintings and the remainder were words.

Procedure

At the beginning of the study phase, participants were told they would view a series of items (paintings and/or words, depending on the experiment) for a brief time each. They were asked to attend to each item and remember it for a later memory test. Stimuli were presented one at a time in the centre of a white background, preceded by a black central fixation cross. Each stimulus was presented for 1 s, except in Experiments 4 (2 or 3 s each) and 5 (2 s each), in which presentation times were longer to allow participants to make 3-point (Experiment 4) or 2-point (Experiment 5) pleasantness judgments for each item (via key press). Experiments 1–3 had 1,400-ms interstimulus intervals (ISIs; including a 1-s fixation cross) and Experiments 5–8 all had 900-ms ISIs (500-ms fixation cross). Experiment 4 had a mix of these two ISI structures.[4] Between the study and test phases, there was a filler/delay period lasting roughly 5 minutes. In some cases, participants answered demographic questions or questions related to experiment-specific hypotheses during this delay interval (e.g., predicting the percentages of paintings and words they expected to successfully recognize; see Lindsay et al., 2015), but these responses are not of interest here. In other experiments, this delay only included a task unrelated to the experiment that was administered solely as a distractor (e.g., participants were asked to write the names of as many countries as they could think of in 5 minutes).[5]

Test phase instructions and structure were similar across all experiments. Participants were told that they would again see a series of items, some of which had been presented in the previous study list and others that had not, and asked to decide whether each item was old/studied or new/unstudied. Old/new decisions were made on a 6-point confidence-weighted scale ranging from 1 (*definitely new*) to 6 (*definitely old*). Test items appeared one at a time and participants responded at their own pace using the number keys. The response scale remained onscreen throughout the test for reference. In Experiment 2, some participants received accuracy feedback throughout the

[3] The original intent was to use four- to eight-letter words, but one three-letter word made it into the set in some experiments. Additionally, one word ("bridge") was accidentally included twice in Experiment 1. The full word set (and accompanying norm data for frequency and other psycholinguistic variables, where available) can be viewed at (osf.io/3qfk5).

[4] Analyses conducted partway through data collection (at $N = 35$) for Experiment 4 showed many participants were performing near ceiling, so both stimulus presentation time and ISIs were reduced (from 3 to 2 s and 1,400 to 1,900 ms, respectively) for the final 17 participants.

[5] Tasks administered during this filler period were mostly completed on paper, so experiment-level details of the exact task and duration have been lost to time for some of the earlier studies.

Table 1 Sample sizes and numbers of replaced hit rates (HRs) and false alarm rates (FARs)

Experiment	*N*		Materials	Ceiling (HR = 1)					Floor (FAR = 0)				
	Total	Analyzed		Whole test	Quartile				Whole test	Quartile			
					1	2	3	4		1	2	3	4
Within subjects													
1	21	21	Paintings	0	2	2	0	0	1	5	4	5	9
			Words	0	4	4	2	3	0	0	0	0	0
2	54	53	Paintings	0	4	3	0	2	0	22	14	14	16
			Words	1	7	9	7	7	0	11	7	3	6
3	39	38	Paintings	0	1	2	0	0	2	14	13	12	12
			Words	0	5	5	9	5	1	7	5	3	3
4	52	51	Paintings	1	15	8	5	4	1	12	19	17	24
			Words	12	32	26	27	25	4	22	20	13	15
5	84	84	Paintings	1	21	9	8	3	7	21	23	27	32
			Words	13	47	47	41	35	1	18	19	16	15
6	48	46	Paintings	0	1	1	0	0	1	7	12	14	10
			Words	0	4	5	2	7	0	2	1	1	4
7a	51	45	Paintings	0	6	3	2	0	2	7	7	13	12
			Words	0	3	5	4	2	1	2	3	2	2
Between subjects													
7b	34	33	Paintings	0	0	0	0	0	0	3	2	1	1
	36	35	Words	0	0	0	0	0	0	1	0	0	0
8	40	40	Paintings	0	0	0	0	1	0	4	2	1	1
	40	37	Words	0	0	0	0	1	0	1	0	0	0

test phase, but data were collapsed across conditions, as this manipulation was not of interest for current purposes.

Analysis details

Unless otherwise specified, all analyses were conducted using R (Version 3.5.2; R Core Team, 2018) in RStudio (Version 1.1.463; RStudio Team, 2016). We relied extensively on *tidyverse* packages (Wickham, 2017) for rearranging, summarizing, and plotting data. Confidence-weighted responses were collapsed to binary old/new judgments by counting responses of 4, 5, or 6 as "old" and responses of 1, 2, or 3 as "new" to enable conventional SDT-based analyses. Hit (HR) and false alarm rates (FAR) were calculated, and rates of 1 or 0 were replaced according to Macmillan and Kaplan (1985; $1–0.5/n_{old}$ & $0.5/n_{new}$, respectively). The number of ceiling and floor replacements per experiment and materials type can be seen in Table 1. HRs and FARs were calculated separately for words and paintings where applicable (i.e., in Experiments 1–7a) and used to calculate sensitivity (d'; $z_{HR} - z_{FAR}$) and response bias (c; $-0.5 \times [z_{HR} + z_{FAR}]$). The above measures were first calculated at the subject level for each experiment and materials type. Participants with d' below 0.2 (for either materials type, in the within-subjects case) were excluded from further analysis. This is admittedly an arbitrary cut-off, but given the high levels of performance generally observed in these experiments, it was chosen as a relatively conservative means of excluding participants who were likely disengaged from the task (or, in the case of the within-subjects experiments, perhaps attending only to one materials type). This criterion led to the exclusion of one participant in a paintings-only condition (1% of all such participants), four participants in a words-only condition (5%), and 10 participants (six for words, four for paintings) from within-subjects experiments (3%). The full trial-level data shared at osf.io/3qfk5 include these participants. Data for one additional participant in Experiment 4 were neither analyzed nor included in this final data file because they were excluded from analyses (for unknown reasons) at the time that experiment was conducted. Postexclusion sample sizes for each experiment can be seen in Table 1.

Test quartile analyses

We divided the 192-item test list into ordered quartiles of 48 items each and calculated hit and false alarm rates, and subsequently c and d', for words and/or paintings at the quartile level for each subject. In addition to the replacements of ceiling-level HRs and floor-level FARs indicated in Table 1, two instances of ceiling FARs for words at the quartile level were also replaced. It should be noted that such replacements were made at undesirably high rates in some cases, especially in Experiments 4 and 5. We outline some of the constraints this imposes on interpretation of these results in the Discussion.

All dependent measures were averaged across subjects within each experiment (and materials type in Experiments 1–7a) and plotted with 95% BCa bootstrap confidence intervals (CIs; Efron, 1987)[6] based on 10,000 bootstrap resamples. Bootstrap analyses were conducted using the *boot* package (Canty & Ripley, 2020; Davison & Hinkley, 1997) in R. Distributions of HRs and FARs were in some cases heavily skewed, so we report corresponding results for medians in the supplementary material.

Mega-analyses

To facilitate evaluation of the overall, cross-experimental trends in these quartile-level analyses, we conducted mega-analyses for each dependent measure. By contrast with typical meta-analytic approaches that rely on experiment- or group-level effect sizes, mega-analysis (also referred to as individual participant/patient data [IPD] meta-analysis) combines participant-level data across experiments (e.g., Cooper & Patall, 2009), preserving the statistical power provided by this large number of observations. In this case, we collapsed data from all eight experiments into two sets based on whether materials type was manipulated within ($N = 338$) or between subjects ($Ns = 73$ and 72 for paintings and words, respectively), based on our previous findings of differences across manipulation types (e.g., in the test-level materials-based bias effect). To account for other across-experiment differences in these measures that were not of particular interest here (e.g., d' scores tended to be very high in Experiments 4 & 5, which included orienting tasks; see Table 2 for test-level summary statistics), each participant's quartile-level scores were converted to z-scores based on the test-level mean and standard deviation for the corresponding experiment and/or materials type.[7]

These scores were then subjected to an analysis of variance (ANOVA) using the *ez* package in R (Lawrence, 2016). To evaluate potential interactions, we conducted 2 (materials type) × 4 (test quartile) ANOVAs (for the within-subjects data this analysis was fully repeated measures, whereas the between-subjects analysis was of course mixed). ANOVA results were also used to generate 99% within-subjects CIs for plotting with the quartile-level means (Loftus & Masson, 1994). We opted for this more stringent alpha of .01 in the mega-analyses because of the large overall sample size and the number of comparisons involved. For within-subjects data these CIs were based on the results of the 2 × 4 ANOVA mentioned above, whereas CIs for the between-subjects data were derived from separate one-way repeated measures ANOVAs conducted for each materials type (i.e., with test quartile as the only independent variable). The results of these one-way ANOVAs are not discussed here but are included on the supplemental OSF page (osf.io/3qfk5/). In all cases, both CIs and p values were corrected for sphericity violations when Mauchly's test was significant at the .05 level. The Hyund–Feldt correction was applied when ε was greater than 0.75; otherwise, the Greenhouse–Geisser correction was applied. Nonsignificant results of interest were followed up with Bayesian analyses conducted using JASP (Version 0.10.2.0; JASP Team, 2019).

Results

Quartile analyses

Figures 1, 2 and 3 show quartile-level means (with 95% BCa bootstrap CIs) for all experiments and dependent measures (c in Fig. 1, d' in Fig. 2, HRs & FARs in Fig. 3). Within-subjects data are shown in panels (a) through (g) and between-subjects data in panels (h) and (i). Differences across experiments may suggest some potentially informative avenues for future study, but our interest here is mainly in overall, across-experiment trends, so we will focus our discussion on the results of the mega-analyses. Readers interested in exploring the experiment-level data further can access the full trial-level data (osf.io/3qfk5).

The results of the mega-analyses are shown in Figs. 4 (c & d') and 5 (HRs & FARs) with 99% CIs. ANOVA results for these mega-analyses are presented in Tables 3 and 4. To reiterate, quartile-level means in these analyses (unlike those in Figs. 1, 2 and 3) are not directly interpretable on the measures' original scales, but represent the means of z-scores calculated for each participant based on experiment-level means and standard deviations for each measure and materials type. To clarify, a positive value of z-transformed c for paintings indicates that c in that test quartile tended to exceed the experiment-level average for paintings. Given the nature of these scores, the main effect of materials here is not of particular interest, but the quartile main effect and

[6] We originally used within-subjects CIs (Loftus & Masson, 1994) derived from repeated-measures ANOVA, but in light of the small sample sizes in some experiments and frequent sphericity violations a bootstrap approach was deemed more appropriate. We thank Caren Rotello for this suggestion (personal communication, November 6, 2019).

[7] One could accomplish basically the same thing by analyzing the unstandardized variables with Experiment as a between-subjects factor in the ANOVAs. We opted for the standardization approach because we think plotting these means and accompanying within-subjects CIs conveys the trends we are interested in more effectively than plotting means of unstandardized values with CIs that include interexperiment variance. However, based on reviewers' comments and interest in this interexperiment variance, we have also included the results of these alternative ANOVAs with experiment included as a between-subjects factor on the supplemental OSF page (https://osf.io/3qfk5/).

Table 2 Test-level means and medians for all dependent variables

Experiment	Materials	Hit rate		False alarm rate		Sensitivity (d')		Response bias (c)	
		Mean (*SD*)	Median	Mean (*SD*)	Median	Mean (*SD*)	Median	Mean (*SD*)	Median
Within Ss									
1	Paintings	0.7 (0.14)	0.71	0.16 (0.13)	0.15	1.76 (0.73)	1.78	0.29 (0.38)	0.29
	Words	0.79 (0.09)	0.79	0.38 (0.1)	0.38	1.18 (0.51)	0.99	−0.26 (0.21)	−0.28
2	Paintings	0.65 (0.14)	0.67	0.13 (0.09)	0.13	1.67 (0.7)	1.71	0.43 (0.26)	0.47
	Words	0.79 (0.13)	0.81	0.3 (0.19)	0.29	1.52 (0.76)	1.26	−0.13 (0.45)	−0.11
3	Paintings	0.67 (0.12)	0.67	0.12 (0.09)	0.10	1.77 (0.69)	1.69	0.43 (0.27)	0.42
	Words	0.79 (0.14)	0.82	0.31 (0.19)	0.28	1.52 (0.71)	1.41	−0.15 (0.47)	−0.16
4	Paintings	0.8 (0.12)	0.83	0.11 (0.09)	0.08	2.29 (0.6)	2.27	0.21 (0.37)	0.24
	Words	0.92 (0.09)	0.96	0.15 (0.14)	0.10	2.86 (0.77)	2.92	−0.22 (0.46)	−0.24
5	Paintings	0.77 (0.12)	0.79	0.14 (0.11)	0.10	2.04 (0.64)	1.98	0.2 (0.38)	0.18
	Words	0.91 (0.12)	0.94	0.21 (0.15)	0.20	2.52 (0.91)	2.69	−0.31 (0.39)	−0.26
6	Paintings	0.59 (0.12)	0.58	0.17 (0.11)	0.14	1.33 (0.55)	1.29	0.42 (0.32)	0.38
	Words	0.76 (0.12)	0.77	0.37 (0.17)	0.38	1.18 (0.53)	1.12	−0.19 (0.42)	−0.23
7a	Paintings	0.63 (0.12)	0.65	0.16 (0.1)	0.15	1.46 (0.71)	1.38	0.36 (0.24)	0.35
	Words	0.71 (0.13)	0.71	0.42 (0.17)	0.42	0.87 (0.5)	0.75	−0.18 (0.41)	−0.22
Between Ss									
7b	Paintings	0.63 (0.13)	0.65	0.21 (0.11)	0.21	1.21 (0.56)	1.07	0.27 (0.29)	0.22
	Words	0.64 (0.14)	0.65	0.31 (0.16)	0.28	0.94 (0.44)	0.91	0.08 (0.37)	0.12
8	Paintings	0.59 (0.16)	0.61	0.24 (0.15)	0.26	1.03 (0.47)	0.96	0.27 (0.43)	0.28
	Words	0.64 (0.13)	0.63	0.38 (0.14)	0.40	0.72 (0.37)	0.58	−0.02 (0.33)	−0.05

interaction terms capture conceptually similar variance as they would in an analysis of raw scores. These analyses showed a significant main effect of test quartile for all variables in the within-subjects data, *c*: $F(3, 1011) = 23.42$, $p < .0001$, $\eta G^2 = 0.016$; hit rates: $F(2.91, 980.55) = 70.8$, $p < .0001$, $\eta G^2 = 0.045$; false alarm rates: $F(3, 1011) = 6.16$, $p = .0004$, $\eta G^2 = 0.004$; d': $F(3, 1011) = 61.59$, $p < .0001$. In the between-subjects analysis, the main effect of test quartile was significant for *c*, $F(2.8, 399.78) = 4.81$, $p = .0034$, $\eta G^2 = 0.011$; d', $F(3, 429) = 25.35$, $p < .0001$, $\eta G^2 = 0.063$; and hit rates, $F(3, 429) = 29.01$, $p < .0001$, $\eta G^2 = 0.058$, but not for false alarm rates, $F(2.87, 411.12) = 1.68$, $p = .1723$.

Given our interest in the possibility of materials-based differences in how responding changes over the course of the recognition test, the interaction between test quartile and materials was the main focus of these analyses. Mega-analysis of the within-subjects data showed an interaction for *c*, $F(3, 1011) = 44.03$, $p < .0001$, $\eta G^2 = 0.027$. As can be seen in Fig. 4a (and in most of the individual experiment-level data in Fig. 1a–g), response bias for paintings tended to increase over the course of the test, whereas bias for words remained relatively stable. The Materials × Quartile interaction was also significant for hit rates, $F(2.9, 977.42) = 26.03$, $p < .0001$, $\eta G^2 = 0.016$, and for false alarm rates, $F(2.96, 995.88) = 15.07$, $p < .0001$, $\eta G^2 = 0.009$, in the within-subjects mega-analyses. For hit rates, this took the form of a steeper across-quartile decline for paintings than words (Fig. 5a), whereas false alarm rates tended to increase over the course of the test for words but not paintings (Fig. 5b). Consistent with a mass of prior research, and not of central interest here, d' declined for both materials types (Fig. 4c). The Materials × Quartile interaction was not significant for d' in the within-subjects data. $F(3, 1011) = 0.48$, $p = .6994$, and supplementary Bayesian analyses (conducted in JASP using the default priors described by Rouder et al., 2012) strongly favoured models without an interaction term over those including an interaction (BF_{excl}= 542.34). In other words, the pattern of decline in overall sensitivity across quartiles was similar for words and paintings.

In the between-subjects mega-analyses, the Materials × Quartile interaction was not significant for any variable, *c*: $F(2.8, 399.78) = 2.32$, $p = .0791$, Fig. 4b; d': $F(3, 429) = 0.44$, $p = .7259$, Fig. 4d; hit rates: $F(3, 429) = 2.38$, $p = .0687$, Fig. 5b; false alarm rates: $F(2.87, 411.12) = 0.5$, $p = .6772$, Fig. 5d. Follow-up Bayesian analyses strongly favoured models excluding the interaction term for false alarm rates (BF_{excl} = 809.71) and for d' (BF_{excl} = 53.67), but evidence against the interaction was more moderate for *c* (BF_{excl} = 5.09) and for hit rates (BF_{excl} = 4.00).

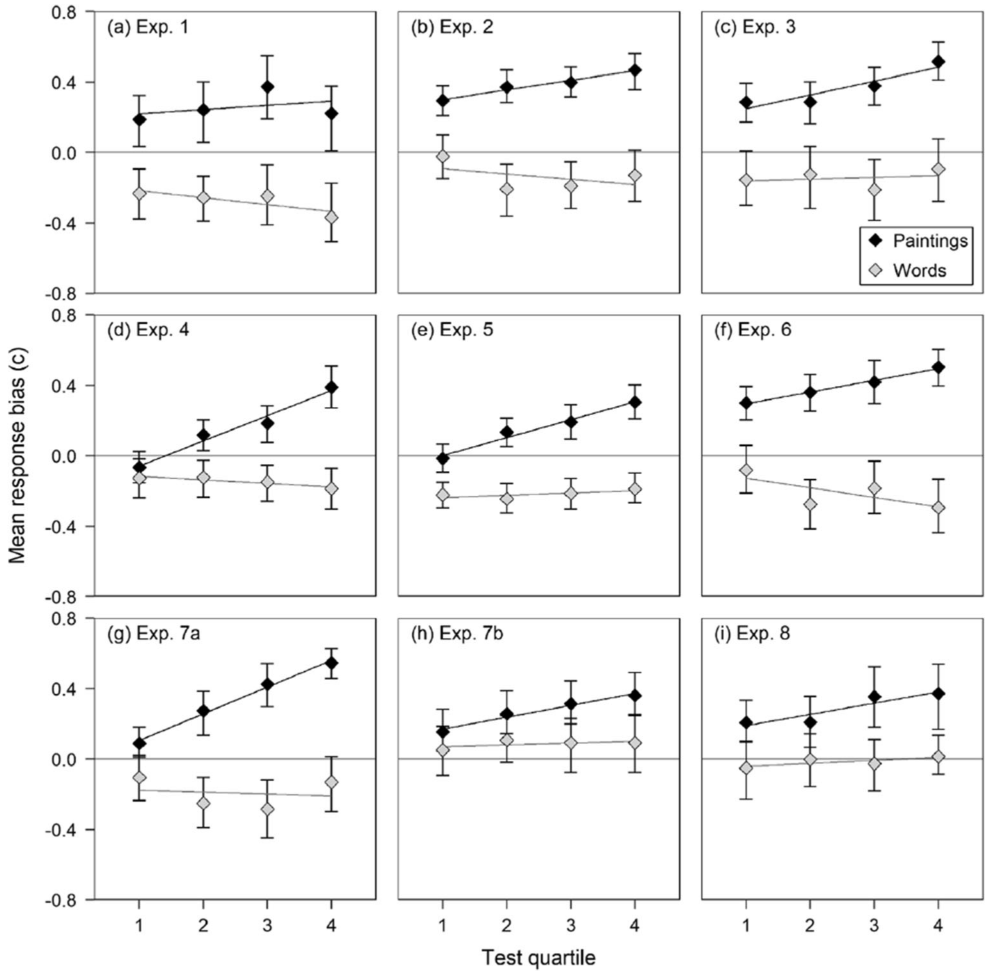

Fig. 1 Mean recognition memory response bias (c) by test quartile and accompanying regression lines for paintings and words in seven experiments in which item type was manipulated within-subjects (**a–g**) and two with a between-subjects manipulation (**h–i**). Error bars are 95% BCa bootstrap confidence intervals (Efron, 1987)

Discussion

We previously reported a consistent pattern of materials-based differences in recognition memory response bias calculated by collapsing across all test trials (Lindsay et al., 2015; Lindsay & Kantner, 2011). Here, we have reanalyzed data from eight such experiments by dividing test trials into ordered quartiles, and found that these materials-based bias differences were usually present across all four quartiles but tended to increase over the course of the test. Specifically, with only a few exceptions (see Fig. 1), mean response bias was conservative for paintings and *more* conservative for paintings than for words in all test quartiles. However, response bias for paintings tended to increase (i.e., using conventional response bias terminology, become more conservative) across quartiles, whereas bias for words showed no consistent pattern across quartiles. Mega-analyses of c scores standardized relative to experiment-level data substantiated this impression for within-subjects data, showing a clear Materials × Quartile interaction (Fig. 4a): C increased monotonically across quartiles for paintings but was approximately flat for words (with the exception of an initial decrease between the first and second quartile). This interaction was not significant for data from experiments in which materials had been manipulated between subjects (Fig. 4b), but there too there was a significant main effect of quartile on bias for paintings and an accompanying pattern of increasing bias across the test.

The results for c show that as the recognition test proceeded, participants (on average) became increasingly more likely to miss a studied painting than to falsely endorse a nonstudied painting, whereas for word stimuli the ratio of these two error rates remained relatively stable across the test. It is important to emphasize that although we have often used conventional "conservative" and "liberal" terminology for referring to response biases throughout this paper, we do not mean to suggest these differences necessarily arise from shifts in participants' underlying decision criteria. We use "conservative" and "liberal" here as convenient, commonly understood descriptors of positive or negative response bias values

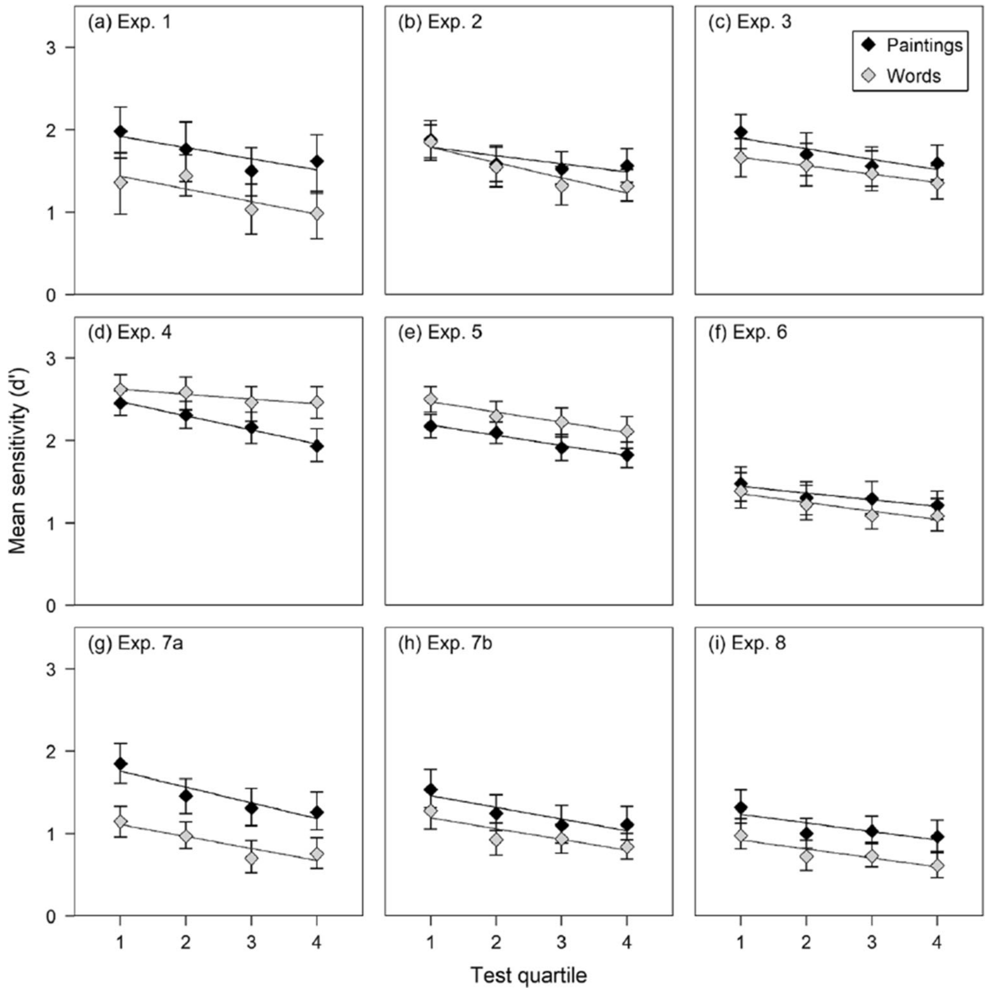

Fig. 2 Mean recognition memory sensitivity (d') by test quartile and accompanying regression lines for painting and word stimuli in seven experiments in which item type was manipulated within-subjects (**a–g**) and two with a between-subjects manipulation (**h–i**). Error bars are 95% BCa bootstrap confidence intervals (Efron, 1987)

(or to describe relative differences in these values across quartiles or materials). But we have no firm evidence that, for example, participants were *being* more conservative in responding to paintings in the sense of evaluating or accumulating evidence of oldness more strictly or reluctantly. The ease of making questionable theoretical leaps in interpreting response bias measures is one reason we think examining the underlying raw response rates is important; possible alternative explanations sometimes become more apparent when these rates are considered.

Analyses of underlying response rates showed that hit rates tended to decrease over the course of the test for both materials types. This decline appeared to be steeper for paintings than words in most within-subjects experiments (Fig. 3a–g), and mega-analyses of standardized HRs showed a significant interaction consistent with this impression (Fig. 5a). The Materials × Quartile interaction was also significant for FARs in the within-subjects data. Here, there was a significant effect of quartile on standardized FARs for words, which tended to increase from the first to final quartile (Fig. 5c). Paintings showed no such pattern. This interaction was not significant in the between-subjects data.

Theoretical implications

The pattern of materials-based differences in how responding changed over the course of the recognition test is consistent with a number of theoretical interpretations. We will discuss a few of these possibilities and some experimental and modeling approaches that may prove fruitful in future efforts to adjudicate among them.

Before we discuss some of the potential theoretical interpretations of our findings, it may be worth explicitly noting some of the differences between the words and paintings used in these experiments. One category of differences could be grouped under the umbrella of "complexity" or "distinctiveness." These intuitively appealing concepts can be difficult to define and are inconsistently operationalized (see Hunt, 2006,

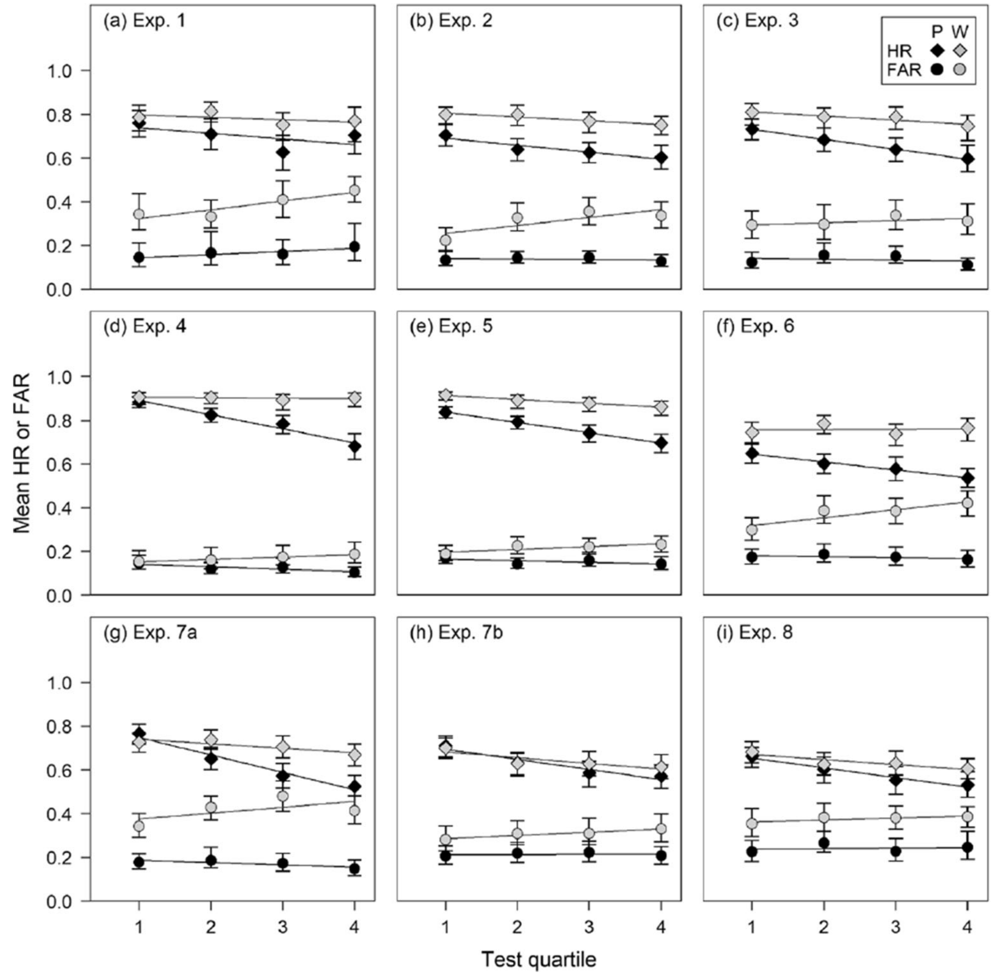

Fig. 3 Mean hit rates (HR) and false alarm rates (FAR) by test quartile and accompanying regression lines for painting (P) and word (W) stimuli in seven experiments in which item type was manipulated within-subjects (**a–g**) and two with a between-subjects manipulation (**h–i**). Error bars are 95% BCa bootstrap confidence intervals (Efron, 1987)

for insightful coverage of the notion of distinctiveness in memory research), but nonetheless provide a useful framework for thinking about some ways memory and decision-making processes might differ across materials. The paintings are inarguably more perceptually complex and diverse than the words, which were all presented in the same plain black text and comprised an inherently limited number of visual features. This kind of complexity has documented effects on recognition performance. For example, making words more perceptually complex by presenting them in varied fonts and colours can eliminate or even reverse the usual picture superiority effect (Ensor et al., 2019).

Whether the paintings or word stimuli we used are more *conceptually* complex or distinctive is more debatable, but here too there are differences that seem likely to have memory implications. Each word represents a discrete, known concept (although some of course have multiple meanings). The paintings more often included multiple concepts and themes—for example, an individual image might include people, trees, and water—but there was substantial overlap across the set (e.g., several paintings featured people, trees, and water). There are more dimensions on which the paintings can differ from each other than the words, and more opportunities for a particular striking feature to stand out at study or test. In this sense the paintings can be thought of as more distinctive. But one could equally argue that the words are more distinctive by virtue of being known entities that map onto existing memory representations.

The distinctiveness heuristic, whereby people are thought to demand more or qualitatively different evidence to endorse an item as "studied" when it belongs to a more distinctive category (Schacter et al., 1999), is worth considering in this context. Dobbins and Kroll (2005), for example, observed a mirror effect whereby photos of familiar locations were more often correctly recognized and rejected than photos of unfamiliar locations, and attributed this to the higher conceptual distinctiveness of well-known scenes leading participants to demand more evidence at test. Perhaps subjects in our experiments tended to view the paintings as more distinctive and hence more memorable than the words and consequently

Table 3 Results of mega-analyses for (z-transformed) hit rates and false alarm rates

Source		df_R	df_E	SS_R	SS_E	F	p	η^2_G
Within subjects								
Hit rates								
2 × 4 RM	Materials	1	337	4.309	780.233	1.861	.173	0.001
	Quartile	2.91	980.55	206.620	983.426	70.804	<.0001[a]	0.045
	Interaction	2.90	977.42	72.637	940.258	26.034	<.0001[a]	0.016
One way	Paintings	2.90	978.07	258.283	1142.273	76.200	<.0001[a]	0.098
	Words	2.94	991.83	20.974	781.411	9.045	<.0001[a]	0.010
False alarm rates								
2 × 4 RM	Materials	1	337	4.693	1014.550	1.559	.213	0.001
	Quartile	3	1011	15.909	870.580	6.158	<.001	0.004
	Interaction	2.96	995.88	34.335	767.944	15.067	< .0001[a]	0.009
One way	Paintings	2.94	990.81	9.179	891.928	3.468	.016[a]	0.005
	Words	3	1011	41.065	746.597	18.536	< .0001	0.020
Between subjects								
Hit rates								
2 × 4 Mixed	Materials	1	143	0.001	562.123	0.0002	.990	<.001
	Quartile	3	429	49.972	246.311	29.012	<.0001	0.058
	Interaction	3	429	4.107	246.311	2.384	.069	0.005
One way	Paintings	3	216	38.774	125.678	22.213	<.0001	0.086
	Words	3	213	15.466	120.633	9.102	<.0001	0.037
False alarm rates								
2 × 4 Mixed	Materials	1	143	0.004	549.001	0.001	.975	<.001
	Quartile	2.87	411.12	3.382	287.451	1.683	.172[a]	0.004
	Interaction	2.87	411.12	0.997	287.451	0.496	.677[a]	0.001
One way	Paintings	3	216	1.604	136.097	0.849	.469	0.004
	Words	2.74	194.30	2.767	151.354	1.298	.277[a]	0.006

[a] *p* (and associated *dfs*) corrected for significant sphericity violation

expected more evidence of oldness before endorsing paintings as studied. That is, maybe at least some subjects have an intuitive and exaggerated expectation of a picture superiority effect or of how well they will remember the paintings in general.

Lindsay et al. (2015) reported studies designed to test this possibility that failed to yield support for it, but in our current view those findings are far from definitive. Participants did, on average, report expecting that they would remember the paintings better than the words, but there was no consistent correlation between these self-reported estimates and subsequent response bias. There are limitations to this correlational approach, however, and memorability judgments made after the study phase may not adequately capture what is going on throughout the test (e.g., Guttentag & Carroll, 1998) We have thus not ruled out the possibility that distinctiveness-driven differences in memory expectations may play a role in these materials-based effects.

Another difference between the paintings and words in these experiments is their preexperimental familiarity. Participants had encountered the words in these experiments many times, whereas they have likely never seen the paintings. The greater familiarity of the words may have contributed to a sense of oldness, leading to relatively higher hit and FA rates for words than for paintings (although this account works less well as an account of conservative bias on paintings tested in the between-subjects design). In addition to familiarity, there are also more qualitative differences between stimuli that are well known prior to the experiment and those that are completely novel. The words in these experiments are already meaningful to participants; they had existing representations in memory, and were embedded in a web of episodic and semantic associations that may come to mind involuntarily and/or be deliberately recruited to facilitate encoding or retrieval. Granted, a painting encountered for the first time may nonetheless bring existing memories to mind, but not a thoroughly consolidated trace or network of the sort that exists for known words.

This difference between words and paintings in preexperimental exposure also means that the kind of judgment that must be made on the recognition test is somewhat distinct for the two item types. For words, an accurate old/new decision requires a source monitoring judgment (in other words, the question is "did I see this word in the specific

Table 4 Results of mega-analyses for (z-transformed) sensitivity and response bias

Source		df_R	df_E	SS_R	SS_E	F	p	η_G^2
Within subjects Sensitivity (d')								
2 × 4 RM	Materials	1	337	2.555	749.745	1.148	.285	.001
	Quartile	3	1011	132.767	726.501	61.586	<.0001	.038
	Interaction	3	1011	0.995	705.404	0.476	.699	<.001
One way	Paintings	2.95	995.16	72.541	729.217	33.524	<.0001[a]	.042
	Words	3	1011	61.221	702.689	29.361	<.0001	.035
Response Bias (c)								
2 × 4 RM	Materials	1	337	27.004	747.392	12.176	.001	.007
	Quartile	3	1011	61.511	885.173	23.418	<.0001	.016
	Interaction	3	1011	104.151	797.099	44.033	<.0001	.027
One way	Paintings	3	1011	158.301	1018.879	52.359	<.0001	.073
	Words	3	1011	7.361	663.393	3.740	.011	.004
Between subjects Sensitivity (d')								
2 × 4 mixed	Materials	1	143	0.132	611.192	0.031	.861	<.001
	Quartile	3	429	65.337	368.639	25.345	<.0001	.063
	Interaction	3	429	1.129	368.639	0.438	.726	.001
One way	Paintings	2.95	995.16	72.541	729.217	33.524	<.0001[a]	.042
	Words	3	1011	61.221	702.689	29.361	<.0001	.035
Response Bias (c)								
2 × 4 RM	Materials	1	337	27.004	747.392	12.176	.001	.007
	Quartile	3	1011	61.511	885.173	23.418	<.0001	.016
	Interaction	3	1011	104.151	797.099	44.033	<.0001	.027
One way	Paintings	3	1011	158.301	1,018.879	52.359	<.0001	.073
	Words	3	1011	7.361	663.393	3.740	.011	.004
Between subjects Sensitivity (d')								
2 × 4 mixed	Materials	1	143	0.132	611.192	0.031	.861	<.001
	Quartile	3	429	65.337	368.639	25.345	<.0001	.063
	Interaction	3	429	1.129	368.639	0.438	.726	.001
One way	Paintings	3	216	26.987	158.852	12.232	<.0001	.057
	Words	3	213	39.394	209.787	13.332	<.0001	.069
Response Bias (c)								
2 × 4 mixed	Materials	1	143	0.010	602.948			
	Quartile	2.80	399.78	9.663	287.365	4.808	0.003[a]	.011
	Interaction	2.80	399.78	4.668	287.365	2.323	.079a	.005
One way	Paintings	3	216	13.308	131.166	7.305	0.0001	.030
	Words	2.62	185.86	1.106	156.199	0.503	0.656[a]	.002

[a] p (and associated dfs) corrected for significant sphericity violation

context of this experiment?"). For paintings, this contextual element is in theory less important; participants need only judge whether they have *ever* seen a particular painting before.

The fact that most of our subjects had never before seen most or all of the paintings may play a role in the differences in how hit and FA rates changed over the course of the test for painting versus words. But we do not think the novelty of the paintings is sufficient as an explanation, because it is not generally the case that novel stimuli produce conservative recognition memory response bias. Indeed, the "pseudoword effect" refers to the observation that hit and FA rates tend to be *higher* for pseudowords (e.g., *hension*, *framble*) than for real words (*tension*, *bramble*; Greene, 2004). Also, we recently found (as yet unpublished) that response bias was neutral on average for

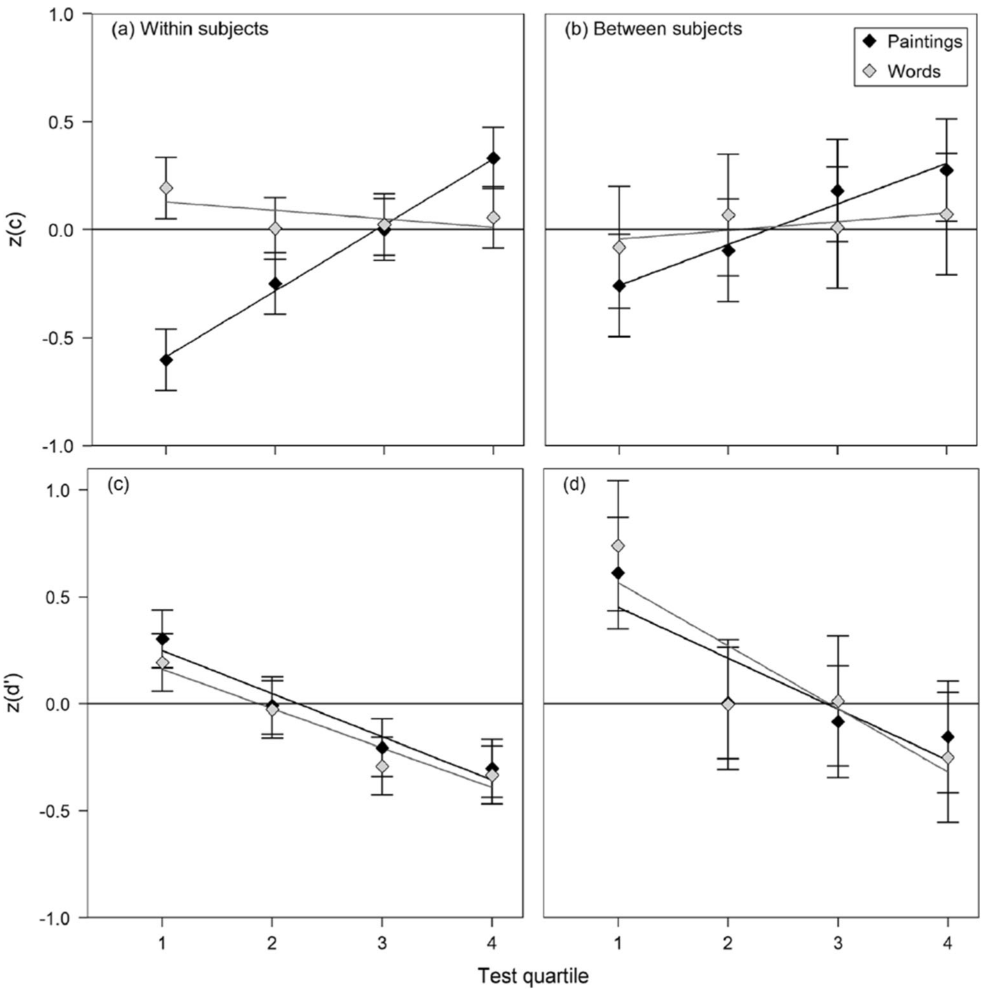

Fig. 4 Means by test quartile (and accompanying regression lines) of *z*-transformed response bias (*c*) and sensitivity (*d′*) scores for painting and word stimuli across seven samples in which materials were manipulated within subjects (**a** and **c**) and two in which materials were manipulated between subjects (**b** and **d**). See text for details on these calculations. Error bars are 99% within-subjects confidence intervals (Loftus & Masson, 1994).

both the words and line drawings in the Snodgrass and Vanderwart (1980) stimulus set using a recognition memory procedure that was otherwise the same as Experiment 7a in this paper. Like the paintings, these line drawings were new to participants (although they depicted familiar objects). So, novelty by itself, at least in this straightforward sense, seems unlikely to account for the MBBE.

Nonetheless, we think it is entirely possible that complexity, distinctiveness, and preexperimental familiarity are all important to understanding the materials-based differences we have observed, perhaps to varying degrees across participants, items, and test trials. With respect to the central findings in this paper, a key question for any of these variables is *how* a materials-level variable that does not itself change over the course of the test could account for these kinds of interactions between materials and test position. One possibility is that such participants' subjective experience related to one or more of these variables might change over the course of the recognition test in ways that affect response bias (e.g., one materials type might seem more or less distinctive or memorable as the test goes on, leading some participants to adjust their decision criteria). Because we want to emphasize that response bias is not *necessarily* involved, however, it is worth considering some more general processes/variables that are known or hypothesized to change over the course of a recognition test.

The across-quartile patterns reported here for both words and paintings fit nicely with recent work on test position effects on recognition of various kinds of stimuli. The typical finding is that hit rates decline over the course of the test, while the pattern for false alarm rates is much more variable: they may increase, decrease, or remain stable (e.g., Criss et al., 2011; Fox et al., 2020; Osth et al., 2018). There is evidence to suggest these effects arise from a complex interplay among multiple mechanisms, with stimulus-, participant-, and experiment-level factors potentially influencing the relative contributions of each. For example, recent work supports a role of both context drift

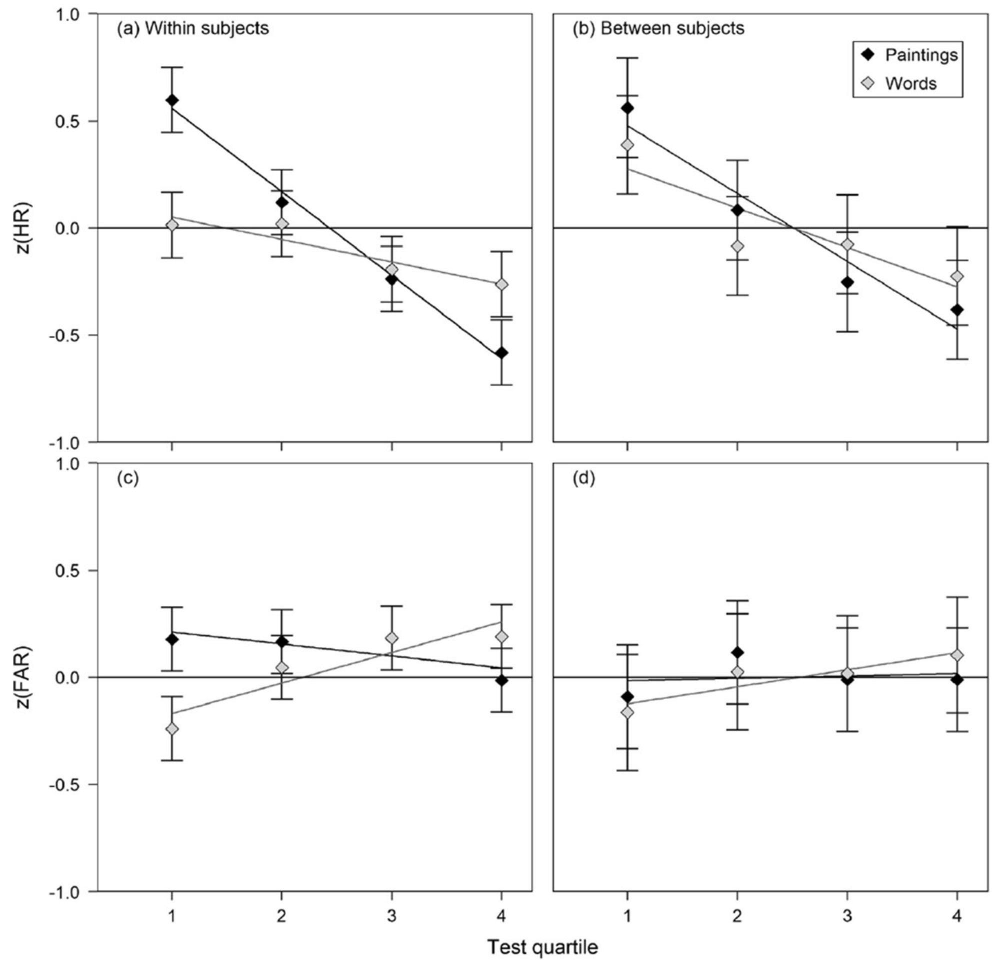

Fig. 5 Means by test quartile (and accompanying regression lines) of *z*-transformed hit rates (HR) and false alarm rates (FAR) for painting and word stimuli across seven samples in which materials were manipulated within subjects (**a** and **c**) and two in which materials were manipulated between subjects (**b** and **d**). See text for details on these calculations. Error bars are 99% within-subjects confidence intervals (Loftus & Masson, 1994)

(Osth et al., 2018; Osth & Dennis, 2015) and item noise (Fox et al., 2020) in test position effects. Context drift as a statistical concept was introduced by Estes (1955). In the context of recognition memory, it can be thought of as the tendency for contextual elements of the memory probe—such as the various kinds of cognitive processing participants are engaged in—to drift farther from those associated with the study episode as the recognition test proceeds. Item noise refers to interference produced by other items encountered in the experimental context (Gillund & Shiffrin, 1984) (by contrast with context noise, which is interference from other contexts in which the current item has been encountered and therefore of particular relevance for words; e.g., Dennis & Humphreys, 2001).

There is some evidence for materials-based differences in the relative roles of these various kinds of interference (Osth et al., 2014; Osth & Dennis, 2015). It is possible that the paintings and words in the experiments are differentially susceptible to various forms of noise such that across-quartile (and overall) response patterns tend to differ between materials. Interference-based mechanisms *could* produce patterns like the ones we have observed without any real change in the decision criterion or the way evidence is evaluated over the course of the test, but it is equally plausible that both kinds of mechanisms play a role. These processes may also interact in complex ways that produce somewhat counterintuitive effects on response rates and other measures (Osth et al., 2018), such that formal modeling will be required to understand the relative contributions of various factors.

The utility of our data in discriminating among some of these potential mechanisms is constrained by a few aspects of the experimental design. Item order within the study and test lists was fully randomized in all of the experiments described here, and as others have pointed out in the context of list length effects (Dennis & Humphreys, 2001; Kinnell & Dennis, 2011), several variables that might change over the course of the test are inherently confounded in such designs. The average retention interval and number of intervening items increase from the first to final test quartile such that decay, contextual drift, and item noise would all be more likely toward the end of the test. Participants may also become fatigued or less attentive as the test proceeds. Other researchers have attempted to disentangle some of these confounds in various ways, such as comparing randomized designs with those in which items are tested in the same or opposite order in which they were studied—minimizing and maximizing, respectively, the extent to which the retention interval/number of intervening item varies across quartiles—and with partially randomized blocked designs that preserve local stimulus context (Averell et al., 2016; Criss et al., 2011). Comparing the results of designs like these with the current results may help narrow down the possible sources of the materials-based differences we have observed.

Alternative analytic approaches may also prove informative in homing in on the theoretical sources of these materials-based bias differences. In drift diffusion models, for example, response bias differences can arise from two parameters with theoretically distinct implications. The usefulness of more complex models in the context of these quartile-level analyses, however, is seriously constrained by the low numbers of trials per cell[8] and by the prevalence of false alarm rates that are near or at floor. This issue also imposes a more general limitation on the conclusions we can draw from these data. Although there was clearly a materials-based difference in how FARs changed over the course of the test (at least in the within-subjects experiments), the low overall frequency of FARs in all quartiles means that for paintings, we can only reasonably rule out the possibility of a systematic across-quartile *increase* in the FAR; a true decrease in this rate might be masked by floor effects. This is particularly relevant to attempts to understand the mechanisms underlying the effects on c, as the FAR is in theory less “contaminated” than the HR by memory effects (e.g., encoding variability) and thus proportionally more sensitive to decisional effects. If participants truly adopt a more conservative decision criterion for paintings over the course of the test, a stable FAR would imply there must also be some mechanism acting on FARs in the opposite direction. As discussed above with reference to various sources of interference, this is entirely plausible, but without stronger evidence against an across-quartile decrease in FARs it is unclear whether such a line of inquiry would be useful. Future efforts might attempt to boost the overall FAR by changing test conditions (e.g., imposing response deadlines) or by changing the stimulus set (e.g., using paintings of a similar style or by only a few artists).[9]

Our results illustrate the importance of using diverse materials to study recognition memory. A great deal of recognition memory research is conducted using verbal materials. There are many advantages to using verbal stimuli, but results obtained with words do not always generalize to other stimulus types (Kinnell & Dennis, 2012; Mulligan, 2013; Osth et al., 2014), and it is important to understand why and under what conditions such materials-based differences arise. Our work also highlights the value of studying response bias. Most previous research on materials-based differences in memory has centered on hit rates or accuracy, but our results demonstrate that limiting comparisons to such measures risks missing potentially informative materials-based response bias effects. As others have emphasized, these effects need not be of specific interest to the researcher to have implications for their results; unknown (or unaccounted for) response bias effects can substantially compromise the validity of inferences based on sensitivity or other accuracy measures (Donaldson, 1993; Grider & Malmberg, 2008; Rotello et al., 2008; Verde & Rotello, 2007; Wiens et al., 1997).

Reassuringly, it is our impression that it has become increasingly common to see some measure of response bias in reports of old/new recognition memory results, even when memory accuracy is of central interest. Although reporting both types of measures provides a more complete picture of the data than either alone, results must still be critically considered with reference to the nature of the data and the model being assumed. With respect to c and d' specifically, the sole advantage of these model-based measures over simpler ones (e.g., raw HRs & FARs, percent accurate)—namely, the ability to separate the contributions of bias and sensitivity to observed responses—only holds under a constrained set of assumptions about the underlying evidence strength distributions. ROC analyses suggest violations of the assumption of equal variance of the “old” and “new” item distributions are the rule, not the exception, in recognition memory data (Mickes et al., 2007; Ratcliff et al., 1992; Yonelinas & Parks, 2007), and numerous authors have demonstrated how misleading inferences based on d' and c can be under such conditions (Dougal & Rotello, 2007; Grider & Malmberg, 2008; S. Rhodes et al., 2018; Verde et al., 2006).

We were particularly concerned that materials-based differences in the extent to which data deviated from equal variance might compromise the results for c and/or d'. To balance competing concerns that group-level ROCs might distort informative differences across individuals (Malejka & Bröder, 2019) or quartiles, whereas ROCs based on fewer observations are more often impossible to fit or difficult to interpret given variation in how individuals use the response scale, we constructed confidence-ratings-based ROCs for each materials type and experiment in three ways: at the group level (collapsing across trials and participants), the quartile level (collapsing across participants within each test quartile), and the participant level (collapsing across trials). All ROCs were fit in R using the

[8] As Caren Rotello pointed out (personal communication, November 6, 2019), this also limits the set of possible hit and false alarm rates, and by extension the possible values of c and d'.

[9] Lindsay and Kantner (2011) Experiment 2 provided some evidence that FARs may be increased by using a set of painting stimuli consisting only of portraits, but bias was nonetheless conservative with that set. Lindsay and Kantner also observed conservative response bias in recognition of snippets of poetry and of Korean folk music.

method described by Vokey (2016); more detail regarding these analyses is included in the supplemental material. We used the resulting zROC slopes to calculate alternative measures of sensitivity (d_a) and response bias (c_a) based on Grider and Malmberg's (2008) demonstration that these measures performed better than other common SDT-based measures under conditions of unequal variance, being less susceptible to variation in zROC slope. These results all looked similar to those in Figs. 1 and 2 with respect to the patterns of central interest (see supplemental material), strengthening our confidence that these results are not spurious artefacts of inappropriate assumptions.

It should be noted that ROCs based on confidence ratings rest on their own controversial assumptions about how participants map their internal states onto discrete responses (Bröder et al., 2013; Bröder & Schütz, 2009; Malmberg, 2002), but Dube and Rotello (2012) argued that much of this concern is unfounded, showing that ROC parameters were generally similar for ratings-based ROCs and those constructed on the basis of bias manipulations. Nonetheless, rather than uncritical reliance on any particular measure of bias or sensitivity—which seems inadvisable given the current state of understanding of recognition memory—we encourage more extensive reporting of data and the analytic process in general. This can ensure data go toward advancing knowledge even if the analyses or conclusions of central interest in the original report prove inappropriate or limited.

We have extended our previous findings of materials-based differences in recognition memory response bias by establishing that the extent of these differences varies as a function of test position. These results point to test length as one possible boundary condition of what has thus far been a robust bias difference between word and painting stimuli, suggesting several avenues for future exploration of the mechanisms underlying these differences. Accompanying ROCs and analyses of underlying hit and false alarm rates illustrate some of the quantitative and interpretive challenges associated with our data and with SDT-based analyses of recognition memory more generally, but also further emphasize the potential for stimulus materials to influence recognition memory in ways that are obscured by focusing primarily on test-level sensitivity measures. The use of diverse stimulus materials and analytic approaches, more careful consideration of dependent measures and their underlying assumptions, and greater attention to response bias—both in terms of its implications for other measures, and as a variable of interest in its own right—can all contribute to a more nuanced understanding of recognition memory and the associated decision-making processes.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13421-021-01227-5.

Author note This research was supported by the Natural Sciences and Engineering Research Council of Canada through a Discovery Grant to D.S.L. and Canada Graduate Scholarships (master's & doctoral) to K.M.F. Justin Kantner developed this line of work (with D.S.L.), coordinated data collection for Experiments 1–7, and was instrumental in efforts to organize data and other research materials for further use. We thank Justin Kantner, Caren M. Rotello, and Michael E. J. Masson for feedback on draft versions of this manuscript, and Jamie-Lee Barden for assistance with proofreading and formatting. We also acknowledge the contributions of a number of former research assistants and honours students to the original data collection process for these studies, including Graeme Austin, Emily Cameron, Jordanna Freeman, Tiffany Hannan, Sarah Kraeutner, Caitlin Malli, Mayumi Okamoto, Priya Rosenberg, and Joe Savidge.

References

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

Aßfalg, A., Bernstein, D. M., & Hockley, W. (2017). The revelation effect: A meta-analytic test of hypotheses. *Psychonomic Bulletin & Review* https://doi.org/10.3758/s13423-017-1227-6

Averell, L., Prince, M., & Heathcote, A. (2016). Fundamental causes of systematic and random variability in recognition memory. *Journal of Memory and Language*, *88*, 51–69. https://doi.org/10.1016/j.jml.2015.12.010

Azimian-Faridani, N., & Wilding, E. L. (2006). The influence of criterion shifts on electrophysiological correlates of recognition memory. *Journal of Cognitive Neuroscience*, *18*(7), 1075–1086. https://doi.org/10.1162/jocn.2006.18.7.1075

Benjamin, A. S. (2013). Where is the criterion noise in recognition? (Almost) everyplace you look: Comment on Kellen, Klauer, and Singmann (2012). *Psychological Review*, *120*(3), 720–726. https://doi.org/10.1037/a0031911

Benjamin, A. S., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language*, *51*(2), 159–172. https://doi.org/10.1016/j.jml.2004.04.001

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review*, *116*(1), 84–115. https://doi.org/10.1037/a0014351

Berch, D. B., & Evans, R. C. (1973). Decision processes in children's recognition memory. *Journal of Experimental Child Psychology*, *16*(1), 148–164. https://doi.org/10.1016/0022-0965(73)90069-6

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

Bröder, A., Kellen, D., Schütz, J., & Rohrmeier, C. (2013). Validating a two-high-threshold measurement model for confidence rating data in recognition. *Memory*, *21*(8), 916–944. https://doi.org/10.1080/09658211.2013.767348

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear—Or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 587–606. https://doi.org/10.1037/a0015279

Brown, S., Steyvers, M., & Hemmer, P. (2007). Modeling experimentally induced strategy shifts: Research report. *Psychological Science*, *18*(1), 40–45. https://doi.org/10.1111/j.1467-9280.2007.01846.x

Canty, A., & Ripley, B. D. (2020). *boot: Bootstrap R (S-Plus) Functions*. https://cran.r-project.org/web/packages/boot/citation.html

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497–505. https://doi.org/10.1080/14640748108400805

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods*, *14*(2), 165–176. https://doi.org/10.1037/a0015565

Cox, G. E., & Shiffrin, R. M. (2012). Criterion setting and the dynamics of recognition memory. *Topics in Cognitive Science*, *4*(1), 135–150. https://doi.org/10.1111/j.1756-8765.2011.01177.x

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

Criss, A. H. (2009). The distribution of subjective memory strength: List strength and response bias. *Cognitive Psychology*, *59*(4), 297–319. https://doi.org/10.1016/j.cogpsych.2009.07.003

Criss, A. H., Aue, W. R., & Kılıç, A. (2014). Age and response bias: Evidence from the strength-based mirror effect. *Quarterly Journal of Experimental Psychology*, *67*(10), 1910–1924. https://doi.org/10.1080/17470218.2013.874037

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

Curran, T., DeBuse, C., & Leynes, P. A. (2007). Conflict and criterion setting in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(1), 2–17. https://doi.org/10.1037/0278-7393.33.1.2

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their applications*. Cambridge: Cambridge University Press. http://statwww.epfl.ch/davison/BMA/

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. https://doi.org/10.1037/a0026330

Defeyter, M. A., Russo, R., & McPartlin, P. L. (2009). The picture superiority effect in recognition memory: A developmental study using the response signal procedure. *Cognitive Development*, *24*(3), 265–273. https://doi.org/10.1016/j.cogdev.2009.05.002

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, *108*(2), 452–478. https://doi.org/10.1037/0033-295X.108.2.452

Dobbins, I. G., & Kroll, N. E. A. (2005). Distinctiveness and the recognition mirror effect: Evidence for an item-based criterion placement heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1186–1198. https://doi.org/10.1037/0278-7393.31.6.1186

Donaldson, W. (1993). Accuracy of d' and A' as estimates of sensitivity. *Bulletin of the Psychonomic Society*, *31*(4), 271–274. https://doi.org/10.3758/BF03334926

Donaldson, W., & Murdock, B. B. (1968). Criterion change in continuous recognition memory. *Journal of Experimental Psychology*, *76*(3), 325–330. https://doi.org/10.1037/h0025510

Dopkins, S., Sargent, J., & Ngo, C. T. (2010). The bias for a recognition judgement depends on the response emitted in a prior recognition judgement. *Memory*, *18*(3), 272–283. https://doi.org/10.1080/09658211003601506

Dougal, S., & Rotello, C. M. (2007). "Remembering" emotional words is based on response bias, not recollection. *Psychonomic Bulletin & Review*, *14*(3), 423–429. https://doi.org/10.3758/BF03194083

Dube, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning Memory & Cognition*, *38*(1), 130–151. https://doi.org/10.1037/a0024957

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185. https://doi.org/10.1080/01621459.1987.10478410

Ensor, T. M., Surprenant, A. M., & Neath, I. (2019). Increasing word distinctiveness eliminates the picture superiority effect in recognition: Evidence for the physical-distinctiveness account. *Memory & Cognition*, *47*(1), 182–193. https://doi.org/10.3758/s13421-018-0858-9

Erdelyi, M. H., Finks, J., & Feigin-Pfau, M. B. (1989). The effect of response bias on recall performance, with some observations on processing bias. *Journal of Experimental PsychologyGeneral*, *118*(3), 245–254. https://doi.org/10.1037/0096-3445.118.3.245

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review*, *62*(3), 145–154. https://doi.org/10.1037/h0048509

Fawcett, J. M., Quinlan, C. K., & Taylor, T. L. (2012). Interplay of the production and picture superiority effects: A signal detection analysis. *Memory*, *20*(7), 655–666. https://doi.org/10.1080/09658211.2012.693510

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. https://doi.org/10.3758/BF03198484

Fox, J., Dennis, S., & Osth, A. F. (2020). Accounting for the build-up of proactive interference across lists in a list length paradigm reveals a dominance of item-noise in recognition memory. *Journal of Memory and Language*, *110*, Article 104065. https://doi.org/10.1016/j.jml.2019.104065

Frithsen, A., Kantner, J., Lopez, B. A., & Miller, M. B. (2018). Cross-task and cross-manipulation stability in shifting the decision criterion. *Memory*, *26*(5), 653–663. https://doi.org/10.1080/09658211.2017.1393090

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. https://doi.org/10.3758/BF03213172

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*(1), 1–67. https://doi.org/10.1037/0033-295X.91.1.1

Goh, W. D. (2005). Talker variability and recognition memory: Instance-specific and voice-specific effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 40–53. https://doi.org/10.1037/0278-7393.31.1.40

Greene, R. L. (2004). Recognition memory for pseudowords. *Journal of Memory and Language*, *50*(3), 259–267. https://doi.org/10.1016/j.jml.2003.12.001

Grider, R. C., & Malmberg, K. J. (2008). Discriminating between changes in bias and changes in accuracy for recognition memory of emotional stimuli. *Memory & Cognition*, *36*(5), 933–946. https://doi.org/10.3758/MC.36.5.933

Guttentag, R., & Carroll, D. (1998). Memorability judgments for high- and low-frequency words. *Memory & Cognition*, *26*(5), 951–958. https://doi.org/10.3758/BF03201175

Han, S., & Dobbins, I. G. (2008). Examining recognition criterion rigidity during testing using a biased-feedback technique: Evidence for adaptive criterion learning. *Memory & Cognition*, *36*(4), 703–715. https://doi.org/10.3758/MC.36.4.703

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

Heit, E., Brockdorff, N., & Lamberts, K. (2003). Adaptive changes of response criterion in recognition memory. *Psychonomic Bulletin & Review*, *10*(3), 718–723. https://doi.org/10.3758/BF03196537

Hilford, A., Glanzer, M., Kim, K., & Maloney, L. T. (2019). One mirror effect: The regularities of recognition memory. *Memory &*

Cognition, *47*(2), 266–278. https://doi.org/10.3758/s13421-018-0864-y
Hintzman, D. L. (2011). Research strategy in the study of memory: Fads, fallacies, and the search for the "Coordinates of Truth". *Perspectives on Psychological Science*, *6*(3), 253–271. https://doi.org/10.1177/1745691611406924
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. https://doi.org/10.1037/0278-7393.21.2.302
Hockley, W. E. (2011). Criterion changes: How flexible are recognition decision processes? In P. A. Higham & J. P. Leboe (Eds.), *Constructions of remembering and metacognition: Essays in honor of Bruce Whittlesea* (pp. 155–166). Palgrave Macmillan UK. https://doi.org/10.1057/9780230305281
Hockley, W. E., & Niewiadomski, M. W. (2007). Strength-based mirror effects in item and associative recognition: Evidence for within-list criterion changes. *Memory & Cognition*, *35*(4), 679–688. https://doi.org/10.3758/BF03193306
Hunt, R. R. (2006). *Distinctiveness and memory*. https://doi.org/10.1093/acprof:oso/9780195169669.001.0001
JASP Team. (2019). JASP (Version 0.10.2). [Computer Software]. https://jasp-stats.org/2019/07/22/announcing-jasp-0-10-2/
Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition*, *40*(8), 1163–1177. https://doi.org/10.3758/s13421-012-0226-0
Kantner, J., & Lindsay, D. S. (2014). Cross-situational consistency in recognition memory response bias. *Psychonomic Bulletin & Review*, *21*(5), 1272–1280. https://doi.org/10.3758/s13423-014-0608-3
Kent, C., Lamberts, K., & Patton, R. (2018). Cue quality and criterion setting in recognition memory. *Memory & Cognition*, *46*(5), 757–769. https://doi.org/10.3758/s13421-018-0796-6
Kinnell, A., & Dennis, S. (2011). The list length effect in recognition memory: An analysis of potential confounds. *Memory & Cognition*, *39*(2), 348–363. https://doi.org/10.3758/s13421-010-0007-6
Kinnell, A., & Dennis, S. (2012). The role of stimulus type in list length effects in recognition memory. *Memory & Cognition*, *40*(3), 311–325. https://doi.org/10.3758/s13421-011-0164-2
Koop, G. J., Criss, A. H., & Pardini, A. M. (2019). A strength-based mirror effect persists even when criterion shifts are unlikely. *Memory & Cognition*, 842–854. https://doi.org/10.3758/s13421-019-00906-8
Lawrence, M. A. (2016). *ez: Easy Analysis and Visualization of Factorial Experiments*. https://cran.r-project.org/package=ez
Ley, R., & Long, K. (1987). A distractor-free test of recognition and false recognition. *Bulletin of the Psychonomic Society*, *25*(6), 411–414. https://doi.org/10.3758/BF03334727
Lindsay, D. S., & Kantner, J. (2011). A search for influences of feedback on recognition of music, poetry, and art. In P. A. Higham & J. P. Leboe (Eds.), *Constructions of remembering and metacognition: Essays in honor of Bruce Whittlesea* (pp. 137–154). Palgrave Macmillan UK. https://doi.org/10.1057/9780230305281_11
Lindsay, D. S., Kantner, J., & Fallow, K. M. (2015). Recognition memory response bias is conservative for paintings and we don't know why. In D. S. Lindsay (Ed.) & C. M. Kelley (Trans.) & A. P. Yonelinas, H. L. Roediger II (Eds.), *Remembering: Attributions, processes, and control in human memory: Essays in honor of Larry Jacoby* (pp. 213–229). Psychology Press.
Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*(4), 476–490. https://doi.org/10.3758/BF03210951
Lukavský, J., & Děchtěrenko, F. (2017). Visual properties and memorising scenes: Effects of image-space sparseness and uniformity. *Attention, Perception, & Psychophysics*. https://doi.org/10.3758/s13414-017-1375-9
Macken, W. J. (2002). Environmental context and recognition: The role of recollection and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(1), 153–161. https://doi.org/10.1037//0278-7393.28.1.153
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. https://doi.org/10.1037/0033-2909.98.1.185
Madigan, S. (1983). Picture memory. In J. C. Yuille (Ed.), *Imagery, Memory, and Cognition: Essays in Honor of Allan Paivio* (pp. 65–89). Erlbaum.
Malejka, S., & Bröder, A. (2019). Exploring the shape of signal-detection distributions in individual recognition ROC data. *Journal of Memory and Language*, *104*(May 2017), 83–107. https://doi.org/10.1016/j.jml.2018.09.001
Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(2), 380–387. https://doi.org/10.1037/0278-7393.28.2.380
Malmberg, K. J., Criss, A. H., Gangwani, T. H., & Shiffrin, R. M. (2012). Overcoming the negative consequences of interference from recognition memory testing. *Psychological Science*, *23*(2), 115–119. https://doi.org/10.1177/0956797611430692
Marken, R. S., & Sandusky, A. J. (1974). Stimulus probability and sequential effect in recognition memory. *Bulletin of the Psychonomic Society*, *4*(1), 49–51.
Marquié, J. C., & Baracat, B. (2000). Effects of age, education, and sex on response bias in a recognition task. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *55*(5), P266–P272.
Megla, E. E., Woodman, G. F., & Maxcey, A. M. (2021). Induced forgetting Is the result of true forgetting, not shifts in decision-making thresholds. *Journal of Cognitive Neuroscience*, 1–13. https://doi.org/10.1162/jocn_a_01701
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. https://doi.org/10.3758/BF03194112
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. https://doi.org/10.1037/h0087363
Miller, M. B., & Kantner, J. (2020). Not all people are cut out for strategic criterion shifting. *Current Directions in Psychological Science*, *29*(1), 9–15. https://doi.org/10.1177/0963721419872747
Mulligan, N. W. (2013). Memory for pictures and actions. In T. J. Perfect & D. S. Lindsay (Eds.), *The SAGE handbook of applied memory* (pp. 20–36). SAGE Publications Ltd. https://doi.org/10.4135/9781446294703.n2
Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review*, *122*(2), 260–311. https://doi.org/10.1037/a0038692
Osth, A. F., Dennis, S., & Kinnell, A. (2014). Stimulus type and the list strength paradigm. *The Quarterly Journal of Experimental Psychology*, *67*(9), 1826–1841. https://doi.org/10.1080/17470218.2013.872824
Osth, A. F., Jansson, A., Dennis, S., & Heathcote, A. (2018). Modeling the dynamics of recognition memory testing with a combined model of retrieval and decision making. *Cognitive Psychology*, *104*(April), 106–142.
Paivio, A., Rogers, T. B., & Smythe, P. C. (1968). Why are pictures easier to recall than words? *Psychonomic Science*, *11*(4), 137–138. https://doi.org/10.3758/BF03331011

Postma, A. (1999). The influence of decision criteria upon remembering and knowing in recognition memory. *Acta Psychologica*, *103*, 65–76. https://doi.org/10.1016/S0001-6918(99)00032-3

Potter, M. C., Staub, A., Rado, J., & O'Connor, D. H. (2002). Recognition memory for briefly presented pictures: The time course of rapid forgetting. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(5), 1163–1175. https://doi.org/10.1037/0096-1523.28.5.1163

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

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

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

Rhodes, S., Cowan, N., Parra, M. A., & Logie, R. H. (2018). Interaction effects on common measures of sensitivity: Choice of measure, Type I error, and power. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-1081-0

Rotello, C. M., Macmillan, N. A., Hicks, J. L., & Hautus, M. J. (2006). Interpreting the effects of response bias on remember-know judgments using signal detection and threshold models. *Memory & Cognition*, *34*(8), 1598–1614. https://doi.org/10.3758/BF03195923

Rotello, C. M., Masson, M. E. J., & Verde, M. F. (2008). Type I error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics*, *70*(2), 389–401. https://doi.org/10.3758/PP

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*. https://doi.org/10.1016/j.jmp.2012.08.001

RStudio Team. (2016). *RStudio: Integrated Development Environment for R*. RStudio, Inc. http://www.rstudio.com/

Schacter, D. L., Israel, L., & Racine, C. (1999). Suppressing false recognition in younger and older adults: The distinctiveness heuristic. *Journal of Memory and Language*, *40*, 1–24.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002a). *E-Prime reference guide*. Psychology Software Tools, Inc.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002b). *E-Prime: User's guide*. Psychology Software Tools, Inc.

Singer, M., & Wixted, J. T. (2006). Effect of delay on recognition decisions: Evidence for a criterion shift. *Memory & Cognition*, *34*(1), 125–137. https://doi.org/10.3758/BF03193392

Singer, M. (2009). Strength-based criterion shifts in recognition memory. *Memory & Cognition, 37*(7), 976–984. https://doi.org/10.3758/MC.37.7.976

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning & Memory*, *6*(2), 174–215. https://doi.org/10.1037/0278-7393.6.2.174

Strack, F., & Forster, J. (1995). Reporting recollective experiences: Direct access to memory systems? *Psychological Science*, *6*(6), 352–358. https://doi.org/10.1111/j.1467-9280.1995.tb00525.x

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

Verde, M. F., MacMillan, N. A, & Rotello, C. M. (2006). Measures of sensitivity based on a single hit rate and false alarm rate: The accuracy, precision, and robustness of d', A_z, and A'. *Perception & Psychophysics*, *68*(4), 643–654. https://doi.org/10.3758/BF03208765

Verde, M. F., & Rotello, C. M. (2004). ROC curves show that the revelation effect is not a single phenomenon. *Psychonomic Bulletin & Review*, *11*(3), 560–566. https://doi.org/10.3758/BF03196611

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

Vokey, J. R. (2016). Single-step simple ROC curve fitting via PCA. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *70*(4), 301–305. https://doi.org/10.1037/cep0000095

Vokey, J. R., & Hockley, W. E. (2012). Unmasking a shady mirror effect: Recognition of normal versus obscured faces. *The Quarterly Journal of Experimental Psychology*, *65*(April 2014), 37–41. https://doi.org/10.1080/17470218.2011.628399

Wallace, W. P. (1978). Recognition failure of recallable words and recognizable words. *: Human Learning and Memory*. https://doi.org/10.1037/0278-7393.4.5.441

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

Wickham, H. (2017). *tidyverse: Easily Install and Load the "Tidyverse."* https://cran.r-project.org/package=tidyverse

Wiens, S., Emmerich, D. S., & Katkin, E. S. (1997). Response bias affects perceptual asymmetry scores and performance measures on a dichotic listening task. *Neuropsychologia*, *35*(11), 1475–1482. https://doi.org/10.1016/S0028-3932(97)00073-0

Witt, J. K., Taylor, J. E. T., Sugovic, M., & Wixted, J. T. (2015). Signal detection measures cannot distinguish perceptual biases from response biases. *Perception*, *44*(3), 289–300. https://doi.org/10.1068/p7908

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