Memory & Cognition (2018) 46:757–769
https://doi.org/10.3758/s13421-018-0796-6

Cue quality and criterion setting in recognition memory

Christopher Kent[1] · Koen Lamberts[2] · Richard Patton[1]

Published online: 2 February 2018

Abstract
Previous studies on how people set and modify decision criteria in old-new recognition tasks (in which they have to decide whether or not a stimulus was seen in a study phase) have almost exclusively focused on properties of the study items, such as presentation frequency or study list length. In contrast, in the three studies reported here, we manipulated the quality of the test cues in a scene-recognition task, either by degrading through Gaussian blurring (Experiment 1) or by limiting presentation duration (Experiment 2 and 3). In Experiments 1 and 2, degradation of the test cue led to worse old-new discrimination. Most importantly, however, participants were more liberal in their responses to degraded cues (i.e., more likely to call the cue "old"), demonstrating strong within-list, item-by-item, criterion shifts. This liberal response bias toward degraded stimuli came at the cost of increasing the false alarm rate while maintaining a constant hit rate. Experiment 3 replicated Experiment 2 with additional stimulus types (words and faces) but did not provide accuracy feedback to participants. The criterion shifts in Experiment 3 were smaller in magnitude than Experiments 1 and 2 and varied in consistency across stimulus type, suggesting, in line with previous studies, that feedback is important for participants to shift their criteria.

Keywords Recognition · Criterion setting · Mirror effect · Cue quality · Signal detection theory

Introduction

People often make old-new recognition decisions about stimuli that differ perceptually from the original visual experience. In addition to changes in viewpoint, occlusion or illumination, the stimulus material itself can be degraded in a number of ways. In this article, we report three experiments in which we studied the effects of stimulus degradation on old-new recognition judgments for visual scenes, words, and faces. In particular, we were interested in studying whether and how people adjust their decision criterion in response to a degraded test presentation. Criterion setting is an important area for understanding how people make recognition judgments and continues to provide a rich test bed for models of recognition memory (e.g., Cox & Shiffrin, 2012; Hicks & Starns, 2014; Starns, Ratcliff, & White, 2012; Starns, White, & Ratcliff, 2010, 2012). However, very few studies have looked at changes in criterion setting due to item specific properties at test, instead manipulating properties at study (e.g., item or list strength).

One study to look at the impact of test stimulus degradation on recognition performance is Wolfe and Kuzmova (2011), who demonstrated that a significant reduction in stimulus resolution (from 256 × 256 pixels at study to 32 × 32 pixels at test) still allowed for efficient old-new recognition, confirming results previously obtained by Uttl, Graf, and Siegenthaler (2007). Still, despite the robust nature of recognition, it is clear that there must be a level of test item degradation so severe that it leads to a significant decline in recognition performance, and we set out to explore what form that decline takes.

Many current accounts of recognition memory are based on some version of signal detection theory (SDT; see Malmberg, 2008). In such accounts, it is usually assumed that the stimulus generates a familiarity signal (corresponding to the value of a random variable with a particular distribution), and if this signal exceeds a criterion value, an "old" decision is made. Recognition models differ in their characterization of the familiarity variable (some models assume that there are other variables at play as well, e.g., Mandler, 1991), but the nature of the task lends itself exceptionally well to a characterization in terms of a signal-criterion comparison.

✉ Christopher Kent
c.kent@bristol.ac.uk

[1] School of Experimental Psychology, University of Bristol, 12a Priory Road, Bristol BS8 1TU, UK

[2] VCs Office, University of York, Heslington, York, UK

Within this framework, degradation of a test stimulus can have several possible effects. Degradation can lead to weaker familiarity signals for old test items, which would result in poorer discriminability of old and new test items. Test item degradation could also affect the variance of the signal distribution. Finally, degradation of the test stimulus could induce change in the criterion that underlies old-new decisions. It is well known that criterion setting can depend on various characteristics of stimulus items and on procedural variables (see Hockley, 2011, for a review). For example, it has been demonstrated that more memorable items are judged against a more conservative criterion than less memorable items (e.g., Hirshman, 1995), and that more liberal criteria are applied to delayed test items compared to immediate test items (Singer & Wixted, 2006). In some circumstances, criterion shifts can occur trial by trial (e.g., Heit, Brockdorff, & Lamberts, 2003; Hockley & Niewiadomski, 2007), although these shifts may lag considerably behind changes in the decision environment (Brown & Steyvers, 2005). Participants' subjective perception of task difficulty (related to perceived memorability of study lists) has also been shown to impact on criterion placement (Bruno, Higham, & Perfect, 2009) and there appear to be reliable individual differences (e.g., Aminoff et al., 2012; Kanter & Lindsay, 2012, 2014). Together, these results led us to expect that, if a criterion shift occurs in response to variation in test item quality, participants will use a more conservative criterion for high-quality test items (i.e., they would need a stronger familiarity signal before declaring a test item "old") than for low-quality test items, where a more liberal criterion would apply. The shift would reflect participants' anticipation of stronger familiarity signals from high-quality old test items than from low-quality old test items (e.g., Brown, Lewis, & Monk, 1977). Such a criterion shift would be compatible with the results of a relevant study by Hockley, Hemsworth, and Consoli (1999). When participants studied normal face stimuli and then carried out a recognition task with normal faces and with degraded faces (wearing sunglasses), a mirror effect occurred (see Glanzer & Adams, 1985), with degraded test stimuli producing lower hit rates and higher false-alarm rates (Hockley et al., 1999).

We carried out three old-new recognition experiments. In the study phase of the experiments, the participants observed a number of images (scenes in all three experiments, and also faces and words in Experiment 3). In the subsequent test phase, images that had been presented at study (Old items) were intermixed with unseen images (New items). The participants were asked to decide for each test item whether it was old or new. In all experiments, some of the test images were degraded. Unlike the experiments in Hockley et al. (1999), the whole stimulus was degraded, similar to the study by Wolfe and Kuzmova (2011). In Experiment 1, the degradation was done through low-pass Gaussian filtering, blurring the images. In Experiments 2 and 3, short exposure durations were used to reduce perceptual quality. At short exposure durations, coarse stimulus information is likely to be more available for further processing than fine-grained information (e.g., Fabre-Thorpe, 2011), and so we expected to find similar degradation effects across experiments. In addition, Experiments 1 and 2 gave trial-by-trial feedback about performance (correct/incorrect) at test, whereas Experiment 3 did not provide participants with feedback.

Experiment 1

Method

Participants Thirty-nine (29 female) students and research staff from the University of Bristol and the University of Warwick participated either in return for course credit or as a volunteer. Mean age was 21:3 years and all reported normal or corrected-to-normal vision.

Materials Stimuli were presented on a Cathode Ray Tube monitor set to 1,152 x 864 controlled via a Pentium class PC running custom written software. Responses were made via a mouse connected to the Universal Serial Bus controller of the PC. Stimuli consisted of 128 digital photographs of real world scenes taken of four subjects (32 images of each; two from each were reserved for presentation at the beginning and end of the study list to control primacy and recency effects): traffic scenes, woodland scenes, buildings, and rivers. For the blurred images we applied a low-pass Gaussian filter with a standard deviation of 25 pixels. The complete set of stimuli is available on request from the first author.

Design and procedure Test cue quality was manipulated within participants. Sixty old and 60 new stimuli were randomly selected from the 120 images for each participant. Of each of the 60 old and 60 new stimuli, 30 were randomly selected to be blurred.

Participants sat alone in a quiet room at a distance of 100 cm from the monitor. The study phase consisted of 68 stimulus presentations. Four stimuli, which were not later tested, were presented at the start of the list and at the end of the list; these were used to reduce the impact of primacy and recency effects. Each study stimulus was presented for 2,000 ms, with an inter-stimulus interval of 500 ms consisting of a neutral gray screen. After the study phase, participants were asked to select a mouse button for their "old" responses (the other button being used for "new" responses). The test phase then started. Each trial started with the presentation of a black central fixation cross on a gray background for 500 ms, followed by a blank gray screen for 100 ms. The test stimulus then appeared, and was displayed until the response was given. Participants were informed they should respond as quickly and as accurately as possible. Once participants had made an old/new response they were presented with a confidence rating screen, in which they clicked on one of four text boxes to indicate how confident they were in the correctness of their response: “Guess”, “Maybe”, “Probably”, and “Definitely”. “Correct”/”Wrong” feedback was then provided centrally for 750 ms. Blurred and Clear stimuli were randomly intermixed. The experiment lasted approximately 15 minutes per participant.

Table 1 Hit and false alarm rates to clear and blurred stimuli in Experiment 1 and Long and Short stimuli in Experiment 2

	Experiment 1		Experiment 2	
	Clear	Blurred	Long	Short
Hit rate	.610	.584	.658	.680
False alarm rate	.227	.442	.192	.533

Results and discussion

We first analyzed the old/new decision data, without taking into account the confidence ratings. Table 1 summarizes the response proportions in the Clear and Blurred conditions, respectively, across all participants. Unlike Hockley et al. (1999), we did not observe a strong mirror effect across non-degraded and degraded test stimuli. The hit rates for clear and blurred old items were very similar, with only slightly more errors in the blurred condition, $t(38) = 1.00$, $p = .33$. However, for new items, the false alarm rate was higher in the blurred condition (.442) than in the clear condition (.227), $t(38) = 8.39$, $p < .001$, $SEM = 0.03$, $d = 1.35$. Sensitivity (d_a) and criterion (c_a) values under a standard Gaussian SDT model were calculated (we used RscorePlus, Harvey, 2010, for all signal detection analyses). As expected, d_a was significantly higher in the clear condition (0.92 95 % CI ± 0.09)[1] than in the blurred condition (0.31 ± .08; difference = 0.61 ±0.09), showing that blurring was effective in reducing cue quality. In addition, there was a significant difference in bias between the conditions (difference = 0.27 ±0.07), with participants using a more conservative criterion in the clear condition (c = 0.23) than in the blurred condition (c = -0.03).

To better understand the nature of the criterion shift, we extended the analysis to include the confidence rating data. The confidence ratings for old and new responses were combined to construct a single 8-point scale, with 1 meaning “definitely new” and 8 meaning “definitely old.” On this scale, a value of 4 corresponded to a "guess" rating following a *new* response, and a value of 5 represented a "guess" rating following an old response. Figure 1 shows the proportions of responses in the eight confidence categories, as a function of stimulus type (clear vs. blurred, and old vs. new). Not surprisingly, the observers were more reluctant to express high confidence in correct responses in the blurred condition than in the clear condition. The data on the 8-point scale were then used to construct *z*-ROC curves, on the basis of transformed hit and false alarm rates across different levels of confidence (see Macmillan & Creelman, 2005). Figure 2 shows the *z*-ROCs for the clear and blurred test items. As expected, the *z*-ROC for the Clear condition shows greater overall discriminability than that for the Blurred condition.

To obtain criterion estimates, a conventional approach is to estimate a separate decision criterion for each pair of adjacent scores on the scale (Macmillan & Creelman, 2005). For an 8-point scale, this implies that seven criteria have to be estimated. The criterion estimates were obtained using a variation of the Marquardt method to find maximum-likelihood parameter estimates (Harvey, 2010). The psychophysical model assumed Gaussian distributions and allowed for unequal variances of the “old” and “new” signal distributions. In each condition (Clear or Blurred), nine parameters were estimated (seven criteria, and the mean and variance of the “old” signal

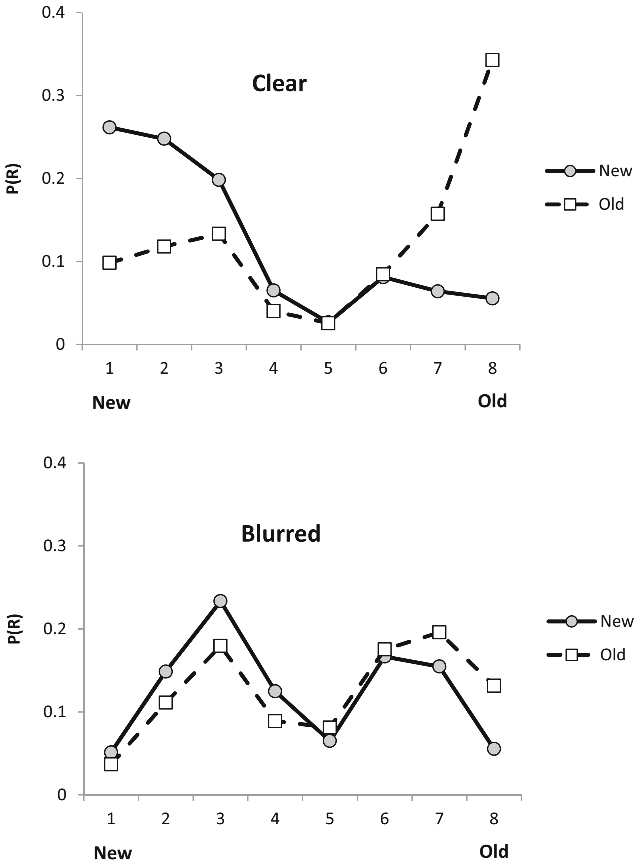

Fig. 1 Proportion of responses at each confidence level for Old Stimuli (broken lines) and New Stimuli (filled lines) in the Clear (top panel) and Blurred (bottom panel) conditions of Experiment 1

[1] We report ±95 % confidence intervals throughout where appropriate.

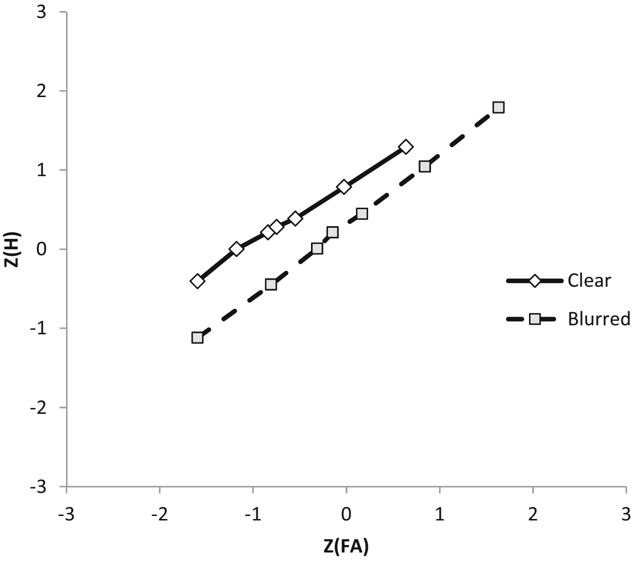

Fig. 2 *Z*-ROC curves calculated on the seven confidence ratings for the Clear (filled lines) and Blurred (broken lines) conditions from Experiment 1

distribution, assuming without loss of generality that the "new" distribution is standard normal). Figure 3 shows an overview of the estimated distributions and criteria for Clear and Blurred test items. The estimated criterion values differ between Clear and Blurred test items (the horizontal bars at the top of each criterion line show the 95 % CI around the estimated value), with generally more conservative settings in the Clear condition than in the Blurred condition.

A crucial question is why the observed criterion values were chosen. Confidence criteria can be set according to different principles (see Stretch & Wixted, 1998). A pattern in which the criteria are spread further apart in the condition with the smaller discriminability is qualitatively compatible with a likelihood-ratio principle (Stretch & Wixted, 1998), according to which participants maintain a constant ratio between the likelihood of a test item being "old" versus "new" for each confidence criterion, across all test conditions. We computed log-likelihood ratios for each of the criteria in the two conditions in Experiment 1, using the equivalent of Equation A4 in Stretch and Wixted (1998). As shown in Fig. 4, the likelihood ratios differ between the Clear and Blurred conditions, and the results therefore do not support the idea that criteria are set to maintain constant likelihood ratios (note that this conclusion rests on the assumption that the standard deviation of the

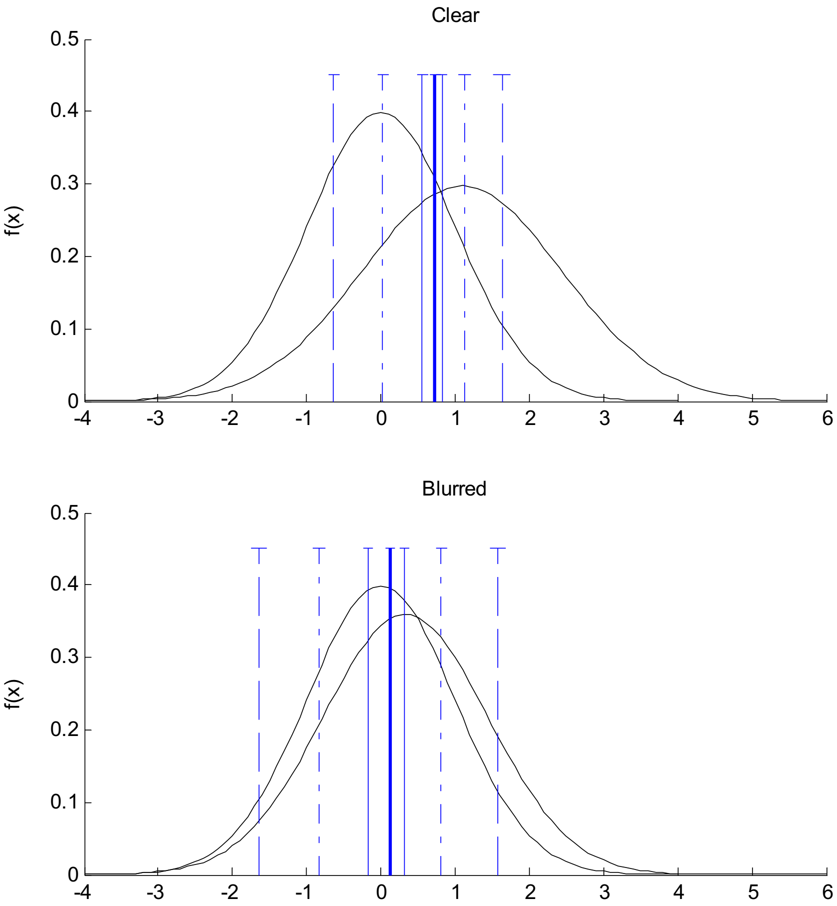

Fig. 3 Estimated distributions and criteria for Clear (top panel) and Blurred (bottom panel) test items from Experiment 1. Distributions to the right are Old Stimuli and distributions to the left New Stimuli. The seven criteria are shown as vertical lines. The horizontal lines show the standard errors of the criterion estimates

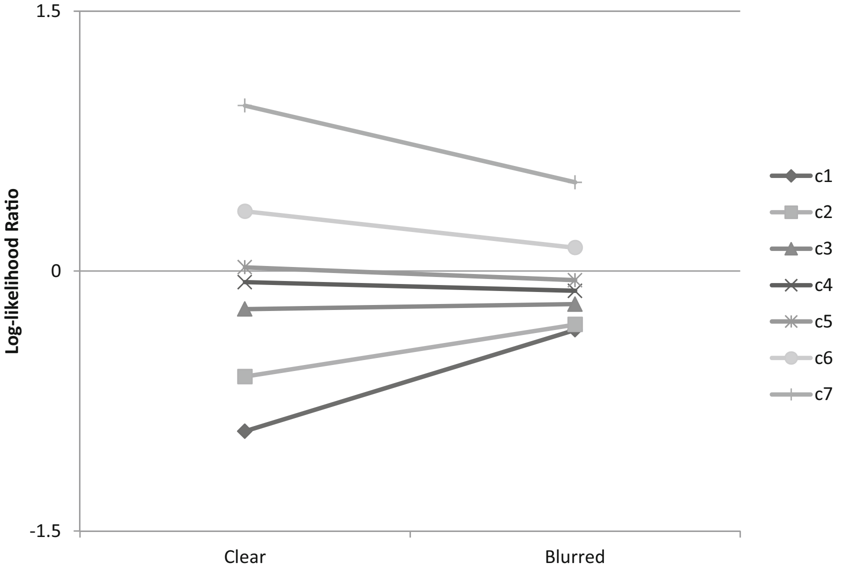

Fig. 4 Log-likelihood ratios for each confidence criterion (c) in the Clear and Blurred conditions of Experiment 1

"new" distribution is the same between conditions). Instead, the criteria seem to reflect the observers' desire to maintain a constant hit rate (as shown in Table 1), combined with a general reduction in confidence for the blurred stimuli. It is remarkable that the observers were willing to tolerate a high false-alarm rate in the blurred condition to maintain a steady hit rate. This suggests that, in the blurred condition, false alarms (saying "old" to new stimuli) were seen as less problematic than misses (i.e., saying "new" to old stimuli). We will consider the reasons for this in the General Discussion.

Experiment 2

Experiment 2 was a designed as a replication of Experiment 1. However, instead of blurring, short exposure duration was used to degrade the stimulus percept.

Method

Participants Twenty-five (22 female) students from the University of Warwick participated in return for course credit. Mean age was 19:3 years and all reported normal or corrected-to-normal vision.

Materials The equipment was identical to Experiment 1. We used the same 128 stimuli, but participants only ever saw the clear stimuli.

Design and procedure The design and procedure were identical to Experiment 1, with the following exception: In the test phase, instead of using blurred stimuli, we displayed clear stimuli in the Short condition for 500 ms, and in the Long condition for 2,000 ms, before removing the stimulus from screen. The stimuli assigned to the Short and Long conditions were randomized per participant.

Results and discussion

Table 1 shows the proportion of new and old responses as a function of exposure duration. Mirroring the effect of blurring in Experiment 1, the hit rate was almost identical at both exposure durations, $t(19) = 0.69$, $p = .50$, but the false-alarm rate was much higher in the Short condition, $t(19) = 9.24$, $p < .001$, $SEM = 0.04$, $d = 2.08$. d_a was significantly lower on Short exposure trials (0.45 ±0.12) than on Long exposure trials (1.12 ±0.13; difference = 0.66, ±0.09). The criterion c also differed between the conditions (c = -0.28 for the Short and c =0.23 for the long condition, difference = 0.51, ±0.11), implying that judgments were more liberal in the short exposure condition, replicating the effect of blurring in Experiment 1.

Figure 5 shows the confidence ratings for new and old stimuli as a function of exposure duration. The ratings were calculated in the same way as for Experiment 1. The pattern of responses is similar to that in Experiment 1. The participants were more reluctant to express high confidence (i.e., respond in categories near 1 or 8) in the short exposure condition than in the long exposure condition. The z-ROCs (see Fig. 6) also look similar to those from Experiment 1.

On the basis of the choices and the confidence ratings, seven criteria were estimated in the same way as for Experiment 1. The criteria are shown in Fig. 7, and the corresponding log-likelihood ratios in Fig. 8. There is a clear shift towards more liberal criteria in the short exposure condition, but there is again no evidence for the

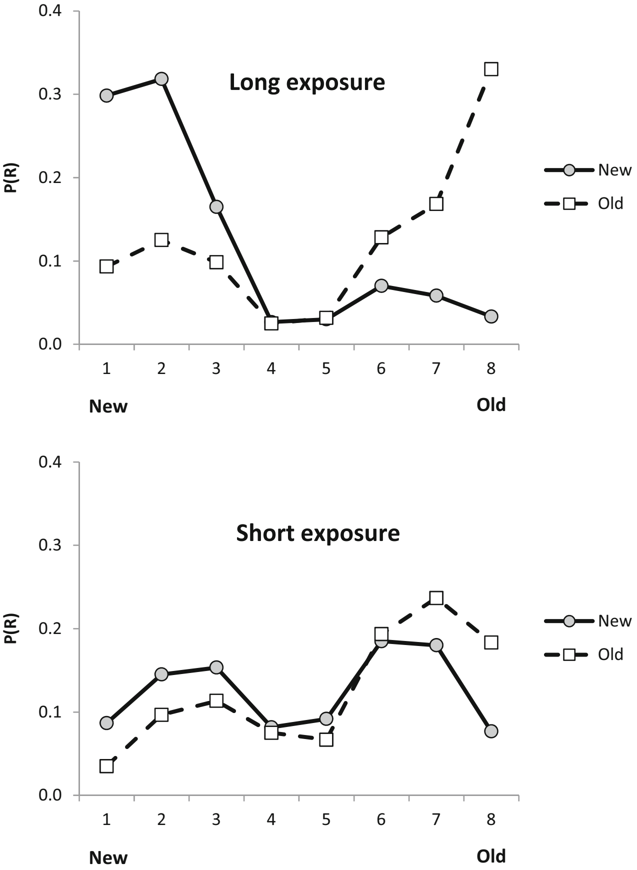

Fig. 5 Proportion of responses at each confidence level for Old Stimuli (broken lines) and New Stimuli (filled lines) in the Long (top panel) and Short (bottom panel) exposure conditions of Experiment 2

likelihood-ratio principle. Altogether, the results of Experiment 2 are remarkably similar to those from Experiment 1, despite the use of a different method for degrading the perceptual quality of the stimuli.

The results from the two experiments and the model-based analyses demonstrate that degradation of test items (by blurring or by limiting exposure duration) affected the criterion for deciding whether a stimulus had been seen before or not. If the test item was degraded, a more liberal criterion setting was used. Moreover, the criterion shift occurred at all reported levels of confidence in the recognition judgments. The question that arises immediately is why the participants shifted criteria in this way. Why were degraded items more likely to be declared "old" than non-degraded items?

We have ruled out an explanation in terms of likelihood-ratio preservation; the participants did not set confidence criteria to maintain a constant ratio of the likelihoods that a test item was old or new. Instead, the effect of the criterion shift was that the hit rate (correct recognition of old items) remained almost constant across all items in each experiment, whereas the false-alarm rate (incorrect recognition of new items) was considerably higher for degraded items. To understand this pattern of results better, it is helpful to consider hit and false-alarm rates for degraded and non-degraded test items at each level of expressed confidence, as shown in Figs. 9 (Experiment 1) and 10 (Experiment 2). In both experiments, hit rates were higher at higher levels of confidence, with relatively small differences in hit rates between non-degraded and degraded test items. False alarm rates, on the other hand, showed a different pattern. In both experiments, the difference in false-alarm rates between non-degraded and degraded items was much greater at high levels of confidence. The observers maintained similar hit rates across levels of degradation at the expense of variation in the false-alarm rates. The criterion shifts we observed are consistent with an interpretation in terms of anticipated strength of the familiarity signal for old items. If the participants expected that old degraded stimuli would not feel as familiar as old non-degraded stimuli, they would shift their criteria to ensure that the weaker familiarity signal still produced a good number of hits for old items. Observers would be prepared to tolerate the higher false-alarm rates for new items to achieve consistent hit rates for old items. In a more general sense, the results would thus provide another example of a criterion shift because of variation in perceived task difficulty (e.g., Bruno et al., 2009).

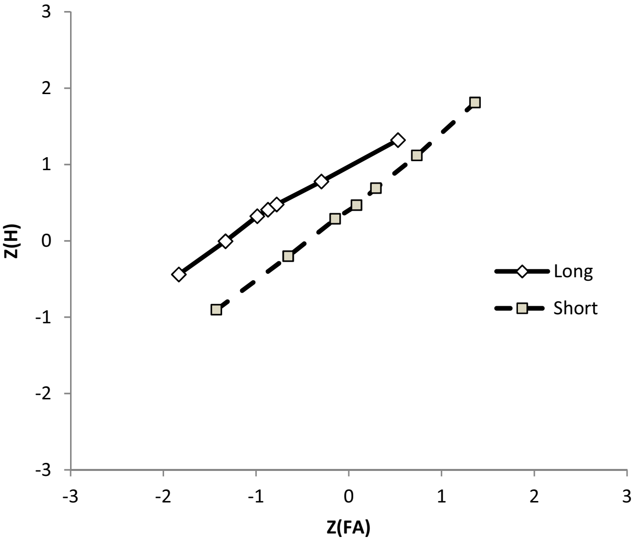

Fig. 6 *Z*-ROC curves calculated on the seven confidence ratings for the Long (filled lines) and Short (broken lines) exposure conditions from Experiment 2

Experiment 3

In Experiments 1 and 2, we gave participants accuracy feedback after each response, which would have provided a clue as to the task difficulty for each item type and may have facilitated the criterion shift. Previous research indicates that feedback is necessary for participants to demonstrate substantial criterion shifts (Estes & Maddox, 1995; Rhodes & Jacoby,

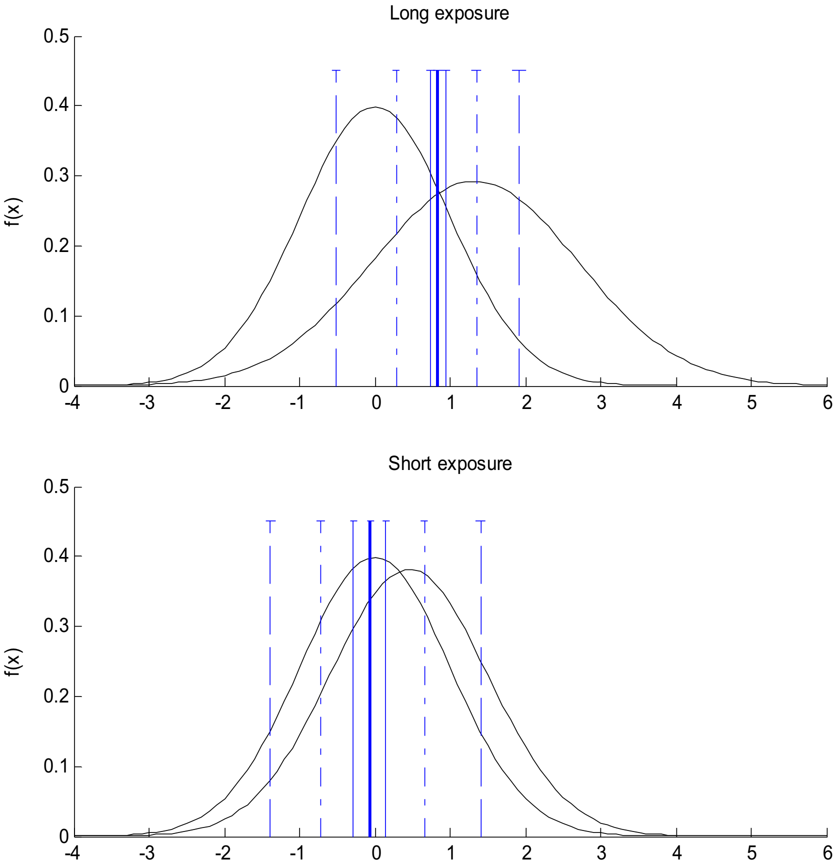

Fig. 7 Estimated distributions and criteria for Long (top panel) and Short (bottom panel) exposure conditions from Experiment 2. Distributions to the right are Old Stimuli and distributions to the left New Stimuli. The seven criteria are shown as vertical lines. The horizontal lines show the standard errors of the criterion estimates

2007; Verde & Rotello, 2007). For example, Verde and Rotello (2007) failed to find a criterion shift in four experiments involving strong and weak study items (through manipulating stimulus frequency and duration) in which the test block was split by item type, even when they cued strength by using different semantic categories for strong and weak items. In a fifth experiment Verde and Rotello provided accuracy feedback and found sizeable criterion shifts. In

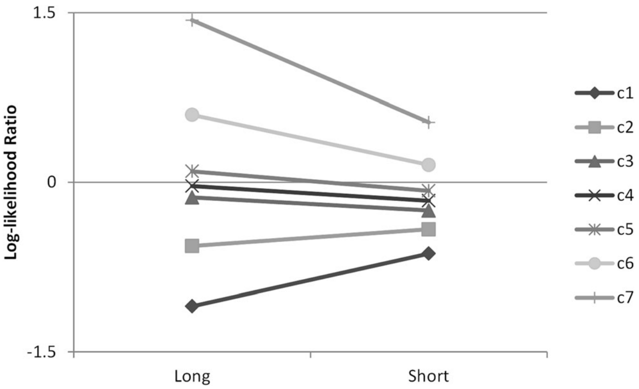

Fig. 8 Log-likelihood ratios for each confidence criterion (c) in the Short and Long exposure conditions of Experiment 2

Experiment 3, we test whether the criterion shifts observed in Experiments 1 and 2 were contingent on the availability of immediate accuracy feedback.

Experiment 3 was a replication of Experiment 2, without accuracy feedback given to participants. In addition, we used three stimulus types (the scenes from Experiments 1 and 2, faces, and words) to further test the generality of the results from Experiments 1 and 2.

Method

Participants Thirty-three (23 female) students from the University of Bristol and members of the Bristol public participated voluntarily. Mean age was 31 years and all participants reported normal or corrected-to-normal vision and were fluent English speakers.

Materials The equipment was identical to that in Experiments 1 and 2. Stimuli consisted of 80 woodlands scenes, 80 faces, and 80 words (randomly assigned to Old and New stimuli for each participant). The scenes were similar to the woodlands scenes used in Experiments 1 and 2. We used both male and

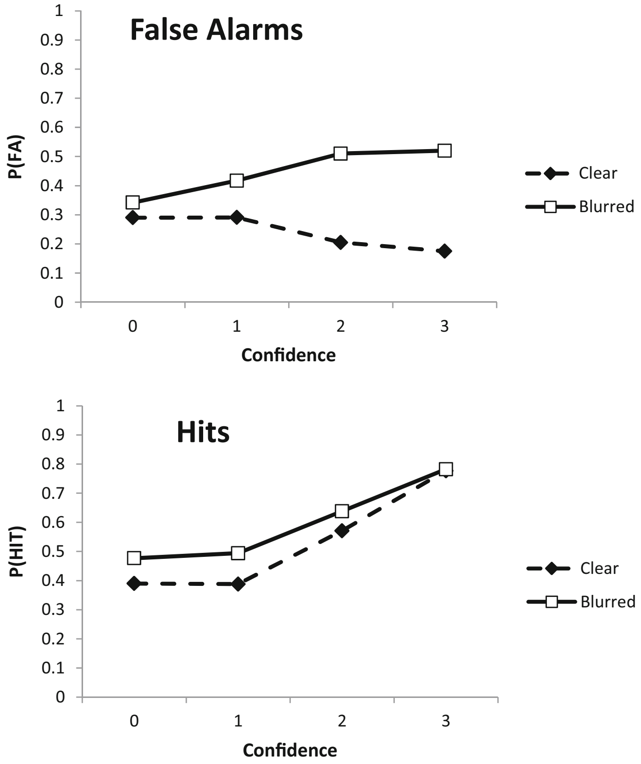

Fig. 9 Proportion of False Alarms (top panel) and Hits (bottom panel) as a function of confidence level for the Clear (broken line) and Blurred (filled line) conditions in Experiment 1

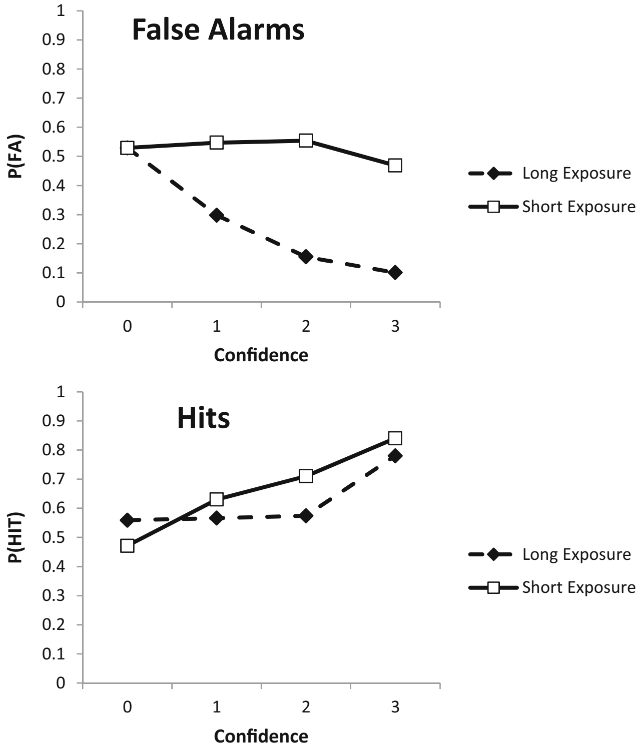

Fig. 10 Proportion of False Alarms (top panel) and Hits (bottom panel) as a function of confidence level for the Long (broken line) and Short (filled line) conditions in Experiment 2

female face stimuli (including hair and external features) chosen from the Glasgow Unfamiliar Face Database (Burton, White, & McNeill, 2010) that were not highly confusable. We selected six-letter words from the MRC Psycholinguistic Database (Coltheart, 1981) with a Kücera-Francis frequency ranging from 10 to 13. Words were presented in the center of the screen in size 48 Arial font.

Procedure The procedure was identical to Experiment 2, with a few modifications. First, in the Short condition, stimuli were shown for 50 ms, instead of 500 ms. Second, participants were asked to respond "New", "Know", or "Remember" by mouse clicking on the relevant box displayed on screen after the stimulus display, instead of responding "New or "Old". For the purposes of this study, we collapse "Know" and "Remember" responses into "Old" responses to enable comparison with Experiments 1 and 2.[2] Most importantly, participants were not given feedback as to the accuracy of their response. Instead the next trial started after a 500-ms inter-trial delay once the confidence judgment had been given. Stimulus type was randomly interleaved at both study and test.

Results and discussion

Table 2 shows the hit and false-alarm rates for the three stimulus types in the Short and Long exposure duration conditions of Experiment 3. A 3 (Stimulus Type) x 2 (Duration) repeated measure analysis of variance (ANOVA) indicated a reliable main effect of Stimulus Type, $F(2, 64) = 7.16$, $p = .002$, $MSE = .029$, $\eta_p^2 = .18$, with a higher hit rate for Words (.72) and Faces (.71) than for Scenes (.62). Importantly, there was a main effect of Duration on hit rate, $F(1, 32) = 17.73$, $p < .001$, $MSE = .016$, $\eta_p^2 = .40$, with a higher hit rate in the Long condition (.72) than in the Short condition (.65). There was also an interaction between Stimulus Type and Duration, $F(2, 64) = 3.38$, $p = .040$, $MSE = .012$, $\eta_p^2 = .096$, with the largest Duration effect for the Face stimuli. Because of the interaction we analyzed each stimulus type separately. There was no reliable difference in hit rate between the Short and Long conditions for the Scene stimuli, $t(32) = 1.40$, $p = .170$, and for the Word stimuli, $t(32) = 2.00$, $p = .054$, replicating the findings from Experiment 1 and Experiment 2. However, for

[2] As expected correct Remember responses (i.e. Remember to Old stimuli) were associated with greater confidence than correct Know responses (i.e., Know to Old stimuli), 62 % "Definitely Confident" versus 15 % "Definitely Confident", for the Remember and Know responses respectively. For Old stimuli across all stimulus types and durations, participants used the Know response on 31 % of trials, and the Remember response on 37 % of trials. Across stimulus types, for all correct "Old" responses (i.e., Remember or Know to Old stimuli) the proportion of Know responses was greater in the Speeded (53 %) condition than in the Unspeeded (39 %) condition. Across durations, for all correct "Old" responses, there were more Know responses for Scenes (57 %) than for Faces (45 %) and for Words (36 %).

Table 2 Proportion of hit and false alarm rates for Experiment 3

	Scenes		Faces		Words	
	Long	Short	Long	Short	Long	Short
Hit rate	.64	.60	.78	.64	.75	.70
False alarm rate	.32	.43	.36	.42	.33	.30

the Face stimuli there was an effect of Duration, $t(32) = 4.47$, $p < .001$, $SEM = .030$, $d = 0.79$, with a higher hit rate in the Long condition (.76) compared with the Short condition (.64).

A similar analysis for the false-alarm rates showed no difference across Stimulus Type, $F(2, 64) = 2.13$, $p = .127$, a main effect of Duration, $F(1, 32) = 6.00$, $p = .020$, $MSE = .019$, $\eta_p^2 = .16$, with more false-alarms in the Short condition (.38) compared to the Long condition (.34), replicating Experiments 1 and 2. There was also an interaction between Stimulus Type and Duration, $F(2, 64) = 4.54$, $MSE = .014$, $\eta_p^2 = .12$, with the difference in false-alarm rate greatest for the Scene stimuli. We analyzed the false-alarm rates separately for each stimulus type. For the Scene stimuli, there were more false alarms in the Short condition (.43) than in the Long condition (.32), $t(32) = 2.65$, $p = .012$, $SEM = .042$, $d = .46$. For the Face stimuli, there were also more false alarms in the Short condition (.42) compared to the Long condition (.36), $t(32) = 2.09$, $p = .044$, $SEM = .030$, $d = .36$. However, for the Word stimuli, there was no difference in false-alarm rates between Durations, $t(32) = 1.11$, $p = .276$.

Overall, the pattern of results is similar to that in Experiments 1 and 2, with the exception of the hit rate for face stimuli and false-alarm rate for the word stimuli. However, the hit rates did have a tendency to be higher for the Long duration condition and false-alarm rate differences were much smaller than those observed in Experiments 1 and 2. The average difference in hit rates between the Short and Long condition was .02 for Experiment 2, and .08 for Experiment 3. The average difference in false-alarm rates was .34 in Experiment 2 and .05 in Experiment 3.

For the Scene stimuli, d_a was higher in the Long condition (0.82 ±.11) than the Short condition (0.37 ±.12; difference = 0.45 ±0.13). For the Face stimuli, d_a was also higher in the Long condition (1.12, ±.12) than in the Short condition (0.53 ±.11; difference = 0.59 ±0.13). However, for the Word stimuli, the difference between the Long condition (1.16 ±0.12) and the Short condition was not statistically reliable (1.06 ±0.12; difference= 0.10 ±0.13). This may reflect the fact that unmasked words can typically be identified after just 60 ms of exposure (e.g., Rayner, Liversedge, White, & Vergilino-Perez, 2003). For the Scene stimuli, there was a reliable difference in bias between the Long condition (c = 0.05 ±0.003) and Short condition (c = -0.04 ±0.003, difference= 0.09 ±0.004) with participants using a more liberal criterion in the short condition. The same pattern of bias was found for the Face stimuli (Short = -0.20 ±0.003; Long = -0.08 ±0.003; difference= 0.11 ±0.004) and Word stimuli (Short = -0.11 ±0.003; Long = 0.002 ±0.003; difference= 0.12 ±0.004). Thus, despite the lack of difference in discriminability between the Long and Short conditions for the Word stimuli, participants still shifted their criteria, presumably because of the perceived increased difficulty of the Short condition.

Figure 11 shows the confidence ratings for each Stimulus type for each Duration condition. For the Scenes and Faces, the pattern is similar to that shown in Fig. 1 (Experiment 1) and Fig. 5 (Experiment 2) with participants more reluctant to express higher levels of confidence (1 or 8) in the Short conditions compared to the Long conditions. However, for the Word stimuli the patterns across both Short and Long exposures were very similar, with only a slightly greater reluctance (a reduction of 6 %) to use the extreme responses in the Short condition. Figure 12 shows the Z-ROC curves for each stimulus type and each exposure duration. Again, for the Scenes and Faces the pattern is very similar to Fig. 2 (Experiment 1) and Fig. 6 (Experiment 2) apart from for the Word stimuli, where the two Z-ROC curves lie on top of each other, reflecting the lack of difference in discriminability between the Long and Short exposures. Figure 13 shows the seven criteria estimates for each exposure and stimulus type and the corresponding log-likelihood ratios are given in Fig. 14. For the Scene and Face stimuli the patterns are similar to Figs. 3 and 4 (Experiment 1) and Figs. 7 and 8 (Experiment 2), except for the central response criteria, which do not shift much between the Short and Long exposures. For the Word stimuli, the criteria were largely the same in both the Short and Long conditions, with the Log Likelihood ratio preserved, except for the two most extreme criteria ("Definitely Old" and "Definitely New"). Overall, the pattern of results from the Scenes and Faces is broadly similar to Experiments 1 and 2, but for Words the test cue degradation appeared to have been less effective, presumably because participants were able to encode the word within the 50-ms presentation window in the Short condition.

General discussion

Across the three experiments, we found evidence for within subject, trial-by-trial, criterion shifts in response to degradation of a test cue. The criterion shifts were largest when feedback was available (Experiments 1 and 2) and diminished when feedback was absent (Experiment 3). Experiments 1 and 2 did not produce mirror effects similar to those obtained by Hockley et al. (1999). However, in a recent follow-up study, Vokey and Hockley (2012) demonstrated that the previously observed mirror effects for partially obscured faces

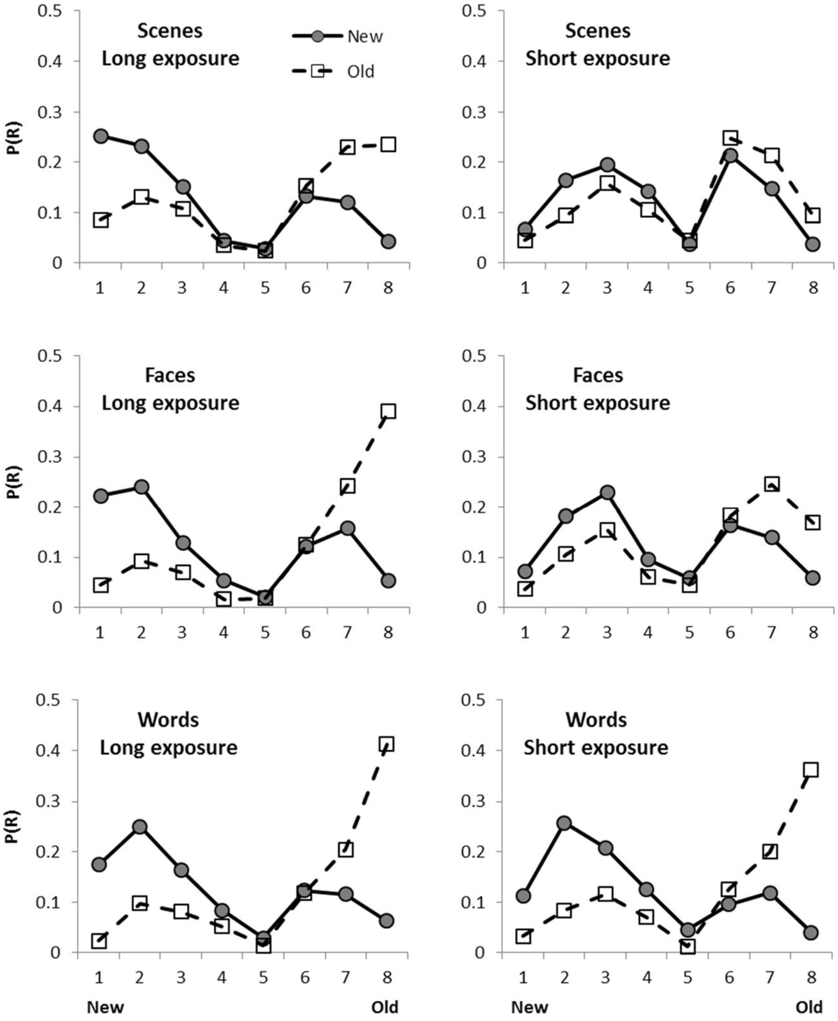

Fig. 11 Proportion of responses at each confidence level for Old Stimuli (broken lines) and New Stimuli (filled lines) in the Long (left panel) and Short (right panel) conditions of Experiment 3 for each Stimulus type (rows)

actually consists of two separable processes (differences in discrimination and changes in decision criteria), and are therefore more complex than initially assumed. Still, Vokey and Hockley (2012) came to the conclusion that their participants were likely to have adopted a more liberal criterion for degraded test items. Analysis of Experiments 1 and 2 apparently provides strong evidence in support of that conclusion, using different stimuli (scenes), degrading the whole stimulus, and using two different methods of degradation, both avoiding the confounds that occurred in the original study by Hockley et al. (1999; see Vokey & Hockley, 2012, for a discussion). Experiment 3, however, showed more typical mirror effects with lower hit rates and higher false alarms in the degraded conditions as well as criterion shifts (except for the Words which showed slightly more false alarms in the non-degraded condition). Experiment 3 therefore provides evidence that feedback is important. When accuracy feedback was not given, the criterion shifts were smaller, but, unlike previous experiments (e.g., Verde & Rotello, 2007), there were still reliable shifts in criteria (at least for the faces and scenes) despite the lack of feedback. Feedback presumably acts to provide a clearer insight into the differential levels of performance across trials types (durations), and hence allows participants to better gauge the need for differential criteria depending on the trial type. This has important implications for real-world applications, where immediate and consistent feedback is usually not available.

In the modelling, we have assumed that familiarity of new degraded and non-degraded test items is the same, by assuming identical means and variances of the no-signal distributions. An alternative assumption (following Vokey & Hockley, 2012) could be that new degraded items are less familiar than new non-degraded items. In the SDT modelling, this assumption would be implemented by allowing the non-signal distribution means and variances to vary between degraded and non-degraded items. However, such a generalized model cannot be identified, which is problematic and the main reason for assuming

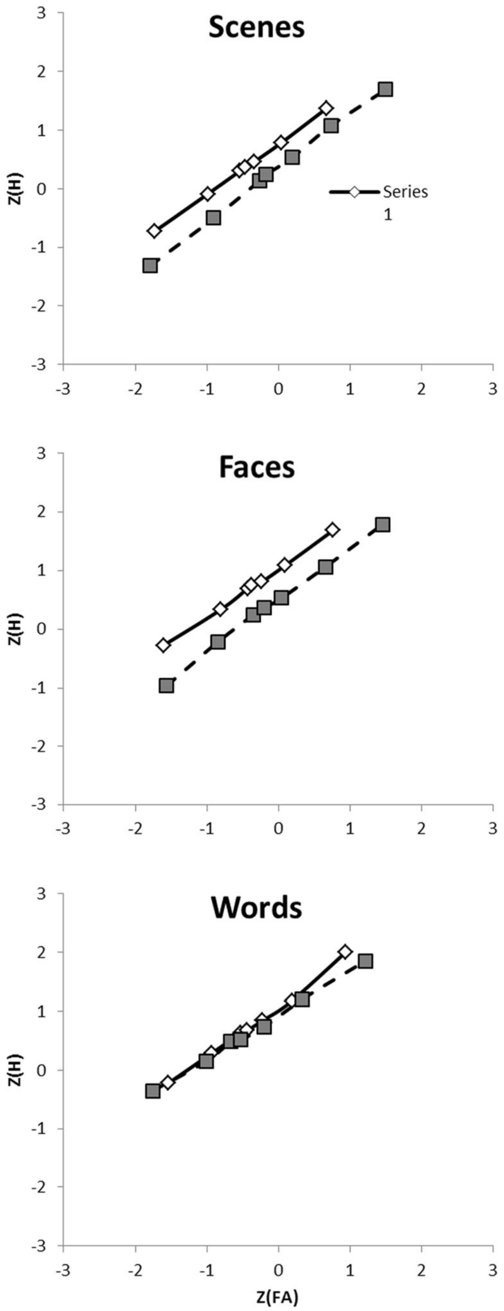

Fig. 12 *Z*-ROC curves calculated on the seven confidence ratings for the Long (filled lines) and Short (broken lines) conditions from Experiment 3 for each Stimulus type

identical no-signal distributions. Regardless, the (untested) distributional identity assumption is by no means problematic for our conclusions about criterion shifts. If it is the case that new degraded stimuli are less familiar than new non-degraded stimuli, the results could only be explained if the criterion shifts were even stronger than we have inferred. In that sense, the equal-familiarity assumption is a safe assumption to make.

However, it is also possible that the New degraded items are in fact *more* familiar than New non-degraded items, perhaps because the degradation removes or prevents encoding of critical distinguishing features.[3] Such a pattern of results would be compatible with differentiation models of episodic memory (e.g., McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997) in which the list strength mirror effect is due to the stronger encoding of more memorable items (increasing the hit rate) which reduces the similarity to New items following a strong list (reduced false alarm rate; see Criss, 2006, for an overview). Thus, when a degraded New item is shown, it appears more familiar than a non-degraded New item, increasing the likelihood of a false alarm. Because there was no manipulation at encoding, all Old items have the same level of encoding and thus, according to differentiation models, the hit rate should not be different between degraded and non-degraded items. The data from Experiments 1 and 2 do not allow us to conclusively rule out this explanation. However, the importance of feedback in modulating the pattern of false alarms and hit rates (from Experiment 3) is consistent with a flexible decision process, rather than as a consequence of differential encoding. Experiment 3 demonstrated a (albeit smaller) criterion shift (measured by *c*) across all stimulus types between the short and long duration conditions. For the scene and face stimuli there was also evidence of individual confidence criteria shifting. The scene stimuli largely followed the pattern of Experiments 1 and 2, with hit rate remaining constant across durations, but the false alarm rate higher in the short duration compared to the long duration condition. The face stimuli however, showed a more typical mirror effect, with both a change in the hit rate and false alarm rate across conditions. Although the word stimuli showed a small criterion shift (measured by *c*), there was no effect on either the hit rate or the false alarm rate, and minor changes in individual confidence criteria. Thus, it appears possible that there are stimulus-specific effects, such that participants can hold multiple criterion across both cue quality and stimulus type. But it is likely that the difference in discriminability between the cue quality conditions determines whether participants maintain multiple criterion within a stimulus type or not rather than stimulus type *per se*.

Better understanding the role of strategic shifts in the decision process is clearly an important goal for future studies in deciding between competing models of episodic memory. Our initial explanation that participants attempt to hold the hit rate constant for scene stimuli (Experiments 1 and 2) does not appear to generalize to other stimulus types (words and faces, Experiment 3). It might be that providing explicit feedback allows participants to shift their criteria for words and faces as well (especially if the differences in discriminability between the cue quality conditions is large). If the words were

[3] We thank two anonymous reviewers for pointing this out.

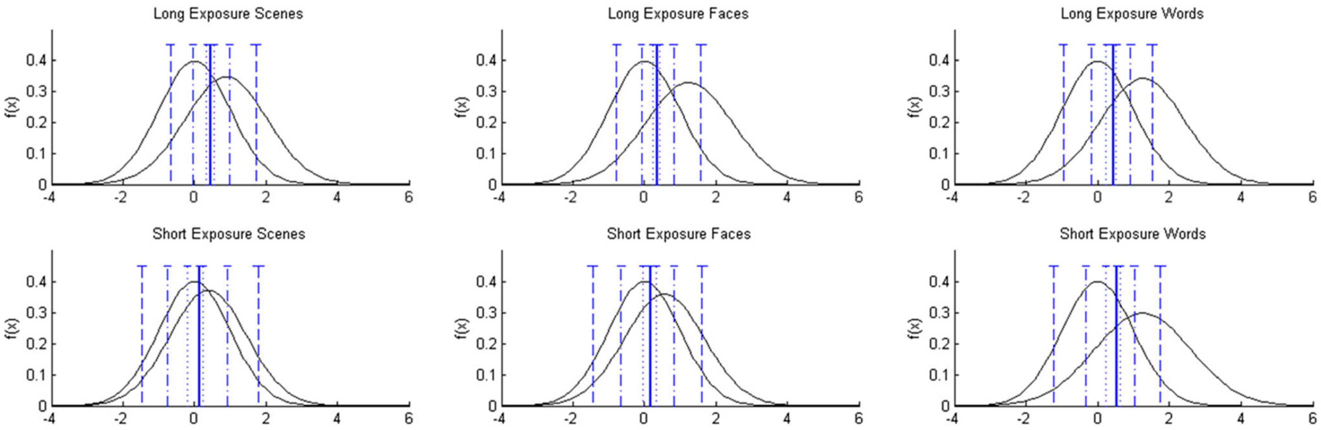

Fig. 13 Estimated distributions and criteria for Long (top row) and Short (bottom row) test items from Experiment 3 for each stimulus type (column). Distributions to the right are Old Stimuli and distributions to the left New Stimuli. The seven criteria are shown as vertical lines. The horizontal lines show the standard errors of the criterion estimates

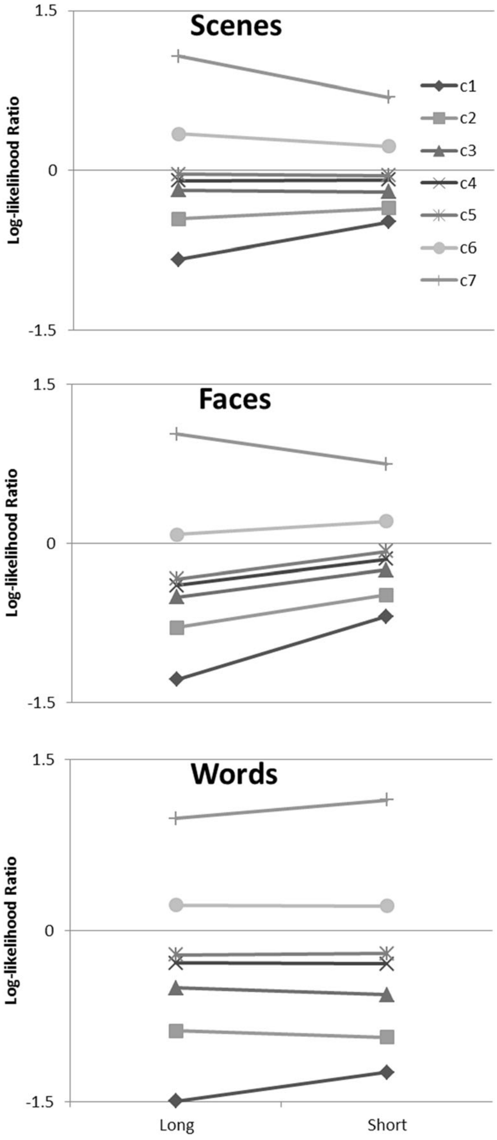

Fig. 14 Log-likelihood ratios for each confidence criterion (c) in the Long and Short conditions of Experiment 3 for each stimulus type

sufficiently dissimilar and the duration insufficiently short to impact encoding at test, then we might expect (as we observed) no differences in discriminability or significant criteria shifts. For faces and scenes, however, where discriminability was lower in the short duration condition, this may encourage a shift in criteria. This does not explain why the mirror effect was observed for faces but not scenes, but this result should be interpreted with caution given that the pattern of hits and false alarm rates for the scene stimuli did move in the direction of a mirror effect. The scene stimuli had the lowest level of discriminability across all experiments, and this is likely to encourage participants to use different criteria, especially with a perceived (and actual) large difference between short and long duration trials. This perceived difference in discriminability is likely to be a factor in determining the size of criteria shifts, and would have been smallest in Experiment 3, perhaps contributing to the smaller and less consistent criterion shifts.

Our results may have important implications for realistic old-new recognition tasks, such as those that may occur in forensic circumstances. While it is well established that encoding factors (such as distance and lighting of an object) can have an impact on recognition judgments in forensic settings (e.g., Loftus & Harley, 2005), it is less well understood how perceptual factors at retrieval impact on recognition. Our findings show that, for example, if a witness were asked to identify a vehicle involved in a hit-and-run collision on the basis of a poor-quality image, the probability of a false-positive identification would be higher than if a high-quality image were used, even in the absence of feedback. Future work will need to establish how this generalizes to other stimulus types.

Acknowledgements This work was supported by an Experimental Psychology Society Grant to Kent. We thank Amy Walters for allowing us use of her photographs as stimuli and Malik Refaat for help in data collection for Experiments 1 and 2. The data for Experiment 3 was collected as part of Richard Patton's MSc dissertation.

References

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

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

Brown, S., & Steyvers, M. (2005). The dynamics of experimentally induced criterion shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 587-599.

Bruno, D., Higham, P.A., & Perfect, T.J. (2009). Global subjective memorability and the strength-based mirror effect in recognition memory. *Memory & Cognition, 37*, 807-819.

Burton, A.M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods, 42*, 286 – 291

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

Cox, G. E., & Shiffrin, R. M. (2012). Criterion setting and the dynamics of recognition memory. *Topics in Cognitive Science, 4*, 135-150.

Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory & Language, 55*, 461-478.

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

Fabre-Thorpe, M. (2011). The characteristics and limits of rapid visual categorization. *Frontiers in Psychology, 2:243*. https://doi.org/10.3389/fpsyg.2011.00243.

Glanzer, M. & Adams, J. K (1985). The mirror effect in recognition memory. *Memory & Cognition, 13*, 8-20.

Harvey, L. O. (2010). *RscorePlus* (version 5.6.1). Downloaded from: http://psych.colorado.edu/~lharvey/Software%20Zip%20Files/RscorePlus_Win.zip

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

Hicks, J. L., & Starns, J. J. (2014). Strength cues and blocking at test promote reliable within-list criterion shifts in recognition memory. *Memory & Cognition*, *42*, 1-13.

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

Hockley, W. E. (2011). Criterion changes: How flexible are recognition decision processes? In P. Higham & J. Leboe (Eds.), Constructions of Remembering and Metacognition: Essays in Honor of Bruce Whittlesea (pp. 155-166). Houndmills: Palgrave Macmillan.

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*, 679-688.

Hockley, W. E., Hemsworth, D. H., & Consoli, A. (1999). Shades of the mirror effect: Recognition of faces with and without sunglasses. *Memory & Cognition, 27*, 128-138.

Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition, 40*, 1163-1177.

Kantner, J., & Lindsay, D. S. (2014). Cross-situational consistency in recognition memory response bias. *Psychonomic Bulletin & Review*, 21, 1270-1280.

Loftus, G. R., & Harley, E. M. (2005). Why is it easier to identify someone close than faraway? *Psychonomic Bulletin & Review, 12*, 43-65.

MacMillan, N. A. & Creelman, C. D. (2005). Detection Theory: A User's Guide *(2nd)*. Mahwah: Lawrence Erlbaum Associates.

Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology, 57*, 335-384.

Mandler, G. (1991). Your face looks familiar but I can't remember your name: A review of dual-process theory. In W. E. Hockley & S. Lewandowsky (Eds.), Relating Theory and Data: Essays in Honor of Bennet B. Murdock (pp. 207–226). Hillsdale: Erlbaum.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*, 724-760.

Rayner, K., Liversedge, S. P., White, S. J., & Vergilino-Perez, D. (2003). Reading disappearing text cognitive control of eye movements. *Psychological Science, 14*, 385-388.

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, & Cognition, 33*, 305-320.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*, 145-166.

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

Starns, J. J., Ratcliff, R., & White, C. N. (2012). Diffusion model drift rates can be influenced by decision processes: An analysis of the strength-based mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 1137.

Starns, J. J., White, C. N., & Ratcliff, R. (2010). A direct test of the differentiation mechanism: REM, BCDMEM, and the strength-based mirror effect in recognition memory. *Journal of Memory and Language*, *63*, 18-34.

Starns, J. J., White, C. N., & Ratcliff, R. (2012). The strength-based mirror effect in subjective strength ratings: The evidence for differentiation can be produced without differentiation. *Memory & Cognition, 40*, 1189-1199.

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

Uttl, B., Graf, P., & Siegenthaler, A. L. (2007). Influence of object size on baseline identification, priming, and explicit memory. *Scandinavian Journal of Psychology, 48*, 281-288.

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

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*, 739-759.

Wolfe, J. M., & Kuzmova, Y. I. (2011). How many pixels make a memory? Picture memory for small pictures. *Psychonomic Bulletin & Review, 18*, 469-475.