

Global subjective memorability and the strength-based mirror effect in recognition memory

DAVIDE BRUNO AND **PHILIP A. HIGHAM**
 University of Southampton, Southampton, England

AND

TIMOTHY J. PERFECT
 University of Plymouth, Plymouth, England

Between-list manipulations of memory strength through repetition commonly generate a mirror effect, with more hits and fewer false alarms for strengthened items. However, this pattern is rarely seen with within-list manipulations of strength. In three experiments, we investigated the conditions under which a within-list mirror effect of strength (items presented once or thrice) is observed. In Experiments 1 and 2, we indirectly manipulated the overall subjective memorability of the studied lists by varying the proportion of nonwords. A within-list mirror effect was observed only in Experiment 2, in which a higher proportion of nonwords was presented in the study list. In Experiment 3, the presentation duration for each item (0.5 sec vs. 3 sec) was manipulated between groups with the purpose of affecting subjective memorability. A within-list mirror effect was observed only for the short presentation durations. Thus, across three experiments, we found the within-list mirror effect only under conditions of poor overall subjective memorability. We propose that when the overall subjective memorability is low, people switch their response strategy on an item-by-item basis and that this generates the observed mirror effect.

In recognition memory experiments, participants first study a list of items, and then, later, in a recognition memory test, they attempt to discriminate previously presented items (*targets*) from novel ones (*distractors*). A commonly used conceptual tool for understanding recognition memory performance is signal detection theory (SDT). According to SDT, targets and distractors on the recognition memory test are each distributed over a psychological strength-of-evidence dimension, with targets having higher mean strength than distractors (Figure 1). To make a recognition decision, participants are assumed to adopt a response criterion (C in Figure 1) somewhere along the strength-of-evidence dimension. If a test item has strength equal to or above the criterion, it is judged *old*; otherwise, it is judged *new*. The proportions of targets and distractors that are called *old* are dubbed the hit rate (HR) and false alarm rate (FAR), respectively.

The mirror effect is a phenomenon of recognition memory in which better *old–new* discrimination in one condition is manifested as both a higher HR and a lower FAR (e.g., Glanzer & Adams, 1985; Glanzer, Adams, Iverson, & Kim, 1993). The consistency with which the mirror effect has been observed with different recognition tasks and with different experimental manipulations led Glanzer et al. to describe it as a “regularity” (p. 546) of recognition memory.

Stretch and Wixted (1998) investigated the causes of the word frequency and the repetition-based mirror ef-

fects and concluded that the latter is a consequence of a shift in the placement of the SDT response criterion. An example of criterion shift is presented in Figure 2. In this case, strong targets (T_s) have been strengthened through study repetitions or increased presentation time so that they have higher strength-of-evidence than weak targets (T_w). In Figure 2, this graphically translates into the T_s distribution being shifted to the right on the strength-of-evidence axis relative to the T_w distribution. The distractor (D) distribution is low on the strength-of-evidence axis because distractors were not presented at study. Only one distractor distribution appears in Figure 2, because it is assumed that the strength manipulation exerts an effect only on targets, as it happens at encoding, and not on distractors. Placing response criteria at the intersection point of the target and distractor distributions for both the weak (C_w) and strong (C_s) conditions (i.e., the point corresponding to the optimal observer, $C = 0$; see Macmillan & Creelman, 2005) creates C_w and C_s , respectively. As can be observed in Figure 2, because of the positioning of the criteria and distributions, more *old* responses are given to T_s (with respect to C_s) than to T_w (with respect to C_w). In addition, fewer *old* responses are given to distractors in the strong condition than in the weak condition. In other words, there is a higher HR and a lower FAR in the strong condition relative to the weak condition—the mirror effect.

D. Bruno, davbruno@psych.umass.edu

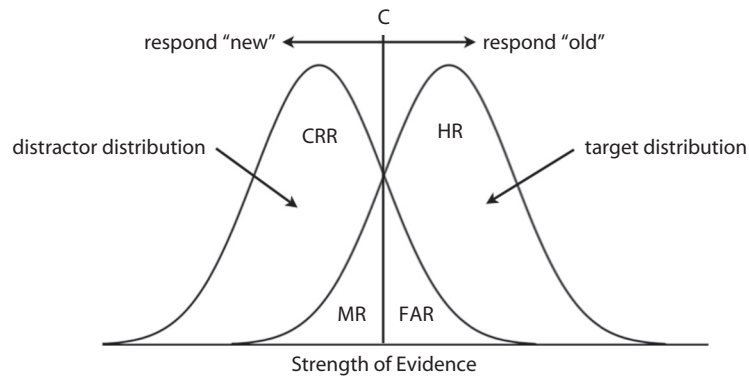


Figure 1. Single-dimension signal detection model for recognition. HR, hit rate; FAR, false alarm rate; MR, miss rate; CRR, correct rejection rate; C, old-new criterion.

Stretch and Wixted (1998) suggested that this explanation of the strength-based mirror effect applied when the strength manipulation occurred between lists. This explanation is rooted in the work of Brown, Lewis, and Monk (1977; Brown, 1976). According to these theorists, participants judge items to be more or less memorable according to several variables, such as the nature of the item itself (e.g., distinctive vs. nondistinctive words), the length of the retention interval, a change in context between study and test, or the number of times a study item is presented. According to Brown et al., if items are judged to be memorable, a strong sense of prior occurrence is expected for them at test. Therefore, recognition performance can be improved (i.e., accuracy will be higher) if the response criterion is placed high on the strength-of-evidence axis relative to if it is placed lower down, mostly because the FAR will decrease substantially.

In the study phase of their first experiment, Stretch and Wixted (1998) presented half of the participants with a list of weak items (each word presented once) and the other half with a list of strong items (each word presented three times). Also, half of the items in each list were high-frequency words and the other half were low-frequency words. The participants were then administered an *old-new* recognition test consisting of low- and high-frequency targets and distractors. The study-test block was then repeated, but the participants who were initially assigned to the weak condition were now assigned to the strong condition and vice versa. Stretch and Wixted predicted that the response criterion would be lower following a weak list than following a strong list, because weak items would be judged as harder to remember (i.e., less memorable). Consistent with this prediction, a mirror effect was found. Because the strength manipulation occurred at study and did not affect the distractor items, Stretch and Wixted assumed that their results were an example of a criterion-shift mirror effect. Indeed, a mirror effect based on different distractor distributions was not deemed possible, given that the strong and weak distractors were “physically identical” (p. 1384) and, hence, should not occupy different positions on the strength-of-evidence dimension.

The mirror effect is observed regularly when strength is manipulated between lists (i.e., strong and weak items are studied in separate lists). In contrast, when strength is manipulated within lists (i.e., strong and weak items are studied and tested together in randomly mixed lists), a mirror effect is rarely observed. According to Stretch and Wixted (1998), participants are unlikely to shift their response criterion on a trial-by-trial basis according to whether a strong or weak item is presented at test. Consistent with this prediction, no mirror effect was observed in any of Stretch and Wixted’s experiments when strength was manipulated within lists. For example, in their Experiment 5, participants studied a list of words, half of which were presented five times in one color (strong condition), and half of which were presented only once in another color (weak condition). The instructions were very clear about the difference in strength between words and also about the association between strength levels and colors. At test, the participants were presented with targets and distractors, and both classes of items were associated with the same colors as at study (distractors in the strong and weak

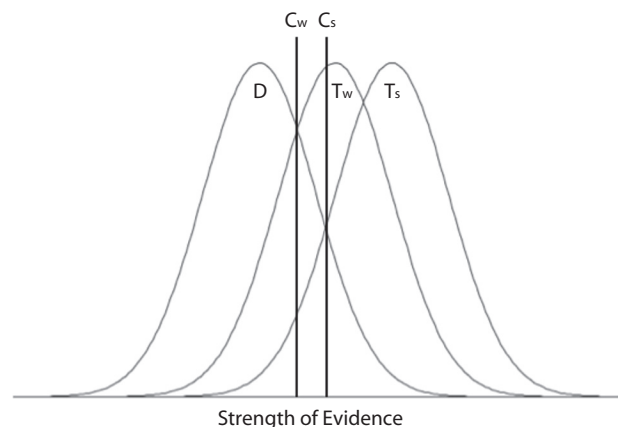


Figure 2. Criterion-shift model of the strength-based mirror effect. D, distractor distribution; C_w , response criterion for weak items; C_s , response criterion for strong items; T_w , weak target distribution; T_s , strong target distribution.

conditions were in the same colors as targets in the strong and weak conditions, respectively). The results showed no strength-based mirror effect; although recognition performance was better for strong words than for weak words as measured by d' , this effect came about because of a difference in the HR between the strength conditions, not because of an FAR difference.

Since Stretch and Wixted's (1998) crucial experiments were published, the strength-based, within-list mirror effect has remained elusive. Morrell, Gaitan, and Wixted (2002, Experiment 1), for example, presented participants with words belonging to two different semantic categories (geographical locations and professions) at study and strengthened words in only one of the categories by repeating them five times. Similar to Stretch and Wixted, no mirror effect was detected: The HR was greater in the strong condition than in the weak condition, but no difference was observed in the FAR between strong distractors from one category and weak distractors from the other.¹ Morrell et al. also increased the salience of the strength manipulation by replacing the list of words representing geographical locations with a list of pictures representing animals (their Experiments 2 and 3), but again, no mirror effect was observed; only the HR differed between the strength conditions. More recently, Higham, Perfect, and Bruno (2009) manipulated strength within lists and additionally used labels at test to indicate whether each test item belonged to the strong or weak condition. These labels ensured that the participants did not forget the association between the strength-defining cue (semantic categories) and strength. Again, however, strength affected only the HR. As in previous research, no effect was observed on the FAR. Stretch and Wixted, Morrell et al., and Higham et al. all concluded that participants typically use overall information about the strength of the items in the test list (i.e., they assess the difference in memorability between strong and weak items) to set an initial decision criterion, but they fail to adjust their criterion on a trial-by-trial basis during test, regardless of variations in the strength of the items.

However, recently, other researchers have reported evidence suggesting that criterion shifts can occur within lists under some circumstances (e.g., Dobbins & Kroll, 2005; Han & Dobbins, 2008; Hockley & Niewiadomski, 2007; Rhodes & Jacoby, 2007; Singer & Wixted, 2006; Verde & Rotello, 2007). For example, Rhodes and Jacoby presented items at test in one of two locations on a screen, with the majority on one side of the screen being old items and the majority on the other side being new. Across three studies, they found that participants shifted their response criterion depending on where an item appeared at test. Dobbins and Kroll constructed lists of items that varied in pre-experimental familiarity, by intermixing photographs of scenes from the participants' home campus and other locations. They found a mirror effect, with a higher HR and lower FAR for familiar scenes. However, they also found that the FAR portion of the mirror effect disappeared when the participants were forced to respond rapidly, or with a 1-week delay prior to the final test. Singer and Wixted

had participants study items from categorized lists (e.g., birds, body parts), with a delay inserted between two study phases, thereby creating strong items (recently studied) and weak items (studied prior to the delay). Strong and weak items appeared on the same test list. With shorter delays (up to 40 min), there was no difference in FAR between the two classes of items. However, for a delay of 2 days between the studied lists, there was a clear separation of FARs for strong and weak items. Verde and Rotello reported a series of studies in which participants studied strong and weak items before an ordered test in which the participants were tested on strong and weak items blocked within a single test list. In their first four studies, their strength manipulations consistently impacted the HR, but not the FAR, as in Stretch and Wixted (1998). However, in their final study, participants received accuracy feedback on a trial-by-trial basis throughout the test. Under these conditions, the participants demonstrated a mirror effect, with lower FAR and higher HR for the strong items.

These demonstrations of differential FAR across classes of items contrast markedly with the failure to demonstrate such a shift based on study repetitions manipulated within a list (Higham et al., 2009; Morrell et al., 2002; Stretch & Wixted, 1998). The aim of the present work is to determine what underpins these apparently discrepant findings. To achieve this aim, it is imperative to bridge the methodological differences between the studies that have produced the varying outcomes. To date, there have been no published demonstrations of within-list mirror effects under experimental conditions similar to those in Stretch and Wixted, Morrell et al., and Higham et al.—that is, when strength is manipulated by repetition, strong and weak items are randomly ordered in the test list, and there is no test feedback. Here, we will present demonstrations of just such a mirror effect pattern in exactly these circumstances.

A key influence on the methodology adopted in the first two experiments reported here was a series of unpublished studies by Tam (2006) investigating recognition memory for words and nonwords. When Tam combined the use of nonwords with a manipulation of strength (repetition) within lists, she observed a strength-based mirror effect with both real words and nonwords. Aside from the presence of the nonwords, the conditions of Tam's study closely resembled those of Stretch and Wixted's (1998), who found no such effect. Here, we attempted to replicate these results in an effort to uncover and identify the important factors that moderate the presence of the mirror effect.

The studies reviewed above collectively demonstrate that people can adjust their response criterion within a list, but that they often do not. In their recent discussion of this literature, Rhodes and Jacoby (2007) argued that participants may not adjust their response criterion because they fail to see the relevance of doing so, even though all the necessary information is available (e.g., in Stretch & Wixted, 1998, strong and weak items were presented in different colors). Participants may view the adjustment of the response criterion as an unnecessary effort that would not benefit their performance. When the relevance of the strength-related dimension is made clear through feed-

back (e.g., Verde & Rotello, 2007) or gross differences between classes of items (e.g., Singer & Wixted, 2006), however, a response criterion shift is observed. It is reasonable, then, to expect that participants can be made to see the advantage of a trial-by-trial adjustment of their own response criteria under certain conditions. In particular, here, we test the idea that one critical predictor is the participants' perception of their own memory for the study list. Specifically, when participants view their memory for the study list as good, there is no motivation or incentive to attend to the nonmemorial cues that indicate strength (e.g., text color), a task that could be effortful and time consuming. Instead, they rely predominantly on memory strength to make recognition judgments, adopting a single criterion throughout the test list. Under such circumstances, a person may judge that lack of memory information is diagnostic of the item's novelty. In comparison, when participants perceive their memory for the study list to be poor, they may be motivated to search for assistance in the recognition task. One potential source of assistance is the nonmemorial cues designating strength (e.g., strong and weak words being presented in different colors). For instance, if the cues indicate that an item belongs to the weak condition, the participants may adjust their response criterion to have low stringency, reasoning that a weak item could easily have been forgotten. Such a strategy would maintain a high HR for the weak targets, despite a perceived lack of memory for the study list. In contrast, if the cues designate high strength, the participants can afford to be more stringent in their responding without compromising performance, despite the fact that the overall conditions for memory performance are perceived to be poor. Although effortful, adoption of different response criteria for strong and weak items would be motivated by a desire to minimize recognition errors under conditions that are perceived to be challenging and would result in a mirror effect.

Because we believe that the incorporation of the nonmemorial strength cues into their recognition judgments stems from the participants' perception of their memory for the study list, we refer to this idea as the *global subjective memorability* (GSM) hypothesis. If this hypothesis has any validity, the reason that Tam (2006) found a mirror effect in her study was that the presence of nonwords in the study list made the participants experience the memory task as difficult, which triggered the utilization of the available test cues and, so, adoption of more than one response criterion at test.

Purpose and Overview of the Experiments

The purpose of this article is to study conditions in which a strength-based mirror effect can be observed within lists and to identify the key factors that predict when such an effect will be observed. Specifically, we are interested in investigating which factors have the effect of soliciting an adjustment of the response criterion during test. Three experiments are reported in this article. In all of the experiments, strength was manipulated within lists by presenting the items either thrice or once at study. The study phase was followed by a test phase in which

strong targets, strong distractors, weak targets, and weak distractors were randomly intermixed as in Stretch and Wixted (1998), Higham et al. (2009), and Morrell et al. (2002). Following Tam (2006), we mixed real words with nonwords in Experiments 1 and 2. Our expectation was that the inclusion of nonwords in the study and test lists would decrease subjective memorability of the study list. Wixted (1992), for instance, showed that participants tended to rate extremely rare words (words with frequency of occurrence of less than once per 7 million) as low in memorability: The median memorability rating on a 1–10 scale for rare words was 4.6, as opposed to 6.4 for low-frequency words and 8.0 for high-frequency words (Wixted, 1992, Experiment 3). (Because the participants were very unlikely to have encountered these rare words prior to the experiment and almost certainly did not know their meaning, the rare words were effectively nonwords.) In Experiment 3, we manipulated subjective memorability of the study list by varying the length of item presentation at study. In each experiment, the key point of interest was the existence of a within-list mirror effect.

EXPERIMENT 1

In Experiment 1, strength (via repetition) was manipulated within lists (i.e., strong targets were presented thrice, and weak targets were presented once), and the word list was made up of 144 words and 40 nonwords. So that participants could discriminate between strong and weak items, all of the real words within each strength level belonged to the same semantic category, as in Morrell et al. (2002) and Higham et al. (2009). Additionally, items were labeled at test to specify their strength category (cf. Higham et al., 2009; Tam, 2006). Although nonwords were included at study and test, recognition performance for them was not analyzed, because our theoretical focus was on the mirror effect for the words. The presence of the nonwords was intended only to reduce overall subjective memorability for the list. Our expectation was that if the mirror effect observed in Singer and Wixted (2006) and Tam was due to overall subjective memorability, it should also emerge here.

Method

Participants. The participants were 40 first-year undergraduate students from the University of Southampton. They took part in the study in a single group at the end of an introductory psychology lecture. They received course credit for their participation or a payment of £5.

Materials. A list of 144 real English words, taken from the MRC psycholinguistic database, was used for the experiment. All of the words were common English nouns, between three and eight letters in length, with a Thorndike–Lorge frequency between 50 and 1,000. Half of the words represented living items (e.g., *parsley*, *goat*), whereas the other half represented nonliving items (e.g., *spoon*, *clock*). In addition, 40 nonwords were selected from Whittlesea and Williams (2000; Tam, 2006) to be mixed in with the real words. As the nonwords only represented a portion of the overall list and mainly served the purpose of reducing the GSM level of the study list, recognition performance for these items was not analyzed.

Design. The design was a 2 (prior presentation: target, distractor) \times 2 (strength: strong, weak) factorial with both factors manipu-

lated within subjects. Ninety-two different items were presented at study. Thirty-six of these items were real words presented thrice (strong condition), 36 were real words presented once (weak condition), and 20 were nonwords presented once, for a total of 164 study trials. For half of the participants, all *living* real words were presented thrice, whereas all *nonliving* real words were presented once. This strength-item type association was reversed for the other half of the participants. The study items were printed in capital letters and in a random order on a single sheet of paper. The words were organized into five columns containing 33 items each, except the far right column, which contained 32 items.

At test, there were 92 targets and 92 distractors. Of the 92 targets, 36 were strong real words, 36 were weak real words, and 20 were nonwords. All real-word targets within a given strength category were either *living* or *nonliving*, which varied according to the counterbalance condition. Consequently, it was possible to use the *living-nonliving* distinction to define strong and weak distractors. Of the 92 distractors, 36 were strong real words, 36 were weak real words, and 20 were nonwords. The test items were printed in capital letters and in a random order on five separate sheets of paper containing 41 items each, except for the final sheet, which contained 20 items.

Four different study lists and two different test lists were used to rotate items through the *living/nonliving*, *old/new*, and strong/weak combinations. A different random order was used for each study and test format. To highlight the strength category distinction for distractors, the test items in the weak and strong conditions were accompanied by different cues. In particular, the participants were given the choice to circle either "1" (presented once) or "0" (*new*) for both targets and distractors in the weak condition, whereas the choices were "3" (presented thrice) or "0" (*new*) for both targets and distractors in the strong condition (Tam, 2006). Space was also provided next to each test item for a 1–6 confidence rating (1, *low*; 6, *high*) about the accuracy of the recognition decision.

Procedure. The participants were provided with one of eight experimental booklets, which contained a consent form, study instructions, a study list, test instructions, and a test list, in that order. The study instructions were as follows:

You are about to take part in an experiment on recognition memory. A list of items is printed on the next page (DON'T turn the page yet). You will be given six minutes to study the list, approximately two seconds per item. Most of the items are real English words, but others are letter strings that look like real words, but aren't actually legal English words (nonwords). Some of the real words will be repeated, whereas other real words only appear once. All of the nonwords only appear once. Your job during the study phase of the experiment is to study each item and to commit it to memory. To do so, work systematically through the list, row by row, pronouncing each item to yourself. Don't spend more than two seconds on any of the items or you will run out of time before studying the whole list. If you finish studying the list before the time is up, start again at the beginning and work systematically through the list again.

The participants were then informed that the real English words represented either something living or something nonliving. Half of the participants were informed that *living* and *nonliving* words would appear once and thrice, respectively, whereas this was reversed for the other half of the participants.²

After completion of the study phase, the participants were requested to read the test instructions, which were as follows (with "X" replaced by "living" and "Y" replaced by "nonliving" for half of the participants and the reverse for the other half):

Some test items are printed on the following pages. Your study list had the X items presented 3 times and the Y items presented 1 time. Nonwords were also presented only 1 time. So For each item on the test list you have to decide either:

For X Items: I saw the word 3 times (3) OR not at all (0)

For Y Items: I saw the word 1 time (1) OR not at all (0)

For nonwords: I saw the nonword 1 time (1) OR not at all (0)

You will now have to decide whether you saw an item 1 time or 3 times. Please make your response by circling the number corresponding to the number of times you saw the item (either "0," "1," or "3"; see examples below). After you give your recognition answer for a word please rate your confidence that the decision you made is correct. To do this, enter any integer value between 1 and 6 (i.e., 1, 2, 3, 4, 5 or 6) in the space provided. Use the following as a guide:

1 means you are *not at all* confident that your decision was correct.

3 or 4 means you are *somewhat* confident that your decision was correct (4 indicating a little more confident than 3).

6 means you are *very* confident that your decision was correct.

Please note that you can be just as confident about a "0" decision as you might be about a "1" or "3" decision. For example, you are probably very confident that your name did not appear on the study list (which it didn't!), because you would have remembered it if it did. So, if one of the test words was your name, you would respond "new" to it on the test, and rate your confidence as "6."

The instruction sheet also contained examples of correct and incorrect methods of filling out the required information for each item. After all of the participants had finished reading the instructions, they were permitted to complete the test at their own pace.

Results and Discussion

The mean HR and FAR were calculated from the number of positive responses ("1" or "3") given to test items. Also, the SDT measure of discrimination (d') was calculated from the HR and FAR (Macmillan & Creelman, 2005). However, since d' is undefined with extreme values (i.e., 1 and 0), it was calculated after applying Snodgrass and Corwin's (1988) transformation to HR and FAR. On the other hand, for statistical tests conducted directly on HRs and FARs, raw scores were used. The untransformed estimates of HR and FAR, and d' derived from the transformed estimates of HR and FAR, are reported in Table 1 as function of strength.

Single-factor repeated measures ANOVAs were conducted on the HR, FAR, and d' . HR and d' were higher in the strong condition than in the weak condition [$F(1,39) = 66.118$, $MS_e = .008$, $p < .001$, $\eta^2 = .629$, and $F(1,39) = 36.529$, $MS_e = .230$, $p < .001$, $\eta^2 = .484$, respectively]. More important, there was no effect of strength on the FAR ($F < 1$). Although strength exerted an effect on the HR, it had no effect on the FAR. This result replicated

Table 1
Mean HR, FAR, and d' and Standard Deviations by Strength Level and Type of Item in Experiment 1

Measure	Strength			
	Strong		Weak	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
HR	0.822	.098	0.658	.178
FAR	0.174	.159	0.170	.110
d'	2.213	.923	1.565	.773

Note—HR, hit rate; FAR, false alarm rate; d' , discrimination. The mean HR and FAR for once-presented nonwords were .723 and .171, respectively.

those of previous research in which strength was manipulated with repetition within lists (e.g., Higham et al., 2009; Morrell et al., 2002; Stretch & Wixted, 1998).

Two mutually exclusive accounts of this null effect on FAR are possible. One is that overall subjective memorability does not drive the appearance of the mirror effect, contrary to our GSM hypothesis. The alternate possibility is that we failed to lower overall subjective memorability sufficiently to trigger a shift in response strategy. We collected confidence-in-accuracy ratings for recognition decisions, and these data may be useful for testing the GSM hypothesis, in that they can be collapsed over experimental conditions to provide a measure of participants' confidence about the list as whole. However, there is no clear prediction associated with a single-point estimate of subjective memorability, and so we reserve discussion on this point until we have reported the data from Experiment 2. In Experiment 2, we sought to test the GSM hypothesis once again by substantially increasing the proportion of nonwords in the list. Our expectation, in line with the GSM hypothesis, is that this increase should reduce subjective memorability overall and that this may be sufficient to trigger a strategy shift and produce a mirror effect. In contrast, the GSM hypothesis would be considerably weakened if overall subjective memorability were significantly lowered without an impact on the FAR rate.

EXPERIMENT 2

The aim of Experiment 2 was to reduce the overall subjective memorability of the study list by increasing the proportion of study trials involving nonwords. The strength manipulation was not applied to nonwords in Experiment 1, so if this design feature were maintained in Experiment 2, it would have meant that a large proportion of trials would not have been analyzed. Consequently, we decided to strengthen half of the nonwords with repetition, consistent with the treatment of the words. This had the benefit of enabling us to test the existence of a within-list mirror effect for the nonwords, as well as for the words (cf. Tam, 2006).

In Experiment 2, 120 items were used overall: Forty of these were real words and 80 were nonwords. However, given the large number of nonwords, the semantic association between *living* and *nonliving* categories and strength levels (cf. Experiment 1) was replaced by a perceptual association. Strong items were presented at test in italics (e.g., *CAR*, *BROOT*) and weak items were underlined (e.g., CALIDON, PEOPLE) for half of the participants, and this association was reversed for the other half.

Method

Participants. Twenty-three undergraduate students from the University of Southampton took part in this study: Eleven received academic credits for their participation, and 12 received a payment of £5. The participants were tested in group sizes between 1 and 7.

Materials. A list of 120 items adapted from Whittlesea and Williams (2000) was used in the experiment. Forty of these items were real words and 80 were nonwords.

Design. The experiment had a 2 (prior presentation: target distractor) \times 2 (strength: strong, weak) \times 2 (item type: real words, non-

words) factorial design. At study, 60 items were presented: Twenty of these items were real words and 40 were nonwords. Half of each target set (10 words and 20 nonwords) was presented only once, whereas the other half was presented three times, which resulted in 120 study trials. The study items were printed in capital letters and in a random order on a single sheet of paper. The items were organized into five columns containing 24 items each.

At test, 120 items were presented. Sixty were the targets shown at study, and 60 were distractors; both were divided into 20 real words and 40 nonwords. For half of the participants, strong targets and distractors were presented in italics, and weak targets and distractors were underlined, whereas this association was reversed for the other half of the participants. The participants were informed about this new association between the perceptual cue and strength in the same way that the participants in Experiment 1 were reminded at test about the association between the semantic cue (*living/nonliving*) and strength. The test items were printed in capital letters in a random order on three separate sheets of paper containing 43 items each, except for the final sheet, which contained 34 items.

Four different study lists and four different test lists were used to rotate items through the italics/underlined, *old/new*, and strong/weak combinations. A different random order was used for each study and test format. To highlight the strength–category distinction for distractors, test items in the weak and strong conditions were accompanied by the same cues used in Experiment 1. Space was provided next to each test item for a 1–6 confidence rating (1, *low*; 6, *high*) about the accuracy of the recognition decision.

Procedure. The procedure of Experiment 2 duplicated that used in Experiment 1.

Results and Discussion

The HR, FAR, and d' for both words and nonwords were calculated in the same manner as in Experiment 1 and are reported in Table 2 as function of strength and item type. Separate 2 (strength: strong, weak) \times 2 (item type: words, nonwords) repeated measures ANOVAs were conducted on HR, FAR, and d' . For the analysis on the HR, the effects of both strength [$F(1,22) = 43.282$, $MS_e = .017$, $p < .001$, $\eta^2 = .666$] and item type [$F(1,22) = 5.570$, $p = .028$, $\eta^2 = .202$] were significant, showing greater HR for strong targets and real words than for weak targets and nonwords, respectively. Follow-up analyses indicated that the effect of strength on the HR was significant for both words [$F(1,22) = 21.471$, $MS_e = .017$, $p < .001$] and nonwords [$F(1,22) = 34.588$, $MS_e = .017$, $p < .001$]. The analysis on the FAR revealed that a mirror effect was obtained; that is, the FAR was greater for weak distractors than for strong distractors [$F(1,22) = 15.891$, $MS_e = .002$, $p = .001$, $\eta^2 = .419$]. Follow-up analyses indicated that the effect of strength on the FAR was significant for both words [$F(1,22) = 43.500$, $MS_e = .002$, $p < .001$] and nonwords [$F(1,22) = 41.500$, $MS_e = .001$, $p < .001$]. Consis-

Table 2
Mean HR, FAR, and d' and Standard Deviations by Strength Level and Type of Item in Experiment 2

Measure	Words				Nonwords			
	Strong		Weak		Strong		Weak	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
HR	0.870	.166	0.691	.200	0.785	.195	0.559	.189
FAR	0.139	.162	0.226	.209	0.185	.169	0.270	.155
d'	2.192	.721	1.302	.625	1.822	.710	0.793	.690

Note—HR, hit rate; FAR, false alarm rate; d' , discrimination.

tent with the observed mirror effect, better discrimination was detected at higher strength [$F(1,22) = 80.918, MS_e = .303, p < .001, \eta^2 = .786$]. A main effect of item type from the ANOVA on d' also indicated that discrimination was better for real words than for nonwords [$F(1,22) = 8.984, p = .007, \eta^2 = .290$]. No other main effect or interaction was significant for any of these ANOVAs [largest $F(1,22) = 2.903, p = .103, \eta^2 = .117$].

Comparison of Experiments 1 and 2. The presence of a strength effect on FAR in Experiment 2 is consistent with the GSM hypothesis. However, the GSM hypothesis requires that this shift be the result of an overall decline in subjective memorability for the study list. In order to test this prediction, we needed an overall measure of subjective memorability for the entire list in each experiment. We took the mean confidence ratings assigned across the entire list as our index of subjective memorability for both experiments. In line with the GSM hypothesis, overall confidence was higher in Experiment 1 ($M = 4.40, SD = .606$) than in Experiment 2 ($M = 4.01, SD = .636$) [$t(61) = 2.408, p = .019$].

Thus, the overall pattern from Experiment 2, in particular the comparison with Experiment 1, is consistent with the GSM hypothesis. In line with the work of Tam (2006), we observed a within-list strength-based mirror effect, even though the participants were made aware of the strength status of items only at test (cf. Stretch & Wixted, 1998). The participants in Experiment 2, but not those in Experiment 1, showed an elevated FAR for weaker items, indicative of an item-by-item change of response criterion. In line with the GSM hypothesis, overall subjective memorability for the list was lower in Experiment 2 than in Experiment 1.

If the GSM account is correct, the same pattern should be obtained with any manipulation that reduces subjective memorability enough to trigger a shift in response strategy. In particular, the mirror effect should not be due to the presence of nonwords in the list. Consequently, in Experiment 3, we manipulated the subjective memorability of the list by manipulating the study time per item at encoding.

EXPERIMENT 3

In Experiments 1 and 2, support for the GSM hypothesis was based on a comparison of performances between experiments and by varying the ratio of words to nonwords. However, the GSM hypothesis is not specifically tied to the use of nonwords, and so, in Experiment 3, we manipulated overall subjective memorability within a single experiment involving only real words.

We also took this opportunity to address other methodological differences between Experiments 1 and 2 that might account for the differing results. In Experiment 1, where no mirror effect was found, item strength was associated with *semantic* category membership (*living* or *nonliving*) and this association was introduced at *study*. Similarly, semantic strength-defining cues were introduced at *study* in Morrell et al. (2002), and they too found

no mirror effects. In contrast, in Experiment 2, where we observed mirror effects, strength was associated with *perceptual* cues (italics or underlining), and these cues were only introduced at *test*. Tam (2006) also introduced her perceptual strength-defining cues (colors) at *test*, and she also found mirror effects. Thus, the discrepant findings between Experiments 1 and 2 (and between Morrell et al. on one hand and Tam on the other) could be due to (1) the fact that the property of the item that is associated with the strength manipulation is critical (i.e., perceptual properties produce the mirror effects whereas semantic properties do not), (2) the timing of the introduction of the cues (i.e., cues introduced at *test* produce the effect, whereas cues introduced at *study* do not), or (3) both of these factors. Potentially, perceptual cues and/or new cues appearing for the first time at *test* are highly salient, causing them into their recognition judgments. To test this idea, in Experiment 3, we reverted to the methodology used in Experiment 1. If the GSM hypothesis is correct, the mirror effect should emerge under conditions of low GSM, regardless of the timing or nature of the strength cues. Conversely, if it is either the timing or the nature (or both) of the cues that is critical, and not GSM, no mirror effect should occur in any condition.³

The variable chosen for the purpose of manipulating GSM in Experiment 3 was the length of item presentation at *study* (presentation rate). Because presentation rate also acts as a strength variable in its own right (e.g., Hockley & Niewiadomski, 2007), it allowed us to test for a between-list mirror effect based on it, as well as the within-list mirror effect based on repetition. A longer presentation rate (e.g., 3 sec/item) should create higher subjective memorability than should a shorter presentation rate (e.g., 0.5 sec/item). Thus, if the GSM hypothesis is correct, a within-list repetition-based mirror effect should be more likely in the condition in which the study list was presented at the faster rate. Since only real words were used in Experiment 3, the semantic association between strength levels and *living/nonliving* categories used in Experiment 1 was reintroduced, so that the participants could benefit from an added cue to item strength.

Method

Participants. The participants were 32 undergraduate students from the University of Southampton. Twenty-two received a payment of £5 for their time, and 10 received course credit. Sixteen were randomly assigned to each of the long and short presentation groups.

Materials. The same 144 real words used in Experiment 1 were used again in Experiment 3.

Design. The experiment used a 2 (prior presentation: target, distractor) \times 2 (strength: strong, weak) \times 2 (presentation rate: long, short) mixed design, with only the presentation rate varied between subjects. The participants in the long and short presentation rate groups were presented with words at *study* for 3 sec and 0.5 sec each, respectively. Only one presentation rate was used per group. Four different study lists were created, each with a dif-

ferent random selection of 36 words from each of the 72 *living* and 72 *nonliving* words. For two of the study lists, *living* words were presented thrice (strong condition) and *nonliving* words were presented once (weak condition), whereas this strength–semantic–category association was reversed for the other two lists. An equal number of participants viewed each study list. At test, all 144 items were presented, of which 72 were targets and 72 were distractors. Four different test lists were used, each corresponding to a given study list. For the 72 distractors on each test, the 36 remaining words from one semantic category (e.g., *living*) were assigned to the strong condition and 36 from the other category (e.g., *nonliving*) were assigned to the weak condition. It was ensured with this assignment that all targets and distractors within a given strength condition were either *living* or *nonliving*, which depended on the counterbalance condition. A different random order of presentation was used for each study and test list.

Procedure. The procedure of Experiment 3 was the same as that of Experiment 1, with the following exceptions. The words were presented to the participants as a slide-show presentation on the screen of a Macintosh computer. The words appeared in the center of the screen, in black 48-point font on a white background. The participants in the long and short presentation rate groups studied each word for 3 and 0.5 sec, respectively. The interstimulus interval was 1 sec in both groups. All timing was controlled by the computer. At test, the words appeared on the computer screen one at a time. The weak test items were accompanied by a cue either to select a button (using the mouse) at the bottom right of the computer screen if the item was new or to select a button at the bottom left of the computer screen if the item had been presented once. The choices for strong test items were *new* and *presented three times*. The same 6-point confidence scale was used, but the responses were collected by requiring the participants to click one of six radio buttons at the bottom of the screen.

Results and Discussion

The HR, FAR, and d' were calculated as in the previous experiments and are reported in Table 3 as function of presentation rate and repetition. A 2 (presentation rate: long, short) \times 2 (repetition: thrice, once) mixed ANOVA conducted on the HR revealed a significant main effect of repetition [$F(1,30) = 51.967, MS_e = .009, p < .001, \eta^2 = .634$], indicating a higher HR for thrice-presented targets than for once-presented targets. The analysis also revealed a marginal effect of presentation rate [$F(1,30) = 3.885,$

$MS_e = .044, p = .058, \eta^2 = .115$]. The interaction was not significant ($F < 1$).

An analogous ANOVA conducted on the FAR revealed both a main effect of repetition [$F(1,30) = 9.201, MS_e = .008, p = .005, \eta^2 = .235$] and a main effect of presentation rate [$F(1,30) = 4.971, MS_e = .035, p = .033, \eta^2 = .142$]. These main effects were qualified by a significant interaction [$F(1,30) = 5.355, MS_e = .008, p = .028, \eta^2 = .151$]. Pairwise comparisons showed that there was no effect of repetition at the long presentation rate ($F < 1$; strong, $M = .109, SD = .121$; weak, $M = .125, SD = .112$) but that there was one at the short presentation rate [$F(1,15) = 13.5, MS_e = .008, p = .002$; strong, $M = .163, SD = .179$; weak, $M = .279, SD = .160$].

An analogous ANOVA conducted on d' indicated both a main effect of repetition [$F(1,30) = 86.544, MS_e = .212, p < .001, \eta^2 = .743$] and a main effect of presentation rate [$F(1,30) = 7.605, MS_e = .1440, p = .010, \eta^2 = .202$]. The interaction was marginally significant [$F(1,30) = 3.591, MS_e = .212, p = .068, \eta^2 = .107$]. The main effects indicated better discrimination for thrice-presented than for once-presented items and better discrimination when the presentation rate was long than when it was short.

Confidence ratings were again analyzed to establish whether a mirror effect emerged only in low GSM conditions, as was predicted by the GSM hypothesis. As previously, mean confidence across all items in the test list was computed. As predicted, mean confidence was marginally higher in the long presentation group ($M = 4.75, SD = .448$) than in the short presentation group ($M = 4.41, SD = .540$) [$t(30) = 1.919, p = .066$].

As anticipated, a repetition-based mirror effect occurred within lists in the short presentation group of Experiment 3, but not in the long presentation group. This pattern is consistent with the idea that the subjective memorability conditions of the experiment drive the likelihood of obtaining a within-list mirror effect. With such a mechanism, the fact that presentation rate moderated the repetition-based mirror effect can be easily explained: The low GSM in the short presentation group caused the participants to attend to available test cues and to adjust their response criteria accordingly, whereas, in the high GSM condition, the test cues were mostly ignored and no criterion shift was effected. Crucially, these data make clear that the previous results from Experiments 1 and 2 and from Tam (2006) are not simply due to the presence of nonwords in the study list. The results from the short presentation group also make clear that a within-list mirror effect can be observed when strength cues are introduced at study and when the property associated with strength is semantic. Thus, the results of Experiment 3 indicate that the failure to observe a mirror effect in Experiment 1 could not have been due solely to the nature of the strength cues or to when the strength cues were introduced.

GENERAL DISCUSSION

To summarize our results, a within-list, strength-based mirror effect was not obtained for words in Experiment 1 when only a small portion of the study trials (12%) and the

Table 3
Mean HR, FAR, and d' and Standard Deviations by Strength Level and Presentation Rate in Experiment 3

Measure	Long Presentation				Short Presentation			
	Thrice		Once		Thrice		Once	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
HR	0.920	.082	0.759	.182	0.828	0.185	0.644	.179
FAR	0.109	.121	0.125	.112	0.163	0.179	0.279	.160
d'	2.926	.954	2.073	.888	2.317	1.131	1.028	.570

Note—HR, hit rate; FAR, false alarm rate; d' , discrimination.

test trials (22%) continued nonwords. However, in Experiment 2, when the majority of the study and test trials were nonwords (both 67%), a mirror effect occurred within lists for both words and nonwords. Finally, in Experiment 3, a mirror effect occurred in the short but not the long presentation group, even though no nonwords were included in any of the lists. For the most part, across experiments (although see discussion below), list-wide confidence in the accuracy of recognition judgments (i.e., confidence averaged across the test list as a whole) was lower in the conditions that produced the mirror effect than in the conditions in which no mirror effect was obtained.

The Global Subjective Memorability Hypothesis

We have hypothesized that the subjective memorability of the study list as a whole partially determines whether a strength-based mirror effect will occur within lists, which we have referred to as the GSM hypothesis. Our argument has been that GSM affects the strategies that participants adopt prior to taking the recognition test. In this section, we will outline the GSM hypothesis in greater detail.

At the most basic level, the question of whether or not a mirror effect occurs within lists comes down to the extent to which participants attend to and use the test cues designating strength (e.g., semantic category, color, italics/underlining, etc.). If these cues are ignored, strong and weak distractors will be treated similarly, and a single FAR will obtain. Conversely, if the cues are attended to, there is a potential that the response criterion will be adjusted accordingly.

According to the GSM hypothesis, if GSM of the study list is high (i.e., the study list is correctly, or incorrectly, considered easy to learn), the cues are ignored, and all test items are judged according to a single default response strategy based on memory strength. The participants will view their memory for the study list as good enough to guarantee successful performance on the basis of memory strength alone, so they do not engage in the effortful task of attending to the nonmemorial strength cues to adjust their response strategies. Therefore, both strong and weak test items (as designated by the nonmemorial strength cues) that are low on the strength-of-evidence dimension (predominantly distractors) are judged to be *new* because their low strength is considered diagnostic of their novelty. Under such circumstances, the participants can be said to be using the *metacognitive* strategy as discussed by Strack and Bless (1994), producing a single low FAR for strong and weak distractors and no mirror effect.

On the other hand, in conditions of low GSM (i.e., the study list is correctly, or incorrectly, considered hard to learn), the participants believe that they should not rely solely on memory strength to make recognition judgments, because it is not necessarily diagnostic of prior presentation. For example, under such conditions, forgetting, as well as novelty, can cause low memory strength for a given item. Forgetting is particularly likely for weak items, so the participants are assumed to go to the added trouble of attending to the nonmemorial test cues designating strength (semantic category, color, etc.) to adjust their response criterion on an item-by-item basis. Doing so will minimize

the number of errors on the memory test. In particular, strong items, which are deemed less likely to have been forgotten than are weak items, are judged according to the same default metacognitive strategy that is used under high GSM conditions. Conversely, a more liberal response criterion is adopted for the weak items that are regarded more likely to have been forgotten because they are judged according to the *presuppositional* strategy (Strack & Bless, 1994; i.e., forgetting is presupposed to be likely). Although effortful, attending to the cues in this way will maintain a high HR for weak items even under poor memory conditions. However, the trade-off is that the FAR for weak items will also be high, exceeding the FAR for strong distractors. This result, coupled with an effect of strength on the HR, produces the strength-based mirror effect.

The GSM hypothesis provides a relatively straightforward explanation for why strength-based mirror effects were observed within lists in some of our experiments but not in others. In Experiment 1, although we attempted to create low GSM by including nonwords at study and test (e.g., Chalmers & Humphreys, 1998; Chalmers, Humphreys & Dennis, 1997; Whittlesea & Williams, 1998; Wixted, 1992), there is actually some evidence that by including only a small proportion of such items, they became distinctive, thus potentially increasing, rather than decreasing, the GSM of the list. For example, the nonword FAR (0.171) was the same as the FAR for weak words (0.174), whereas the (once-presented) nonword HR (0.723) exceeded the HR for the (once-presented) weak words (0.658). In contrast, in Experiment 2, the inclusion of a greater proportion of nonwords in the study and test lists had the desired effect: Nonword discrimination was lower than that for words across both strength categories, and overall confidence was lower than that in Experiment 1. As predicted by the GSM hypothesis, this lower confidence and accuracy was associated with within-list mirror effects for both classes of items. Finally, in Experiment 3, in which GSM was manipulated with presentation rate within the same experiment, a mirror effect occurred in the short presentation condition (with low GSM and confidence) but not in the long presentation condition (with high GSM and confidence).

Other Potential Accounts

In this section, we consider some alternative explanations of our results. Singer and Wixted (2006) discussed whether substantial differences in d' might promote within-list criterion shifts, leading to a mirror effect. In their studies, within-list mirror effects occurred only when a long (2-day) retention interval was used, which also corresponded to a large difference in d' between the short and long retention interval conditions. However, despite this result, a broader reading of the literature led them to conclude that large d' differences were neither necessary nor sufficient to promote criterion shifts. Indeed, our own data concur with this conclusion. The average difference in d' between the strong and weak conditions when the mirror effect occurred was 0.94, which was numerically smaller than the average d' difference when the mirror effect was absent (0.97). Singer and Wixted noted that delay

may be a particularly salient variable as far as participants are concerned, being subject to metacognitive reasoning about its effects on memory. Thus, when there is a large difference in the study–test delay’s for the items in the test list, participants attend to the cues designating strength and adjust their response criterion accordingly.

In our view, Singer and Wixted’s (2006) analysis is quite compatible with the GSM hypothesis. In essence, their argument was that it is not overall objective memorability or even between-class differences in objective memorability (i.e., d') that determines whether the strength-defining cues are used to set the criterion during the recognition test. Instead, the critical factor is *subjective* memorability, which is reliant on metacognitive theories about the way that memory works and the variables that influence it. It is intriguing to us that long retention intervals are considered important enough to cause participants to attend to test cues, but that repetition is not, at least under most circumstances. On the basis of our results, repetition will lead to test-cue utilization only if the testing situation is considered particularly challenging and confidence is low, such as when the study list consisted mostly of nonwords (Experiment 2) or when the presentation rate was particularly short (Experiment 3). A fruitful avenue of future research will be to investigate metacognitive theories about recognition memory, because these seem to be the key to understanding when mirror effects will occur within lists and when they will not.

Perhaps it is not surprising that differences in d' between the strength conditions are unable to account for participants’ willingness to incorporate test cues into their recognition judgments. After all, such an explanation assumes that the participants mostly know when they are performing well and when they are not. Although this monitoring ability may be reasonably good, there will certainly be cases when the correlation between overall recognition performance and the assessment of that performance breaks down. Consider, instead, a dual-process memorability account of within-list mirror effects in which the participants monitor *recollection*, rather than the amount of *old–new discrimination*. Unlike simple *old–new* discrimination, recollection has been shown to be associated with excellent metacognitive monitoring (e.g., Higham, 2002; Higham et al., 2009; Payne, Jacoby, & Lambert, 2004); people know when they are recollecting information and when they are failing to do so, except perhaps under circumstances especially designed to produce considerable amounts of false recollection (e.g., Roediger & McDermott, 1995). Under this dual-process hypothesis, if there is a large difference in the amount of recollection between the strength conditions, the participants will attend to the strength-defining cues and a mirror effect will be observed. One rationale for such behavior is that if there is a large amount of recollection for strong targets but very little for weak targets, the strong targets may subjectively stand out from the other (weak) targets and distractors. Hence, attending to the strength cues could be helpful under these circumstances, because any items designated *strong* which do not elicit a large amount of recollection can be easily and confidently rejected.

The main problem with this account is that recollection and recognition discrimination are highly correlated; if recollection is high, it is likely that discrimination will also be, because recollection generally leads to confident, accurate recognition decisions (e.g., Yonelinas, 1994). Although we did not specifically measure recollection, it would be difficult to explain how there was a bigger recollection difference between the strength conditions in our experiments that produced a mirror effect than between the conditions that did not when the corresponding discrimination (d') difference was numerically smaller.

Other explanations of within-list criterion shifts have focused on the importance of feedback (e.g., Han & Dobbins, 2008; Rhodes & Jacoby, 2007; Verde & Rotello, 2007). When feedback on recognition accuracy is provided, participants are more likely to adopt a conservative criterion on easy trials and a liberal one on hard trials and to shift the criterion on a trial-by-trial basis. Although we did not explicitly provide feedback in our experiments, participants presumably spontaneously monitor their own test performance, especially when prompted to provide retrospective confidence judgments about the correctness of their responses. This monitoring process no doubt provides implicit feedback that is more likely to be accurate, and possibly relied on more heavily, when the memory conditions of the experiment are good. If so, the GSM hypothesis prediction is that criterion shifts are more likely under objectively good memory conditions than under poor ones. However, as was noted above, there is no evidence in our research or in the research discussed by Singer and Wixted (2006) that within-list criterion shifts are moderated by d' .

Rhodes and Jacoby (2007) found that, in the absence of feedback, participants were more likely to shift their criterion according to the placement of a test item on the computer monitor that was associated with a particular old-item base-rate probability if that placement had just switched from the previous trial, rather than if it had remained the same. In the context of our experiments, an analogous result would be that strength-based criterion shifts would be more likely if the currently judged item belonged to a different strength condition than the previously judged item. However, we found no evidence for this pattern of results. For example, the participants in the short and long presentation groups of Experiment 3 rated the test items in exactly the same order (four different random orders in each group), yet a mirror effect was observed in the former but not in the latter group.

Finally, it must be considered that a criterion shift is not necessary for the observation of the mirror effect. In contrast to the criterion-shift account of the mirror effect, a differential-distribution explanation argues that the effect occurs because there are two distractor distributions positioned at different points on the strength-of-evidence scale (e.g., Criss, 2006; Criss & McClelland, 2006; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997): The weak distractor distribution is, on average, higher on the strength-of-evidence scale than the strong distractor distribution. Hence, a higher FAR is observed for weak than for strong distractors. However, as was noted, dis-

tractors are not presented at study and not strengthened, and, thus, it is not clear why the distractor distributions should shift as a function of strength. One possibility is that distractors and targets are similar on the basis of an association established at encoding, so that, once a target is strengthened, an associated distractor may also be strengthened (e.g., *raccoon* is a target presented thrice, and *badger* is a potentially strong distractor, since both belong to the same taxonomic category [animals]; Higham et al., 2009). However, the results of Experiment 2 make this explanation less plausible, since no association is established at encoding between targets and distractors of the same strength level. In fact, both cues (i.e., the italics/underlined cue and the test tag) only appeared at test. Therefore, a criterion-shift account of the mirror effect seems the most economical and probable in this case.

Limitations of the GSM Hypothesis

Although we believe that the GSM hypothesis is currently the best available account of our experimental results and of those of the other studies discussed above, it is not without limitations and problems as it currently stands. First, a comparison of mean confidence ratings in Experiments 1 and 3 presents an important anomaly: The average confidence rating in Experiment 1 ($M = 4.40$), where no mirror effect was observed, is nearly identical to the average confidence rating in the short presentation rate group of Experiment 3 ($M = 4.41$), where a clear mirror effect was found.⁴ This finding appears to contradict the claim of the GSM hypothesis that subjective memorability of the study list affects strategic control at test, since participants who are equally confident in their own memory performance produce opposite patterns of behavior.

However, this anomalous result is confounded, in that nonwords were used in Experiment 1 but not in Experiment 3. If the nonword ratings are eliminated from the analysis and only confidence ratings for words are compared between experiments, the confidence data are wholly consistent with the GSM hypothesis: The mean confidence ratings for the conditions where a mirror effect was observed (Experiment 2 and the short presentation rate group in Experiment 3) are lower ($M = 4.43$ and $M = 4.41$, respectively) than the mean confidence ratings for the conditions where mirror effects were not found (in Experiment 1 and in the long presentation rate group of Experiment 3: $M = 4.51$ and $M = 4.75$, respectively).

However, these comparisons, as well as those between Experiments 1 and 2 that include nonwords reported above, must be treated with caution for a number of reasons. First, even if only confidence ratings for words are compared, there are still a number of methodological differences between the three experiments, rendering direct contrasts potentially problematic. For example, Experiments 1 and 2 were conducted using pen-and-paper format, whereas Experiment 3 was computerized. Moreover, the presentation of the study list was massed in Experiments 1 and 2, whereas it was item by item in Experiment 3. Comparing confidence without concern for these methodological differences presupposes an absolutist view

of confidence ratings that ignores the experimental contexts in which the ratings are given. Nonetheless, despite the fact that the present experiments fail to provide definitive evidence in support of the GSM hypothesis, there are no competing theories in the recognition literature that can potentially explain all the vagaries of strength-based criterion shifts within lists in such a parsimonious manner. Consequently, we believe that the GSM hypothesis is at least a reasonable contender and is worthy of further investigation in subsequent research.

This subsequent research might utilize other types of metacognitive judgments to test the GSM hypothesis more fully. For instance, global judgments of study-list memorability could be collected after study and prior to test and/or at various points throughout the test list. These ratings, if paired with retrospective confidence judgments, could be indicative of whether metacognitive/presuppositional strategies are adopted prior to the test or whether there are strategic shifts at certain points during testing. Another possibility is to administer a post hoc questionnaire to enquire more specifically about conscious strategies that the participants might be using. We have not specified whether the adoption of specific response strategies during the recognition test is conscious and deliberate or whether it occurs at a more intuitive, perhaps unconscious, level (Han & Dobbins, 2008). Metacognitive measures, in addition to retrospective confidence ratings, might help to elucidate this issue, along with several others.

CONCLUSION

Despite some limitations outlined above, the GSM hypothesis appears to be the only account that is able to explain our present demonstrations of strength-based mirror effects within lists. The GSM hypothesis suggests that, under subjectively poor memory conditions, people are distrusting of the diagnostic value of the memory information for test items, and they are willing to use nonmemorial test cues to facilitate the recognition decision. This idea draws similarities with the outshining hypothesis proposed by Smith (1988). The reinstatement at test of meaningless contextual cues (e.g., a smell, an item's position on a computer screen) associated with study items has a strong effect on recall, but its effect on recognition is inconsistent. According to the outshining principle, the test probe at recognition is such a strong cue in itself that it outshines the effect of contextual cues, rendering them ineffective at altering performance. In contrast, contextual cues help recall, because no test probes are provided. Similarly, it appears that nonmemorial test cues are outshone by memory information proper in recognition tests with high GSM but are employed by participants in recognition tests with low GSM.

AUTHOR NOTE

Preparation of this article was supported by Grant RES-000-23-1375 from the Economic and Social Science Research Council (ESRC). Portions of this research were presented at the 24th BPS Annual Cognitive Section Conference at the University of Aberdeen, August 20–22, 2007,

and at the 48th Annual Meeting of The Psychonomic Society, Long Beach, CA, November 15–18, 2007. D. Bruno is now at the University of Massachusetts, Amherst, MA. Correspondence concerning this article should be addressed to D. Bruno, Department of Psychology, Tobin Hall, University of Massachusetts, Amherst, MA 01003 (e-mail: davbruno@psych.umass.edu).

REFERENCES

- BROWN, J. (1976). An analysis of recognition and recall and of problems in their comparison. In J. Brown (Ed.), *Recall and recognition* (pp. 1-35). New York: Wiley.
- BROWN, J., LEWIS, V. J., & MONK, A. F. (1977). Memorability, word frequency, and negative recognition. *Quarterly Journal of Experimental Psychology*, **29**, 461-473.
- CHALMERS, K. A., & HUMPHREYS, M. S. (1998). Role of generalized and episode specific memories in the word frequency effect in recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 610-632.
- CHALMERS, K. A., HUMPHREYS, M. S., & DENNIS, S. (1997). A naturalistic study of the word frequency effect in episodic recognition. *Memory & Cognition*, **25**, 780-784.
- 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.
- CRISS, A. H., & MCCLELLAND, J. L. (2006). Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM). *Journal of Memory & Language*, **55**, 447-460.
- 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, & Cognition*, **31**, 1186-1198.
- GLANZER, M., & ADAMS, J. K. (1985). The mirror effect in recognition memory: Data and theory. *Memory & Cognition*, **13**, 8-20.
- GLANZER, M., ADAMS, J. K., IVERSON, G. J., & KIM, K. (1993). The regularities of recognition memory. *Psychological Review*, **100**, 546-567.
- HAN, S., & DOBBINS, I. G. (2008). Examining recognition criterion rigidity during testing using a biased-feedback technique: Evidence for adaptive criterion learning. *Memory & Cognition*, **36**, 703-715.
- HIGHAM, P. A. (2002). Strong cues are not necessarily weak: Thomson and Tulving (1970) and the encoding specificity principle revisited. *Memory & Cognition*, **30**, 67-80.
- HIGHAM, P. A., PERFECT, T. J., & BRUNO, D. (2009). Investigating strength and frequency effects in recognition memory using type-2 signal detection theory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **35**, 57-80.
- 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.
- MACMILLAN, N. A., & CREELMAN, C. D. (2005). *Detection theory: A user's guide*. Mahwah, NJ: 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.
- MORRELL, H. E. R., GAITAN, S., & WIXTED, J. T. (2002). On the nature of the decision axis in signal-detection-based models of recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 1095-1110.
- PAYNE, B. K., JACOBY, L. L., & LAMBERT, A. J. (2004). Memory monitoring and the control of stereotype distortion. *Journal of Experimental Social Psychology*, **40**, 52-64.
- 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.
- ROEDIGER, H. L., III, & McDERMOTT, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 803-814.
- SINGER, M., & WIXTED, J. T. (2006). Effect of delay on recognition decisions: Evidence for a criterion shift. *Memory & Cognition*, **34**, 125-137.
- SHIFFRIN, R. M., & STEYVERS, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, **4**, 145-166.
- SMITH, S. M. (1988). Environmental context-dependent memory. In G. Davies & D. Thomson (Eds.), *Memory in context: Context in memory* (p. 13-33). New York: Wiley.
- SNODGRASS, J. G., & CORWIN, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, **117**, 34-50.
- STRACK, F., & BLESS, H. (1994). Memory for nonoccurrences: Metacognitive and presuppositional strategies. *Journal of Memory & Language*, **33**, 203-217.
- 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, & Cognition*, **24**, 1379-1396.
- TAM, H. H. Y. (2006). *Fluency-based production and memorability-based reduction of false alarms in recognition memory*. Unpublished doctoral dissertation, University of Southampton, UK.
- VERDE, M. F., & ROTELLO, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition*, **35**, 254-262.
- WHITTLESEA, B. W. A., & WILLIAMS, L. D. (1998). Why do strangers feel familiar, but friends don't? A discrepancy-attribution account of feelings of familiarity. *Acta Psychologica*, **98**, 141-165.
- WHITTLESEA, B. W. A., & WILLIAMS, L. D. (2000). The source of feelings of familiarity: The discrepancy-attribution hypothesis. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **26**, 547-565.
- WIXTED, J. T. (1992). Subjective memorability and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **18**, 681-690.
- YONELINAS, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 1341-1354.

NOTES

1. It should be noted that because distractors are not presented at study, they cannot be strengthened. Distractors are only labeled *strong* or *weak* on the basis of their association to strong or weak targets, respectively.

2. We would have preferred to present the study list on an item-by-item basis, which is the more common method in the literature. However, because the participants were tested in a large single group, this method was not possible. The greatest danger with the methodology that we used is that our repetition (strength) manipulation would have been nullified because the participants may have studied once-presented (weak) items more than once or studied thrice-presented items less than three times. To lessen the likelihood of this occurring with any consistency, the items were randomly ordered in the study list, and different random orders were used for different participants. The randomization technique appears to have worked, because, as will become clear, a robust and consistent strength effect was observed on the HR in conditions in which this methodology was used. Nonetheless, to eliminate any potential problems that this methodology might have created, single-item presentation was used in the study phase of Experiment 3.

3. Stretch and Wixted (1998) used perceptual cues to define the strength categories and found no strength-based mirror effects. Consequently, it seems unlikely that it was the perceptual versus semantic nature of the cues that accounts for the differences between the results of Experiments 1 and 2. Nonetheless, it seemed prudent to hold the nature of the cues constant in Experiment 3 to eliminate it as a potential contributor to the observed pattern of results.

4. We thank an anonymous reviewer for noting this comparison.