Journal of Memory and Language 65 (2011) 363–377

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

# Quality trumps quantity at reducing memory errors: Implications for retrieval monitoring and mirror effects

Jason M. Scimeca [a], Ian M. McDonough [b], David A. Gallo [b,*]

[a] Brown University, Providence, RI 02912, USA
[b] University of Chicago, Chicago, IL 60637, USA

ARTICLE INFO

*Article history:*
Received 8 January 2011
revision received 16 April 2011
Available online 25 May 2011

*Keywords:*
False recognition
Mirror effect
Retrieval monitoring
Distinctiveness heuristic

ABSTRACT

Memories have qualitative properties (e.g., the different kinds of features or details that can be retrieved) and quantitative properties (e.g., the frequency and/or strength of retrieval). Here we investigated the relative contribution of these two properties to the retrieval monitoring process. Participants studied a list of words, and memory for these words was enhanced either by studying an associated picture or by word repetition. Subsequent memory tests required participants to selectively monitor retrieval for these different kinds of stimuli. Compared to words that were studied only once, test words associated with either pictures or repetitions were more likely to be correctly recognized, but critically, false recognition was reduced only when monitoring memory for picture recollections. Subjective judgments and speeded tests indicated that study repetition increased the number of test words that elicited recollection and familiarity (a quantitative difference), but studying pictures maximized the recollection of unique or distinctive details (a qualitative difference). These results indicate that memory quality is more critical than quantity for retrieval monitoring accuracy.



## Introduction

In a seminal review, Koriat, Goldsmith, and Pansky (2000) drew a distinction between quantity-oriented and quality-oriented approaches to understanding memory. Quantity-oriented approaches are concerned with factors that affect the amount of remembered information, whereas quality-oriented approaches are concerned with factors that affect the accuracy of remembered information. They further argued that research on memory accuracy and distortion in the late 20th century represented a historical shift from quantity to quality-oriented approaches, and that metacognitive monitoring processes should play a central role in quality-oriented approaches. Metacognitive monitoring processes are involved in consciously controlled aspects of memory reconstruction, such as selecting the appropriate retrieval strategies and decision criteria.

Along these lines, a goal of recent research has been to understand the role that monitoring processes play in avoiding false memory effects at retrieval (Gallo, 2010; Mitchell & Johnson, 2009). Much of this research has focused on false recognition, or the incorrect acceptance of nontarget information on a recognition memory test. A major challenge has been to explain why some encoding manipulations on target information reduce the size of false recognition effects, even though the items that are used to measure false recognition effects (lures) are not themselves encoded in the same way as the target information. As described below, both quantitative and qualitative approaches have been used to explain these sorts of effects, but there has been little attempt to directly compare the potential contributions of both quantitative factors and qualitative factors to the retrieval monitoring process.

* Corresponding author. Address: Department of Psychology, University of Chicago, 5848 S. University Ave., Chicago, IL 60637, USA.
*E-mail address:* dgallo@uchicago.edu (D.A. Gallo).

0749-596X/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2011.04.008

## Quantitative factors

Signal detection models of recognition memory are a classic example of theories that emphasize quantitative factors. These models assume that memories can be mapped along a single continuum at retrieval, which is often described as memory strength or familiarity (Lockhart & Murdock, 1970). In these theories the main role of monitoring or decision processes is to set a response criterion along the retrieval continuum. More recently, multidimensional signal detection theories have been developed to allow for qualitative differences in memory retrieval (Banks, 2000; Rotello, Macmillan, & Reder, 2004; see also Wixted & Mickes, 2010). However, exactly how qualitatively different types of retrieved information are combined in a memory decision has not been extensively investigated within this approach, and most applications of signal detection models tend to be restricted to situations where a single continuum of retrieval strength may provide an adequate description (e.g., recognition memory for word lists, cf. Wixted, 2007). This theoretical framework has primarily been used to advance a quantitative approach to understanding memory.

With respect to false recognition effects, a quantitative approach has been extensively applied to the mirror effect (Glanzer & Adams, 1985; Greene, 2007), or the finding that some encoding manipulations simultaneously increase the correct recognition of targets (hits) and decrease the false recognition of lures (false alarms). In this context much attention has been paid to strength-based manipulations whereby items receive more study time either through repetition or presentation duration (e.g., Cary & Reder, 2003; Hirshman, 1995; Hockley & Niewiadomski, 2007; Kim & Glanzer, 1993). For example, Stretch and Wixted (1998) had some participants study a list of items three times and had others study a list of items once. Study repetition not only increased hits to repeated items, ostensibly by enhancing their memory strength, but also decreased false alarms to nonstudied items in some conditions. The target repetition effect on false alarms was attributed to the use of a more conservative response criterion along the strength continuum at retrieval, corresponding to the quantitative differences in target strength. Such false recognition effects are not always obtained in repetition studies (e.g., Bruno, Higham, & Perfect, 2009; Verde & Rotello, 2007), but the significant effect in some conditions is theoretically important (for recent discussion, see Starns, White, & Ratcliff, 2010). To the extent that these false recognition effects are driven by strength-based criterion shifts, they suggest that quantitative memory differences can affect the retrieval monitoring process.[1]

More qualitative approaches also have been applied to mirror effects, in the form of dual process theories. These theories make a qualitative distinction between recollection, or the retrieval of specific details from the study phase that are associated with a test item, and familiarity, or a decontextualized feeling of oldness towards the test item itself (Mandler, 1980; Yonelinas, 2002). Cary and Reder (2003) analyzed subjective judgments of recollection and familiarity (e.g., Rajaram, 1993; Tulving, 1985) with respect to the strength-based mirror effect. They found that repetition increased recollection judgments for studied items, but decreased familiarity judgments associated with nonstudied items. These findings suggest that participants were less reliant on familiarity when recollection was high, thereby decreasing false recognition (also see Joordens & Hockley, 2000).

Dual process theories make a qualitative distinction between recollection and familiarity, but many applications of dual process theories to false recognition effects focus on quantitative factors. From this perspective, the false recognition portion of the mirror effect may be attributed either to a qualitative shift away from familiarity-based responding and towards recollection-based responding, or to a quantitative shift in one's response criterion along the familiarity dimension. Sometimes it is attributed to both (see Cary & Reder, 2003; Gallo, Weiss, & Schacter, 2004). More generally, dual process theories emphasize the role of recollection in controlling or overriding the influences of familiarity (e.g., Brainerd, Reyna, Wright, & Mojardin, 2003; Hintzman & Curran, 1994; Jacoby, 1991), thereby associating recollection with more accurate retrieval monitoring. These theories tend to focus on the relative level or amount of recollection compared to familiarity (e.g., how many test items elicit the experience of recollection), again emphasizing quantitative differences. Most dual-processes are mute with respect to how the recollection of qualitatively different kinds of details could affect retrieval monitoring accuracy, such as those details originating from different kinds of stimuli or different sources of studied information.

## Qualitative factors

Relative to these other frameworks, the source-monitoring framework more heavily emphasizes qualitative memory differences across different types or sources of information (Johnson, Hashtroudi, & Lindsay, 1993). According to this framework, memory decisions often rely on metacognitive expectations. When making these decisions, people rely not only on the quantitative properties of memories (e.g., the frequency and/or strength of retrieval), but also on the qualitative properties of memories that they expect to vary across different sources of information (e.g., the recollection of different kinds of details or features). The extent that different quantitative and qualitative properties will contribute to a memory decision is assumed to depend, in part, on the perceived usefulness of those properties to accurate responding (Bink, Marsh, & Hicks, 1999). Within this framework, retrieval monitoring involves the selecting of memory sources (or dimensions) that may be appropriate for a decision, and the setting of corresponding decision criteria along each of these dimensions.

With respect to false recognition effects, this more qualitative approach has been extensively applied in the

[1] Study repetitions also can reduce false recognition in list-based exclusion tasks (Jacoby, 1999), but these effects may be due to the use of a recall-to-reject process that is qualitatively different than the criterion-setting processes of interest in the present paper (i.e., disqualifying vs. diagnostic monitoring, see Gallo, 2010).

context of the distinctiveness heuristic (Schacter & Wiseman, 2006). This theory stems from work showing that false recognition is reduced when memory is tested for pictures relative to words (e.g., Dodson & Schacter, 2002; Israel & Schacter, 1997; Schacter, Israel, & Racine, 1999). According to the distinctiveness heuristic theory, participants expect more distinctive recollections for pictures than words because pictures contain qualitatively richer attributes (e.g., more detailed and unique configurations of visual features across items, cf. Nelson, 1979). These expectations influence decision processes at test, so that participants are more likely to rely on distinctive recollections when tested for pictures and more vague recollections and/or familiarity when tested for words. Because nonstudied items are more likely to elicit vague memories than distinctive ones, they are less likely to be falsely recognized when tested for pictures. Importantly, it is assumed that this decision process is driven by qualitative differences in the to-be-remembered features, as opposed to quantitative differences along a single strength or familiarity continuum.

Although the distinctiveness heuristic theory appeals to qualitative factors, Gallo et al. (2004) raised the alternative possibility that testing memory for pictures may have suppressed false recognition in these studies via quantitative factors (i.e., a familiarity-based criterion shift as described in the mirror effect literature). These two possibilities were difficult to disentangle in prior studies, because they used standard old/new recognition memory tests that did not control or measure the relative contributions of quantitative or qualitative factors on performance. To more directly test the distinctiveness heuristic theory, Gallo and colleagues developed the criterial recollection task (Gallo et al., 2004; Gallo, McDonough, & Scimeca, 2010). The purpose of the criterial recollection task was to exert more experimental control over the types of information that participants use to make memory decisions, relative to standard old/new recognition tests, thereby advancing theories of retrieval monitoring.

In the original task (Gallo et al., 2004), participants studied black words (e.g., hydrant) that were immediately followed by the same word in red font or by a corresponding picture of the object, manipulated within-participants. Memory was subsequently tested using black words as retrieval cues, with different retrieval instructions across test blocks. On the picture test participants needed to accept test words that were associated with a picture at study, whereas on the red word test they needed to accept test words that were associated with a red font at study. Because some items were studied as both red words and pictures, the recollection of the noncriterial format (e.g., red words on the picture test) was relatively uninformative with respect to whether the item had been studied in the criterial format (e.g., pictures on the picture test). Thus, participants could not use a recall-to-reject strategy, and instead had to selectively search their memory for the to-be-recollected information on each test. By holding the retrieval cues constant (black words) and varying the retrieval orientation across the test blocks, differences in memory performance could be attributed to differences in the to-be-recollected stimuli and corresponding retrieval monitoring processes.

Gallo et al. (2004) found that the correct recognition of targets was increased and the false recognition of lures was reduced when participants were tested for pictures compared to red words, even though the same kinds of items needed to be discriminated across the two tests. This pattern is conceptually similar to the mirror effect in standard old/new recognition memory, but critically, the false recognition suppression effect persisted even after red words were repeated at study, thereby increasing their quantitative memory strength beyond that for pictures (for extensions to other kinds of stimuli see Gallo, Meadow, Johnson, & Foster, 2008; McDonough & Gallo, 2008). These studies demonstrate that qualitative factors (e.g., expecting distinctive picture recollections) can affect false recognition independent from quantitative factors, and that qualitative factors have a powerful effect on the retrieval monitoring process.

Although these criterial recollection studies indicate that recollection quality can affect retrieval monitoring accuracy, they leave open the possibility that memory quantity or strength also might affect retrieval monitoring accuracy. This possibility remains because these studies only compared false recognition across conditions that were assumed to vary in recollective distinctiveness (i.e., a qualitative difference). In order to independently assess the effects of memory quantity or strength on retrieval monitoring, memory quantity would need to be manipulated within a single experiment, while holding recollection quality relatively constant. The distinctiveness heuristic theory makes no predictions under these conditions, because that approach focuses on qualitative factors. However, quantitative approaches would predict that testing stronger memories would reduce false recognition compared to a condition where memory was tested for relatively weaker memories. The aforementioned mirror effect studies are consistent with this idea, but as discussed, the use of standard old/new recognition memory tests makes it difficult to disentangle qualitative and quantitative factors in those studies (we return to this issue in the 'General discussion'). To our knowledge, no previous study has directly investigated the possibility that enhancing memory quantity can have similar effects on false recognition as enhancing memory quality, using a task that explicitly requires the monitoring of these different properties of memory at retrieval.

## Current experiments

The goal of the present studies was to independently evaluate the impact of quantitative and qualitative aspects of memory on retrieval monitoring accuracy, and also to directly compare the two. Our working assumptions were that familiarity can only vary quantitatively, but that recollection can vary either quantitatively or qualitatively. For example, some study conditions may enhance the number of studied words that elicit a strong feeling of familiarity at test, even though the type of information retrieved from memory (i.e., a feeling of oldness) is qualitatively the same across conditions. Similarly, some study conditions may enhance the number of studied words that elicit

the recollection of specific details at test, independent of the kinds of details that are recollected (e.g., an associated font color versus a picture). Here the term "strength" refers to such quantitative differences in memory, and is operationalized as the number of test words in a condition that elicit either successful recollection or a strong sense of familiarity. By contrast, we use the term "distinctiveness" to refer to qualitative differences in recollection across conditions, or the extent that one recollection contains features that make it unique relative to others (e.g., the configuration of visual details in a picture helps to differentiate it from other items in memory). Even if the number of words that elicit recollection is held constant across two different conditions (e.g., font color is recollected with the same frequency as pictures), so that recollection is equally strong in this quantitative sense, the quality of the recollected details from one condition might be more distinctive (e.g., pictures) compared to the other (e.g., font color).[2]

In the current studies we used a modified version of the criterial recollection task, explicitly requiring participants to monitor memory for stimuli that primarily differed either qualitatively or quantitatively. Unlike prior uses of this task, we did not attempt to equate targets and lures on memory strength, so that both qualitative and quantitative memory differences could potentially contribute to the decision process. To achieve this end we tested different kinds of targets across test blocks, but we also tested lures that were drawn from the same nontarget source. This procedure allowed target memory to vary across test blocks, but held the familiarity of the lures constant. As a result, any differences in false recognition to these lures across tests blocks could be attributed to differences in retrieval expectations elicited by the to-be-remembered (target) information and corresponding monitoring accuracy. Either recollection or familiarity could potentially contribute to discrimination between targets and lures in this task, but because some of the lures were studied (and hence familiar), we assumed that the recollection of certain features would be the most accurate basis for responding. We also manipulated the speed with which participants made memory judgments to test the extent that participants had attempted to use recollection relative to familiarity.

Experiment 1 aimed to manipulate retrieval quality, extending previous findings with pictures and words to this modified procedure. As in prior work, we assumed that studying pictures with their corresponding words would primarily increase the quality of the recollected information elicited by these words at test (i.e., the number of unique details or features in the recollection), although it also may increase memory quantity (i.e., the number of items eliciting such a recollection and/or the number of items eliciting a strong feeling of familiarity). Experiment 2 aimed to manipulate retrieval quantity while minimizing differences in quality, investigating whether study repetition of words would be sufficient to reduce false recognition under comparable conditions to Experiment 1. We assumed that repeating studied words would primarily increase the number of items eliciting recollection and/or familiarity at test (a quantitative difference), but relative to pictures, repeating words would have relatively less effect on recollection quality. Experiment 3 compared these two manipulations under the same testing conditions, to more directly evaluate these assumptions about the relative effects of memory quality and quantity. We also conducted secondary experiments that used speeded old/new recognition memory tests and self-paced subjective judgments as additional manipulation checks of our assumptions. Of particular interest were subjective ratings of "unique recollected details," described in the methods of these secondary experiments, which were designed to assess the qualitative distinctiveness of retrieved memories.

## Experiment 1A

Participants studied a list of red words and a list of pictures, which served as targets on the subsequent memory tests. They also studied a list of words in black font, in order to provide a common source of familiar lures on the subsequent memory tests. Following the study phase, memory for red words and pictures was tested in separate test blocks, presenting all test words in black font as retrieval cues. The test instructions manipulated retrieval orientation across test blocks, so that participants had to focus on remembering red words (on the red word test) or pictures (on the picture test) when making their memory decisions.

This experiment had two goals. The first goal was to replicate prior work using this modified version of the task, demonstrating that false recognition would be reduced when participants monitored memory for pictures (the picture test block) compared to red words (the red word test block). The second goal was to provide additional evidence that these false recognition differences across test blocks are due to recollection-based responding. To test this hypothesis in the current experiment, participants completed self-paced and speeded tests. Previous research (Dodson & Hege, 2005; Hay & Jacoby, 1996; Yonelinas, 2002) has found that speeded test decisions are more likely to affect recollection than familiarity. If the predicted false recognition differences across test blocks under self-paced conditions are due to recollection-based retrieval monitoring, then these false recognition differences should be minimized on the speeded tests.

### Method

#### Participants

Sixteen University of Chicago undergraduates participated for course credit or $5. All participants spoke English

[2] If the recollection of a given detail is conceptualized as all-or-none (a discrete process), then it is straightforward to compare the number of items that elicit the recollection of one kind of detail to the number of items that elicit the recollection of a qualitatively different kind of detail. If it is instead assumed that the recollection of a given detail can vary in strength or vividness (a continuous process), then the number of items that elicit the recollection of one kind of detail can only be compared to the number of items that elicit the recollection of another kind of detail if the same criterion is used to categorize a "successful" recollection across the two different kinds of details (e.g., claiming to recollect each detail with the same level of confidence).

fluently and reported normal or corrected-to-normal color vision.

*Materials and design*

Stimuli were 256 easily recognizable pictures (e.g., acorn, suitcase, typewriter) and their corresponding verbal labels (words) from Gallo et al. (2004). Pictures were colored images of objects from various Internet sources, with the object cropped from any surrounding context. Pictures and words were presented on a white background on the computer screen. These items were divided into 16 sets of 16 items, which were counterbalanced across the study conditions (black words, red words, pictures, nonstudied) and the test conditions (described below).

Study items were presented in blocked format. The black word block was presented first to equate processing of the black words across all participants. The black word block consisted of 128 words presented in black font. Half of these black words (64) were only studied in the black word block, in order to serve as familiar lures for the subsequent memory tests (i.e., items that were studied, but not in either of the criterial formats). For the other half of the black words, 32 were subsequently studied in red font (i.e., in the red font block), and 32 were subsequently studied as a picture (i.e. in the picture block). These "both items" were included so that membership in the black word block was not mutually exclusive with membership in the other study blocks, so that participants could not reject black words at test by recollecting that the item had been presented in the black word block. Following the black word block, the ordering of the red word block and picture block was counterbalanced across participants. The red word block consisted of 64 items presented in red font, half of which had earlier been presented in the black word block. The picture block consisted of 64 items presented as green words and pictures, half of which had earlier been presented in the black word block. The green words served as labels for the pictures, so that these same words would serve as retrieval cues for the pictures on the subsequent tests.[3] Within each study block, the order of items was randomized.

Participants completed four different test blocks, with test items always presented as words in black font. Half of the participants completed the two speeded tests before the two self-paced tests, and half of the participants completed the two self-paced tests before the two speeded tests. Within each test speed, the order of the test type (picture test or red word test) was counterbalanced across participants. During each test, participants were tested on 64 items (randomly arranged). There were 16 items of each type in each test. The red word test contained items that were studied as red words, black words, both, or nonstudied, whereas the picture test contained items studied as pictures, black words, both, or nonstudied. Items were only tested once.

*Procedure*

Participants were instructed that they would study several lists of items, and to pay attention to the presentation format of each item for subsequent memory tests. They were informed that some words would be presented in black font, some would be presented in red font, and some would be presented with corresponding pictures. They were further instructed that some items would be presented in multiple formats. Each black word and red word was presented for 2 s. For each picture item, its corresponding word label was presented in green font for 500 ms, followed by the actual picture for 1.5 s. This label was necessary to cue the item at test.

Following the study phase, the experimenter read test instructions aloud to the participant. On the red word tests, participants were told that they would be tested on items that had either been studied as red words only, as black words only, as both red and black words, or were not studied. Participants were informed that items studied as pictures would not appear on this test. On this test, participants were to press "Yes" if they remembered studying the item as a red word (i.e., it might have been studied as a red word only, or as both a red word and a black word), otherwise "No." Participants were told that remembering whether or not they had studied an item as a black word was irrelevant on the test, because some items were studied as both black words and as red words, and others were not. The instructions for the picture test were similar, except they were adjusted to correspond to pictures rather than red words.

On the self-paced test, participants were instructed that there was no time limit on how quickly they must respond to test items and asked to be as accurate as possible. On the speeded test, participants were instructed that they would have less than one second to respond after the test item appeared on the screen. Participants were asked to respond within the time limit and to be as accurate as possible. A tempo procedure was used for the speeded test (cf. Balota, Burgess, Cortese, & Adams, 2002). For each trial, a series of visual arrows appeared on the screen at 700 ms intervals. The test item appeared 700 ms after the second set of visual arrows was displayed. If participants responded in less than 800 ms, a "good" feedback appeared on the screen and they were allowed to proceed to the next trial. If participants took longer than 800 ms to respond, a "too slow" feedback appeared on the screen for three seconds before they were allowed to proceed to the next trial. After both "good" and "too slow" feedback, participants proceeded to the next trial by pressing the spacebar.

*Results and discussion*

Data from Experiment 1A are presented in Table 1. We first analyzed results from the self-paced test, highlighting two key points. First, as instructed, participants attempted to endorse items that had been associated with the criterial information at study. On the picture test, hits to picture items were greater than false alarms to black words (.63

[3] Because pictures were always associated with the same word in green font, the recollection of either a picture or a font color could have supported performance on the picture test. However, by design, the recollection of font color also could have supported performance on the red word test, so that any performance differences between the tests can be attributed to picture recollections. Experiment 3 provides a replication of this study while varying font color in a different way.

**Table 1**
Proportion of "yes" responses on criterial recollection tests in Experiment 1A.

| | Self-paced | Speeded |
|---|---|---|
| *Red word (1x) test* | | |
| Both targets (red) | .66 (.04) | .59 (.05) |
| Red targets (1x) | .51 (.05) | .47 (.05) |
| Black lures (1x) | .46 (.04) | .48 (.05) |
| New lures (0x) | .11 (.02) | .24 (.03) |
| *Picture (1x) test* | | |
| Both targets (picture) | .72 (.03) | .56 (.05) |
| Picture targets (1x) | .63 (.05) | .52 (.05) |
| Black lures (1x) | .18 (.03) | .35 (.05) |
| New lures (0x) | .07 (.02) | .21 (.04) |

*Notes.* The standard error of each mean is in parenthesis. 1x = studied once, 0x = not studied. "Both targets" were studied twice, once as a black word and once as a red word ("red"), or once as a black word and once as a picture ("picture").

and .18, $t(15) = 6.99$, $SEM = .063$, $p < .001$, $d = 2.70$). This effect was not significant on the red word test (.51 and .46, $p = .49$), indicating that it was very difficult to recollect font color. Second, studying items as black words elevated responses on each of these tests, demonstrating familiarity-based confusions. False alarms to black words were greater than those to new items (i.e., nonstudied words) on the picture test (.18 and .07, $t(15) = 4.53$, $SEM = .025$, $p < .001$, $d = .93$) and the red word test (.46 and .11, $t(15) = 9.12$, $SEM = .039$, $p < .001$, $d = 2.57$). Also, hits to both items (i.e., targets that had also been studied as black words) were greater than hits to pictures (.72 and .63, $t(15) = 2.67$, $SEM = .037$, $p < .05$, $d = .57$), and similarly for the red word test (.66 and .51, $t(15) = 4.28$, $SEM = .036$, $p < .001$, $d = .81$).

We next compared self-paced performance across the two test blocks. Hits to picture items on the picture test (.63) were greater than hits to red words on the red word test (.51), $t(15) = 2.80$, $SEM = .041$, $p < .05$, $d = .60$, demonstrating the picture superiority effect (e.g., Jenkins, Neale, & Deno, 1967; Nelson, Reed, & Walling, 1976). In contrast, false alarms to black words on the picture test (.18) were lower than false alarms to black words on the red word test (.46), $t(15) = 5.85$, $SEM = .047$, $p < .001$, $d = 1.85$, demonstrating the false recognition suppression effect often associated with pictures (e.g., Dodson & Schacter, 2002; Gallo et al., 2004; Schacter et al., 1999). Taken together, these findings represent a mirror effect. Because the lures were equated on familiarity across the test blocks, these results suggest that the false recognition portion of this mirror effect was driven by a retrieval monitoring process.

Consider next the results from the speeded tests, from which we included all responses in the analysis (regardless of response latency) to avoid item-selection effects. These tests generally decreased hit rates and increased false alarm rates compared to the self-paced tests, consistent with the idea that they reduced recollection-based responding. To simplify the analysis of these results we computed source discrimination scores, or the difference between criterial hits (correct "yes" responses to either picture items or red word items on the respective test) and black word false alarms. A 2 (test) × 2 (speed) ANOVA revealed an effect of test, $F(1, 15) = 31.15$, $MSE = .04$, $p < .001$, and speed, $F(1, 15) = 15.89$, $MSE = .03$, $p < .01$, and a marginal interaction, $F(1, 15) = 4.35$, $MSE = .04$, $p < .06$. Speed tended to reduce accuracy more on the picture test, where recollection was assumed to be greatest, although this effect is difficult to interpret because source discrimination was very low on the red word test. Additional evidence for differences in recollection and familiarity are provided in Experiment 1B.

Response latency analysis confirmed that responses were faster on the speeded test (mean of all responses = 582 ms) than on the self-paced tests (1466 ms), $F(1, 30) = 99.53$, $MSE = 125{,}851$, $p < .001$, with an average of 94% of the responses occurring before the 800 ms "too slow" prompt. Participants responded quite rapidly on the speeded test, most likely because they were only penalized for responding too slow. We also found a test x speed interaction, $F(1, 30) = 12.53$, $MSE = 22{,}216$, $p = .001$. Self-paced responses were slower for the red word test (1543 ms) than for the picture test (1391 ms), $t(15) = 2.12$, $SEM = 71$, $p = .05$, suggesting that the monitoring process was more effortful on the red word test (cf. Gallo et al., 2004, 2010). In contrast, speeded responses were slower for the picture test (638 ms) than the red word test (526 ms), $t(15) = 5.40$, $SEM = 21$, $p < .001$. This result potentially indicates that participants were more likely to try to use recollection on the speeded picture test than on the speeded red word test.

## Experiment 1B

In this experiment, participants completed the same study phase as in Experiment 1A, but instead of completing criterial recollection tasks, participants completed memory tests designed to investigate the subjective differences between memory for pictures and words. As one measure of familiarity, participants first completed a speeded old/new recognition test that should have reduced the impact of recollection. They then completed a self-paced test consisting of a series of judgments: an "old"/"new" recognition judgment, an "actually recollect"/"very familiar" judgment for items judged "old", and a numerical rating of the relative amount of "unique recollected details" for items that were judged "actually recollect." The "actually recollect"/"very familiar" judgment is analogous to the "remember"/"know" judgment that has frequently been used in recognition memory research (Rajaram; 1993; Tulving, 1985), but because these judgments reflect the overall quantity of recollection and familiarity, we included the rating of "unique recollected details" to more directly assess qualitative differences in recollective distinctiveness. We assume that "actually recollect" judgments reflect the number of items that elicit a recollection, which is potentially independent from the quality of these recollections. In contrast, we assume that ratings of "unique recollected details" gauge the relative level of distinctiveness of the recollected items, on average, which is potentially independent from the number of items eliciting a recollection.

### Method

#### Participants

Seventeen undergraduates participated, using the same recruitment procedures as the prior experiment. Data from one participant were discarded due to failure to complete the task.

#### Materials and design

Stimuli were the same 256 items used in Experiment 1A. These items were divided into 16 sets of 16 items, which were counterbalanced such that each participant in Experiment 1B studied a list of items that was identical to the items studied by a participant in Experiment 1A. By keeping the study phase identical to Experiment 1A, these methods ensured that the relative memory differences across the stimuli would be preserved. To measure these memory differences, participants completed two memory tests. The first was a speeded old/new recognition test. The second test consisted of a series of subjective memorial judgments, described in detail below. During each test, participants were tested on 112 items (randomly arranged). Each test contained 16 items from each of the classes of studied items (items studied as a black word only, as a red word only, as a picture only, as both a black word and red word, and as both a black word and picture), as well as 32 nonstudied items. Because we had fewer test blocks in this experiment than in Experiment 1A, we did not test the remaining 32 items that had been studied as black words only. Test items always were presented as words in black font.

#### Procedure

The study phase and instructions of Experiment 1B were identical to those of Experiment 1A. Participants completed two memory tests. Participants were told that items could have been studied in one format, studied in multiple formats, or never studied. The first test was a speeded old/new recognition memory test, which used a similar 700 ms tempo procedure as in Experiment 1A. However, in addition to the "too slow" prompt for responses over 800 ms, we also added a "too fast" prompt for responses faster than 600 ms in order to help participants reach the 700 ms tempo. The second test was self-paced and consisted of a series of judgments. The first judgment was an old/new recognition judgment. If participants responded "new", the trial ended; if they responded "old", they made a second decision. For the second judgment (hereafter referred to as the recollect-familiar judgment), participants made one of two judgments: actually recollect (AR) or very familiar (VF). If participants responded "very familiar," the trial ended. If they responded "actually recollect," they made a third judgment. On the third judgment, hereafter referred to as the unique recollected details (URD) rating, participants used a 1–7 scale to rate the amount of unique details they could recollect for each item.

The specific instructions for these subjective judgments were based on prior research (see Gallo et al., 2008; McDonough & Gallo, 2008). For the recollect-familiar judgment, the participants were instructed to respond "actually recollect" if they could recall, or bring to mind, a specific memory of the item's presentation during the study phase (such as the font color or what they were thinking about), and to respond "very familiar" if they thought the test item was studied but were unable to recollect any specific details about the items presentation during the study phase. Participants were instructed to avoid using confidence to decide between these two judgments. For the rating of unique recollected detail, participants were instructed to rate the amount of unique details that they could retrieve for each recollected item. Unique details were defined as those that were "different from other items in the study phase, independent of how strong or vivid those recollections may be." They were told to use the entire 1–7 scale, with low ratings indicating that they recollected very few unique details (i.e., the recollection of general features that were shared with memories of other items), and high ratings indicating that they recollected many unique details for the item (i.e., a distinctive recollection that contained several thoughts or features that were uniquely associated with the item). We assume that participants used this scale to make relative ratings of recollective distinctiveness across items, as opposed to a literal count of the absolute number of unique features, although either interpretation would suffice and we left it to them to interpret the scale at this level. Participants were asked to repeat the instructions to the experimenter to ensure that they understood them.

#### Results and discussion

The results from Experiment 1B are presented in Table 2. The most critical comparison for our hypothesis was between test words that had been studied in red font and test words that had been studied as pictures. As predicted, participants were more likely to respond "actually recollect" for picture items (.48) than for red word items (.34), $t(15) = 2.51$, $SEM = .058$, $p < .05$, $d = .52$. This result is consistent with the hypothesis that pictures enhance recollection. However, it may be that these judgments were driven more by quantitative differences in recollection (i.e., the number of items eliciting recollection) rather than by qualitative differences in recollection (i.e., the average distinctiveness of recollected items). The unique recollected detail ratings were intended to more specifically assess recollection quality. On these ratings picture items were rated more highly (4.46) than red word items (3.33), $t(13) = 3.61$, $SEM = .319$, $p < .01$, $d = .70$.[4] A similar pattern was observed for items that had been studied as both pictures and black words (4.52) compared to those that had been studied as both red words and black words (3.89), although this effect failed to reach significance ($p = .07$). Overall these ratings were consistent with the idea that pictures elicited more distinctive recollections than red words.

The recollect-familiar judgments also provided an estimate of the relative levels of familiarity of the items. A direct comparison of the familiarity judgments between red

[4] Two participants had no ratings of uniquely recollected details for red words.

**Table 2**
Proportion of speeded old/new recognition responses and self-paced subjective responses in Experiment 1B.

| | Speeded test | Self-paced test | | |
|---|---|---|---|---|
| | "Old" | "AR"/"VF" | IRK | "URD" |
| Both targets (picture) | .66 (.05) | .56 (.07)/.16 (.04) | .33 (.07) | 4.52 (.38) |
| Both targets (red) | .64 (.05) | .46 (.06)/.18 (.04) | .32 (.06) | 3.89 (.33) |
| Picture targets (1x) | .64 (.04) | .48 (.07)/.14 (.03) | .26 (.05) | 4.46 (.38) |
| Red targets (1x) | .51 (.06) | .34 (.07)/.19 (.04) | .29 (.04) | 3.33 (.45) |
| Black targets (1x) | .48 (.05) | .31 (.06)/.15 (.03) | .22 (.04) | 3.82 (.42) |
| New lures (0x) | .30 (.04) | .03 (.02)/.13 (.02) | .14 (.03) | – |

*Notes*. The standard error of each mean is in parenthesis. 1x = studied once, 0x = not studied. "Both targets" were studied twice, once as a black word and once as a picture ("picture"), or once as a black word and once as a red word ("red"). AR = actually recollect, VF = very familiar, IRK = familiarity estimate from the independent-remember-know adjustment, URD = mean ratings of unique recollected details (1–7 scale). False recognition of new lures was too low to elicit meaningful recollection ratings.

word items (.19) and picture items (.14) revealed no difference, $p = .10$. However, according to dual-process models, raw proportions of familiarity judgments can misrepresent familiarity, because familiarity judgments can be made only in the absence of recollection. We therefore used the independent remember-know procedure (IRK) to correct for this discrepancy (i.e., $F = F/(1 - R)$, calculated for each participant and then averaged, see Yonelinas, 2002). Using the IRK procedure, there still was no difference between red word items and picture items ($p = .59$), nor was there a difference between the corresponding both items ($p = .60$). In contrast, the speeded old/new recognition test revealed that picture hits (.64) were greater than red words hits (.51), $t(15) = 2.59$, $SEM = .051$, $p < .05$, $d = .67$. To the extent that speeded recognition relies mostly on familiarity, these results suggest that studying pictures made the test cues more familiar than the other items, although some recollection may have persisted on the speeded test. (As in Experiment 1A, we included all responses in our analysis of the speeded test.) A similar difference on the speeded test was not observed between the corresponding both items ($p = .94$).[5] Participants were good at keeping the 700 ms response tempo, as the mean response latency across all items was 692 ms (standard error = 13 ms), with 86% of the responses occurring before the 800 ms "too slow" prompt.

In sum, subjective judgments of unique recollected details were greater for pictures than for red words, consistent with the assumption that this manipulation affected recollection quality. These findings bolster the idea that the false recognition differences across test blocks in Experiment 1A (self-paced) were due to differences in recollective distinctiveness. Nevertheless, studying pictures also might have enhanced quantitative aspects of memory, relative to words studied only once. Although subjective estimates of familiarity did not differ between these items, differences were observed on the speeded old/new test. Experiment 2 was designed to more directly assess the potential effects of memory quantity on performance, using a manipulation that we assumed would have relatively little effect on recollection quality.

[5] A reviewer noted that, relative to recognition on the self-paced test (AR + VF), recognition on the speeded test appeared to be reduced to the same extent for red words and pictures. This finding may appear inconsistent with the idea that pictures elicited more recollection, but because the speeded test always preceded the self-paced test in this experiment the two were not designed to be directly comparable.

## Experiment 2A

Experiment 2A investigated the role of quantitative memory differences in retrieval monitoring, while minimizing qualitative differences across items. Participants in this experiment again studied black words and red words, as in Experiment 1A. However, instead of studying pictures, the green words that had been associated with pictures in Experiment 1A were replaced with green words that were repeated multiple times at study. We assumed that repeating study words would lead to stronger memories than presenting words only once, representing a quantitative memory difference (i.e., green words would be more familiar and/or recollected more frequently). However, we assumed that there should be relatively little qualitative differences between the two (i.e., similar types of to-be-recollected features), at least as compared to the large qualitative differences in perceptual detail between red words and pictures. Participants again completed self-paced and speeded criterial recollection tasks.

We predicted that participants would show increased recognition of studied items on the green word test compared to the red word test, owing to the study repetitions, but predictions for false recognition differed between the quantitative and qualitative approaches. If quantitative differences can drive retrieval monitoring, then false recognition should be reduced on the green word test compared to the red word test, analogous to the picture/red word effect observed in Experiment 1A and to the false alarm portion of the strength-based mirror effect described in the Introduction. In fact, because repetition is likely to enhance the quantitative amount of both recollection and familiarity, performance on the green word test could benefit from either recollection or familiarity-based responding, thereby providing a strong test of the idea that quantitative memory differences can affect retrieval monitoring. In contrast, if qualitative differences are the more critical factor for retrieval monitoring, then the false recognition effect observed in Experiment 1A might not be found in Experiment 2A, because study repetition is not assumed to affect memory quality as much as quantity.

*Participants*

Eighteen undergraduates participated using the same recruitment procedures as the prior experiments. The data from one participant were discarded due to failure to follow task instructions, and the data from another participant were lost due to computer malfunction.

*Materials and design*

Stimuli were the same 256 words used in Experiment 1A. The study phase was identical to that of Experiment 1A, except that instead of studying a picture block, participants studied a green word block. The green word block consisted of 64 words presented in green font, with half of these words presented earlier in black font. Green words were repeated three times (non-consecutively) at study.

*Procedure*

Instructions were similar to those used in Experiment 1A, but they were adapted for the green word/red word manipulation. Participants were instructed that they would study several lists of words for a subsequent memory test. They were instructed to pay attention to the color of the font of each word because their memory would be tested for font color. Participants were informed that words would be presented in black font, green font, or in red font. They also were told that some words would be presented in both black font and green font, and that other words would be presented in both black font and red font. Every word in the green word block was presented three times, randomly distributed throughout the block. Each study trial was presented for two seconds. Participants took the same tests as in Experiment 1A, with the same tempo procedures, except they completed green word tests instead of picture tests. Test items always were presented as words in black font.

*Results and discussion*

Data for Experiment 2A are presented in Table 3. Within each of the self-paced tests performance was similar to Experiment 1A. On the green word test, hits to green word items were greater than false alarms to black word items (.71 and .42, $t(15) = 5.79$, $SEM = 0.051$, $p < .001$, $d = 1.55$), and on the red word test, hits to red word items were greater than false alarms to black word items (.50 and .41, $t(15) = 1.95$, $SEM = .044$, $p = .07$, $d = .42$). False alarms to black words were greater than those to nonstudied items on the green word test (.42 and .17, $t(15) = 7.30$, $SEM = .034$, $p < .001$, $d = 1.24$) and the red word test (.41 and .19, $t(15) = 6.06$, $SEM = .036$, $p < .001$, $d = 1.08$). On the green word test, hits to both items were marginally greater than those to green words (.78 and .71, $t(15) = 1.78$, $SEM = .035$, $p = .096$, $d = .38$), and the analogous effect was significant for the red word test (.64 and .50, $t(15) = 6.26$, $SEM = .022$, $p < .001$, $d = .69$).

We next compared self-paced performance across the two test blocks. Hits to green words on the green word test (.71) were greater than hits to red words on the red word test (.50), $t(1,15) = 4.52$, $SEM = .048$, $p < .001$, $d = 1.16$, demonstrating the typical repetition effect on memory for studied items. Critically, false alarms to black words on the green word test (.42) were not different from false alarms to black words on the red word test (.41), $p = .75$. The lack of a false recognition difference is a major departure from the results obtained in Experiment 1A, and also does not mimic the false recognition effects we would expect from a strength-based mirror effect. This finding indicates that quantitative differences in memory did not affect retrieval monitoring processes in the same way as qualitative differences.

**Table 3**
Proportion of "yes" responses on criterial recollection tests in Experiment 2A.

| | Self-paced | Speeded |
|---|---|---|
| *Red word (1x) test* | | |
| Both targets (red) | .64 (.05) | .62 (.06) |
| Red targets (1x) | .50 (.05) | .48 (.07) |
| Black lures (1x) | .41 (.05) | .39 (.04) |
| New lures (0x) | .19 (.05) | .23 (.05) |
| *Green word (3x) test* | | |
| Both targets (green) | .78 (.04) | .69 (.04) |
| Green targets (3x) | .71 (.04) | .62 (.04) |
| Black lures (1x) | .42 (.05) | .46 (.05) |
| New lures (0x) | .17 (.04) | .28 (.06) |

*Notes.* The standard error of each mean is in parenthesis. 1x = studied once, 3x = studied thrice, 0x = not studied. Some "both targets" were studied twice, once as a black word and once as a red word ("red"), whereas others were studied four times, once as black word and thrice as a green word ("green").

Consider next the results from the speeded tests. A 2 (test) × 2 (speed) ANOVA on source discrimination scores revealed an effect of test, $F(1, 15) = 7.34$, $MSE = .04$, $p < .05$, no effect of speed, $F(1, 15) = 1.91$, $MSE = .04$, $p = .19$, and a significant interaction, $F(1, 15) = 4.77$, $MSE = .02$, $p < .05$. The interaction indicates that speed only reduced discrimination on the green word test, suggesting that the self-paced green word test was more likely to rely on recollection. By contrast, discrimination on the red word test was unaffected by response speed, suggesting that this test was primarily driven by familiarity. Although the red words and black words were each presented once at study, the black word list always preceded the other study lists. This design feature may have reduced the familiarity of the black words by the time they were tested, especially because the repetition of green words added to the length of the intervening study lists.

Analysis of response latencies confirmed that responses were faster on the speeded test (mean of all responses = 549 ms) than on the self-paced tests (1335 ms), $F(1, 30) = 40.40$, $MSE = 244{,}685$, $p < .001$, with an average of 97% of the responses occurring before the 800 ms "too slow" prompt. There was no effect of test and no interaction (both $F$s < 1), although a post hoc analysis revealed that speeded responses were slower for the green word test (572 ms) than the red word test (526 ms), $t(15) = 3.20$, $SEM = 14$, $p < .01$. Analogous to the pattern observed in Experiment 1A, this result potentially indicates that participants were more likely to try to use recollection

on the speeded green word test. In contrast, self-paced responses did not differ between the red word test (1385 ms) and the green word test (1285 ms), $t < 1$.

## Experiment 2B

Participants completed the same study phase as in Experiment 2A, but were given memory tests designed to investigate the subjective memory differences between green words and red words. These test procedures were parallel to those of Experiment 1B.

### Method

#### Participants

Sixteen young adults participated, using the same recruitment procedures as the prior experiments.

#### Procedure

Stimuli were the same 256 used in Experiment 2A. These items were divided into 16 sets of 16 items, which were counterbalanced such that each participant in Experiment 2B studied a list of items that was identical to a participant in Experiment 2A. The study phase and instructions of Experiment 2B were identical to those of Experiment 2A. Participants completed the same two memory tests as in Experiment 1B, with the same tempo procedure used in that experiment.

### Results and discussion

Data from Experiment 2B are presented in Table 4. The most critical comparison for our hypothesis was between words that had been studied three times in green font and words that had been studied once in red font. Participants were more likely to claim to "actually recollect" green word items (.47) than red word items (.24), $t(15) = 4.31$, $SEM = .053$, $p < .001$, $d = 1.167$, suggesting that repeating green words enhanced recollection. However, this increase may have reflected a quantitative difference (i.e., recollecting green words more frequently than red words, but with similar quality) or a qualitative difference (i.e., more distinctive recollections for green words). Consistent with a quantitative difference, the unique recollected details ratings revealed no significant difference between the green word items and red word items, nor was there a difference between the corresponding both items (both $p$'s > .50). Unlike the significant effect in ratings of unique recollected details that we found with the picture/word manipulation in Experiment 1B, these results suggest that repeating words did not significantly enhance recollection quality.

Differences in familiarity were observed between green words and red words. Familiarity estimates (using the IRK procedure) indicated that green word items (.61) were more familiar than red word items (.47), $t(15) = 2.93$, $SEM = .050$, $p < .05$, $d = .54$, and a similar effect was found on the speeded recognition test, .75 and .65, $t(15) = 1.74$, $SEM = .053$, $p = .10$, $d = .48$. Similar patterns were found in the corresponding both items, although only the difference on the speeded test was significant. As in Experiment 1B, participants were good at keeping the 700 ms response tempo on the speeded test. The mean response latency collapsing across items was 696 ms (standard error = 22 ms), with 86% of the responses occurring before the 800 ms "too slow" prompt.

In sum, repeating the green words three times at study enhanced recollection and familiarity compared to red words, but these effects were primarily limited to quantitative increases as measured by subjective ratings. When recollection quality was measured, using the unique recollected details rating, there were no significant differences between green words and red words. These results stand in contrast to those found when comparing pictures to red words, which showed both quantitative and qualitative memory differences.

## Experiment 3

The main goal of Experiment 3 was to replicate the effects in Experiments 1 and 2 in a single experimental design. A comparison of Experiments 1A and 2A suggests that false recognition was lower on the picture test compared to the green word test (.18 and .42, $t(30) = 3.80$, $p = .001$), even though hits to green words were somewhat greater than hits to pictures (.71 and .63, $t(30) = 1.47$, $p = .15$). If this pattern can be replicated in a single study, it would bolster the conclusion that qualitative differences are more likely than quantitative differences to suppress false recognition. We also sought to replicate our

**Table 4**
Proportion of speeded old/new recognition responses and self-paced subjective responses in Experiment 2B.

| | Speeded test | Self-paced test | | |
|---|---|---|---|---|
| | "Old" | "AR"/"VF" | IRK | "URD" |
| Both targets (green) | .79 (.03) | .54 (.06)/.24 (.04) | .55 (.08) | 3.53 (.29) |
| Both targets (red) | .67 (.05) | .41 (.05)/.27 (.03) | .49 (.06) | 3.32 (.26) |
| Green targets (3x) | .75 (.04) | .47 (.06)/.29 (.04) | .61 (.07) | 3.51 (.27) |
| Red targets (1x) | .65 (.05) | .24 (.04)/.34 (.05) | .47 (.07) | 3.38 (.38) |
| Black targets (1x) | .60 (.04) | .25 (.03)/.26 (.04) | .36 (.06) | 3.34 (.26) |
| New lures (0x) | .36 (.06) | .04 (.01)/.14 (.03) | .15 (.04) | – |

*Notes*. The standard error of each mean is in parenthesis. 1x = studied once, 3x = studied thrice, 0x = not studied. Some "both targets" were studied twice, once as a black word and once as a red word ("red"), whereas others were studied four times, once as black word and thrice as a green word ("green"). AR = actually recollect, VF = very familiar, IRK = familiarity estimate from the independent-remember-know adjustment, URD = mean ratings of unique recollected details (1–7 scale). False recognition of new lures was too low to elicit meaningful recollection ratings.

subjective judgment effects in a single group of participants, given that subjective judgments about recollection and familiarity can be sensitive to test context and instructions (e.g., Bodner & Lindsay, 2003; Geraci & McCabe, 2006). The current design allowed participants to compare their memory for all of the different kinds of stimuli when making the subjective judgments, thereby maximizing consistency in their use of the judgments.

A secondary goal for Experiment 3 was to make the importance of repetitions more explicit at test. Although we had repeated green words during the study phase of Experiment 2A, we did not reiterate in the green word test instructions that the target items were repeated. To investigate whether or not such instruction would matter, in Experiment 3 we reiterated at test that all of the green words had been repeatedly studied for half the participants, whereas we used instructions that did not reiterate repetition in the other half. To anticipate, the same results were obtained regardless of whether repetitions were emphasized or not in the test instructions. We also simplified the experimental design by excluding the speeded tests and reducing the types of studied items in this experiment.

## Method

### Participants

Fifty-eight young adults participated using the same recruitment procedures as the prior experiments. The data from two participants were lost due to computer malfunction.

### Materials and design

Stimuli were 224 of the items used in the prior experiments. These items were divided into 14 sets of 16 items, which were counterbalanced across the study conditions (black words, red words, green words, pictures, and nonstudied) and the test types (criterial recollection and manipulation check). Items were studied in a single mixed block. Sixty-four items were presented in black font only, 32 items in red font, 32 items in green font, and 32 items as pictures. The items presented in green font were repeated three times at study. At study, all red words, green words, and pictures were immediately preceded by the same item in black font (500 ms), so that presentation as a black word was common to all of these items.

Participants completed four test blocks. Three of the blocks consisted of a self-paced criterial recollection test: a red word test, a green word test, and a picture test. Each of these tests contained 48 items: 16 that had been studied in the criterial format, 16 that had been studied only as black words, and 16 nonstudied items. The other test block consisted of a self-paced old/new recognition judgment, followed by the recollect-familiar judgment and the rating of unique recollected details. This test contained 80 items: 16 items that had been studied in each of three criterial formats, 16 studied only as black words, and 16 new items. The order of the three criterial recollection test blocks was counterbalanced across participants, as was whether the old/new recognition memory test with subjective ratings occurred first or last. Test items always were presented as words in black font.

### Procedure

Study instructions were similar to those used in Experiments 1 and 2, but were adapted to inform participants that they would be studying items in all four formats (black words, red words, green words, and pictures). Items studied as black words only were presented for 2 s in black font. For all other items, the corresponding word was presented in black font for 500 ms, followed by the item in the criterial format. Because we presented a brief black word for every item in this experiment, and intermixed all of the studied items into in a single list, we did not include a separate set of items studied in multiple formats (e.g., the "both" items of prior experiments). Following the study phase, the experimenter read test instructions aloud to the participant. The instructions for each test were similar to the instructions for the criterial recollection tests and subjective judgments described previously, except that half the participants were reminded that the green words were repeated at study and the red words were not. All responses were self-paced.

### Results and discussion

Data from the criterial recollection tests of Experiment 3 are presented in Table 5. Preliminary analyses revealed that the new instructional manipulation (color instructions vs. repetition instructions) had no significant effects, and so for simplicity we only report analyses that were collapsed across this manipulation. As can be seen from the table, these two instructional conditions resulted in the same pattern of false recognition differences across the test blocks (i.e., red word test = green word test > picture test), thereby representing two independent replications of our most important results.

Comparing the picture test to the red word test, hits to picture items (.60) were greater than hits to red word items (.42), $t(55) = 5.61$, $SEM = .032$, $p < .001$, $d = .86$, whereas false alarms to black words were lower on the pic-

**Table 5**
Proportion of "yes" responses on self-paced criterial recollection tests in Experiment 3.

| | Color instructions | Repetition instructions |
|---|---|---|
| *Red word (1x) test* | | |
| Red targets (1x) | .38 (.04) | .46 (.04) |
| Black lures (1x) | .26 (.03) | .20 (.03) |
| New lures (0x) | .08 (.02) | .08 (.02) |
| *Green word (3x) test* | | |
| Green targets (3x) | .67 (.04) | .68 (.05) |
| Black lures (1x) | .25 (.03) | .17 (.03) |
| New lures (0x) | .11 (03) | .09 (.02) |
| *Picture (1x) test* | | |
| Picture targets (1x) | .60 (.04) | .60 (.03) |
| Black lures (1x) | .04 (.01) | .02 (.01) |
| New lures (0x) | .03 (.02) | .02 (.01) |

*Notes.* The standard error of each mean is in parenthesis. 1x = studied once, 3x = studied thrice, 0x = not studied. "Both targets" were not needed in this design.

ture test (.03) than on the red word test (.23), $t(55) = 9.69$, $SEM = .021$, $p < .001$, $d = 1.56$. This comparison replicated the mirror-effect pattern observed in Experiment 1A, and extended this pattern to a situation where memory for red words was relatively greater. Comparing the green word test to the red word test, hits to green words (.67) were higher than hits to red words (.42), $t(55) = 8.46$, $SEM = .029$, $p < .001$, $d = 1.09$, whereas false alarms to black words were equivalent on the green word test (.21) and the red word test (.23), $p = .34$. This comparison replicated the pattern observed in Experiment 2A, and again indicates that study repetition does not yield a strength-based mirror effect on this kind of retrieval monitoring task.

The design of Experiment 3 also allowed for a direct comparison of the picture test and the green word test. In this experiment, the pictures were presented once at study and the green words were presented three times, and each kind of item was preceded by a black word (thereby equating this factor across the conditions). Here we observed that false alarms to black words on the picture test (.03) were lower than those on the green word test (.21), $t(55) = 8.69$, $SEM = .021$, $p < .001$, $d = 1.54$. These results again suggest that qualitative memory differences (pictures > repetitions) are critical for obtaining a false recognition suppression effect. With respect to false recognition of the black words, the only difference between these two tests was the retrieval orientation elicited by the test instructions and the nature of the targets. We also found that hits to pictures (.60) were lower than hits to green words (.67), $t(55) = 2.39$, $SEM = .023$, $p = .02$, $d = .32$, potentially reflecting a quantitative memory difference between these two types of targets that was in the opposite direction of the qualitative difference (repetitions > pictures, see next).

Subjective judgments from the standard old/new recognition test are presented in Table 6. These results suggest three main conclusions. The first conclusion is that repetitions of green words enhanced familiarity relative to the other stimuli, representing a quantitative difference. Participants were most likely to respond "old" to green words than to pictures (.75 and .68, $t(55) = 2.84$, $SEM = .028$, $p < .01$, $d = .36$), and each of these items was recognized more often than red words and black words ($p$'s < .001). Moreover, the results of the independent remember-know procedure (IRK) confirmed that the green words were more familiar than pictures (.51 and .36, $t(55) = 3.15$, $SEM = .046$, $p < .01$, $d = .50$), as well as the other items (all $p$'s < .001). The second conclusion is that both repetition and pictures appeared to increase quantitative aspects of recollection. "Actually recollect" judgments revealed no difference between green words and pictures (.51 and .54, $p = .43$), even though these items were judged to be "actually recollected" more frequently than the other items ($p$'s < .01). The third conclusion is that pictures led to the highest ratings of unique recollected details, representing a qualitative difference. Pictures were rated more highly than green words (4.48 and 3.76, $t(55) = 4.56$, $SEM = .16$, $p < .001$, $d = .51$), as well as the other items. Green words also were rated more highly than red and black words ($p$'s < .05), indicating that study repetition also increased recollective detail, albeit not as much as studying pictures.

Overall, these subjective results replicate the patterns found in the previous experiments, and indicate that these differences persist even when the different classes of items are presented to the same participants in the same experimental session. Word repetition enhanced recollection quantity to the same level as pictures, and enhanced familiarity even more, leading to memories that were quantitatively stronger. By contrast, studying pictures led to the highest levels of qualitative distinctiveness, as measured by ratings of unique recollected details.

## General discussion

The current experiments provided three key results. First, false recognition was consistently reduced when tested for pictures compared to font color. This finding replicates prior work on recollective distinctiveness, extends this finding to a situation where the lures were equated across conditions, and provides a basis upon which to evaluate the second key result. The second key result was that testing memory for repeated words consistently failed to reduce false recognition, even though repetitions did affect hit rates and the task was formally identical to the picture condition. These patterns were obtained when pictures and repetitions were independently manipulated (Experiments 1 and 2) and also when they were manipulated in the same experiment (Experiment 3). Finally, our subjective measures confirmed that pictures enhanced memory quality more than words (i.e., distinctive or unique recollected details), even though study repetition significantly increased memory quantity (i.e., the number of words eliciting recollection and/or familiarity). Considered as a whole, these findings indicate that differences in recollec-

**Table 6**
Proportion of self-paced old/new recognition responses and subjective responses in Experiment 3.

| | "Old" | "AR"/"VF" | IRK | "URD" |
|---|---|---|---|---|
| Green targets (3x) | .75 (.03) | .54 (.04)/.21 (.02) | .51 (.04) | 3.76 (.17) |
| Picture targets (1x) | .68 (.03) | .51 (.03)/.17 (.02) | .36 (.03) | 4.48 (.20) |
| Red targets (1x) | .55 (.03) | .33 (.03)/.22 (.02) | .35 (.03) | 3.45 (.20) |
| Black targets (1x) | .51 (.03) | .28 (.03)/.23 (.02) | .33 (.03) | 3.29 (.18) |
| New lures (0x) | .12 (.01) | .02 (.01)/.10 (.01) | .10 (.01) | – |

*Notes.* The standard error of each mean is in parenthesis. 1x = studied once, 3x = studied thrice, 0x = not studied. AR = actually recollect, VF = very familiar, IRK = familiarity estimate from the independent-remember-know adjustment, URD = mean ratings of unique recollected details (1–7 scale). These data are collapsed across the instructional conditions. False recognition of new lures was too low to elicit meaningful recollection ratings.

tion quality led to robust and reliable differences in false recognition, whereas differences in memory quantity had no measurable effect on false recognition.

### *Implications for retrieval monitoring*

To interpret these results we first consider why testing memory for pictures reduced false recognition. In Experiments 1 and 3 we found increased hits and decreased false alarms on the picture test compared to the red word test, analogous to the strength-based mirror effect that is sometimes observed in old/new recognition memory (e.g., Hirshman, 1995). The false recognition portion of this pattern has been attributed to the use of a more conservative response criterion on the picture test (cf. Schacter et al., 1999), but importantly, we assume that this kind of shift reflects a greater reliance on qualitative differences in recollection, as opposed to quantitative differences in recollection or familiarity. This assumption is based on our prior work with the criterial recollection task as well as the patterns of subjective judgments in the current studies. However, there also was some evidence that pictures also enhanced more quantitative aspects of memory relative to once-presented words in the current study. To more definitively determine whether quantitative memory differences could affect false recognition, independent from qualitative differences, we turned to the effects of study repetitions.

Does monitoring memory for repeated words elicit the same type of false recognition suppression effect as monitoring memory for pictures? Based on prior research on the mirror effect, and quantitative approaches that appeal to strength-based criterion shifts, there were good reasons to expect that the answer to this question would be "Yes." Nevertheless, the answer was unequivocally "No." In Experiments 2 and 3, study repetitions of words increased quantitative memory strength relative to once-presented words, as indexed by hit rates and by subjective judgments that gauged the frequency of recollection and familiarity, but there was no decrease in false recognition when tested for repeated words compared to once-presented words. These false recognition patterns are particularly striking given the relative ease with which participants suppressed false recognition when tested for pictures, and even though repeated words were at least as strong in memory as pictures, as measured by hit rates, "actually recollect" judgments, and familiarity estimates.

One explanation for the lack of an effect of repetition on false recognition is that, when tested for stimuli that elicit relatively low levels of recollective distinctiveness (such as words), participants adopted a single response criterion regardless of differences in memory strength of the targets across the test blocks. This explanation is consistent with findings from standard old/new recognition memory tests that participants are relatively unwilling to shift response criteria in response to strength differences at test, at least under some conditions (e.g., Stretch & Wixted, 1998; Verde & Rotello, 2007; see Dobbins & Han, 2007). Our results provide a particularly compelling demonstration of this finding, because our testing conditions and instructions made the strength differences obvious to participants. The idea of an invariant response criterion is consistent with the assumption that recollection and familiarity can be blurred or summed into a single "strength-like" experience when recognition memory is tested for words (cf. Wixted, 2007). However, it also could be that word repetitions separately enhanced the quantity of both recollection and familiarity for targets, while false recognition of lures was primarily driven by familiarity and a constant familiarity-based criterion. Either way, our results suggest that false recognition was driven by a constant response criterion under conditions of low recollective distinctiveness.[6]

In contrast to the repetition manipulation, it is clear that participants did change their response criterion to suppress false recognition when tested for pictures, demonstrating that they were willing and able to shift their criteria across the test blocks in our task. Critically, this pattern is counterintuitive from a quantitative perspective that fails to take qualitative aspects of memory into account, because study repetitions led to greater overall memory strength (e.g., old/new recognition) than pictures, but failed to suppress false recognition in the same way. In fact, our subjective judgments indicated that the only memorial dimension where pictures were consistently greater than repeated words was in the judgment of unique recollected details. These findings suggest that qualitative differences in recollection enhanced the accuracy of retrieval monitoring more than quantitative differences, or at least were perceived to be sufficiently more diagnostic by participants to significantly suppress false recognition.

### *Implications for mirror effects*

Our task was specifically designed to study the retrieval monitoring process, which necessarily varied across our test blocks, and not to resolve prior inconsistencies with respect to the strength-based mirror effect in standard recognition memory. Nevertheless, our results do raise the question as to why some studies have found reduced false recognition following study repetitions (e.g., Benjamin, 2001; Starns, Hicks, & Marsh, 2006; Stretch & Wixted, 1998), whereas we did not. One possible explanation for this discrepancy is that strength-based criterion shifts are more likely to occur when participants primarily rely on familiarity, as may be the case on standard recognition memory tests for words. Although we found evidence that familiarity affected performance in the criterial recollection task, our task design and instructions encouraged participants to strategically rely on the recollection of specific features (e.g., font color or pictures). When participants are instructed to attempt to recollect specific features, they may be less likely to consider quantitative memory differ-

[6] A reviewer wondered whether an even stronger manipulation of quantitative strength (e.g., repeating pictures 100 times) would be sufficient to suppress false recognition. This is an open question, but note that increasing study repetitions might eventually enhance qualitative aspects of memory as well (e.g., providing more opportunities to generate unique associations or to process additional visual features). Our use of three repetitions of words was chosen because it is comparable to other studies in the memory literature, and because prior work suggested that it might lead to "stronger" memories than pictures on some dimensions, but still produce less recollective distinctiveness. The extent that these distinctions apply to other manipulations or combinations of stimuli requires additional research.

ences in the monitoring process that affects false recognition.

Another possibility is that the false alarm portion of the strength-based mirror effect in standard recognition memory is not due to strength-based criterion shifts. For example, differentiation models propose that the relative strength of nonstudied items changes as a function of the strength of target items, so that the two are less confusable when targets are strengthen by repetition (e.g., Criss, 2006). More elaborate discussion of these models is outside the present scope (see Starns et al., 2010), but to the extent that differentiation relies on an automatic global matching process that is independent from retrieval orientation, it should not have differentially affected the strength of the lures in our task because all of our test items were preceded by a single study phase. Thus, differentiation models might explain why repetition causes a mirror effect in standard recognition memory, and they also might explain why repetition did not cause a mirror effect in our task.

A final possibility is that prior manipulations of memory strength in the mirror effect literature might have inadvertently affected recollection quality. For example, Singer (2009) found that study repetitions resulted in a mirror effect, but only when these repetitions were associated with a deep processing task at encoding. Deep processing is likely to enhance recollection quality, and this increase in recollection quality may have suppressed false recognition via the same retrieval monitoring mechanisms described in the current study. In fact, Gallo et al. (2008) provided evidence for this hypothesis by manipulating levels of processing in the criterial recollection task. Because studies using standard recognition memory tests do not typically control for the relative contributions of recollection and familiarity, or differences in recollection quality, this hypothesis might apply to other strength-based mirror effects in standard recognition memory tests as well.

Our results also suggest boundary conditions for strength-based theories of the mirror effect on standard recognition memory tests. For example, Bruno et al. (2009) argued that people are more likely to shift their response criterion on standard recognition tests under conditions of poor global subjective memorability for the study list, making these conditions the most likely to result in mirror effect patterns in false recognition. It is unclear how this theory would extend to the current task, because the repetition of study words enhanced their memory strength relative to pictures by some measures, but pictures elicited more distinctive recollections by other measures. The extent that one condition led to greater global subjective memorability than the other depends on the aspect of memory under consideration. Concepts such as global memorability or overall memory strength are difficult to apply to situations where the kinds of to-be-remembered information vary.

More generally, strength-based theories have not been adequately applied to situations where memories are allowed to vary both qualitatively and quantitatively. The current studies represent one such situation, but this situation also seems to hold for most of the complex events that need to be remembered in real life. Multi-dimensional frameworks may provide better descriptions of these sorts of situations, such as the ideas endorsed by the source-monitoring framework (e.g., Johnson et al., 1993) and the mechanics described in some multi-dimensional signal detection theories (e.g., Rotello et al., 2004). These frameworks leave open the possibility that either quantitative or qualitative memory differences can affect retrieval monitoring or decision processes, but exactly how these two factors interact when making memory decisions has been underspecified in prior work. Our study makes a significant advance in this regard, suggesting that only qualitative memory differences affect the accuracy of the retrieval monitoring processes. To the extent that memory quality plays a larger role in the retrieval monitoring process than memory quantity, in both the lab and in life, differences in memory quality need to be more formally considered by existing theoretical frameworks.

## Acknowledgment

This research was funded by a University of Chicago College Research Opportunity Grant to JMS. We thank Ilana Bergelson, Melissa Mongrella, and Meghan Putty for assistance collecting data and Maria McElwain for helpful comments.

## References

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

Banks, W. P. (2000). Recognition and source memory as multivariate decision processes. *Psychological Science, 11*, 267–273.

Benjamin, A. S. (2001). On the dual effects of repetition on false recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 27*, 941–947.

Bink, M. L., Marsh, R. L., & Hicks, J. L. (1999). An alternative conceptualization to memory "strength" in reality monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 804–809.

Bodner, G., & Lindsay, D. S. (2003). Remembering and knowing in context. *Journal of Memory and Language, 48*, 563–580.

Brainerd, C. J., Reyna, V. F., Wright, R., & Mojardin, A. H. (2003). Recollection-rejection: False memory editing in children and adults. *Psychological Review, 110*, 762–784.

Bruno, D., Higham, P. A., & Perfect, T. J. (2009). Global subjective memorability and the strength-based mirror effect in recognition memory. *Memory & Cognition, 37*, 807–818.

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language, 49*, 231–248.

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.

Dobbins, I. G., & Han, S. (2007). What constitutes a model of item-based memory decisions? In *The psychology of learning and motivation*. In A. Benjamin & B. Ross (Eds.). *Strategic and nonstrategic influences on memory attribution* (Vol. 48, pp. 95–144). London: Elsevier.

Dodson, C. S., & Hege, A. C. G. (2005). Speeded retrieval abolishes the false-memory suppression effect: Evidence for the distinctiveness heuristic. *Psychonomic Bulletin & Review, 12*, 726–731.

Dodson, C. S., & Schacter, D. L. (2002). When false recognition meets metacognition: The distinctiveness heuristic. *Journal of Memory and Language, 46*, 782–803.

Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition, 37*, 831–846.

Gallo, D. A., McDonough, I. M., & Scimeca, J. (2010). Dissociating source memory decisions in prefrontal cortex: fMRI of diagnostic and disqualifying monitoring. *Journal of Cognitive Neuroscience, 22*, 955–969.

Gallo, D. A., Meadow, N. G., Johnson, E. L., & Foster, K. T. (2008). Deep levels of processing elicit a distinctiveness heuristic: Evidence from the criterial recollection task. *Journal of Memory and Language, 58*, 1095–1111.

Gallo, D. A., Weiss, J. A., & Schacter, D. L. (2004). Reducing false recognition with criterial recollection tests: Distinctiveness heuristic versus criterion shifts. *Journal of Memory and Language, 51*, 473–493.

Geraci, L., & McCabe, D. P. (2006). Examining the basis for illusory recollection: The role of remember/know instructions. *Psychonomic Bulletin & Review, 13*, 466–473.

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

Greene, R. L. (2007). Foxes, hedgehogs, and mirror effects: The role of general principles in memory research. In J. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger III* (pp. 53–66). Psychology Press.

Hay, J. F., & Jacoby, L. L. (1996). Separating habit and recollection: Memory slips, process dissociations and probability matching. *Journal of Experimental Psychology: Learning, Memory and Cognition, 22*, 1323–1335.

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

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

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.

Israel, L., & Schacter, D. L. (1997). Pictorial encoding reduces false recognition of semantic associates. *Psychonomic Bulletin & Review, 4*, 577–581.

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

Jacoby, L. L. (1999). Ironic effects of repetition: Measuring age-related differences in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 3–22.

Jenkins, J. R., Neale, D. C., & Deno, S. L. (1967). Differential memory for picture and word stimuli. *Journal of Educational Psychology, 58*, 303–307.

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

Joordens, S., & Hockley, W. E. (2000). Recollection and familiarity through the looking glass: When old does not mirror new. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 1534–1555.

Kim, K., & Glanzer, G. (1993). Speed versus accuracy instructions, study time, and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 638–652.

Koriat, A., Goldsmith, M., & Pansky, A. (2000). Toward a psychology of memory accuracy. *Annual Review of Psychology, 51*, 481–537.

Lockhart, R. S., & Murdock, B. J. (1970). Memory and the theory of signal detection. *Psychological Bulletin, 74*, 100–109.

Mandler, G. (1980). Recognizing: The judgment of previous occurrences. *Psychological Review, 87*, 252–271.

McDonough, I. M., & Gallo, D. A. (2008). Autobiographical elaboration reduces memory distortion: Cognitive operations and the distinctiveness heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1430–1445.

Mitchell, K. J., & Johnson, M. K. (2009). Source monitoring 15 years later: What have we learned from fMRI about the neural mechanisms of source memory? *Psychological Bulletin, 135*, 638–677.

Nelson, D. L. (1979). Remembering pictures and words: Appearance, significance, and name. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 45–76). Hillsdale, NJ: Erlbaum.

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

Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory & Cognition, 21*, 89–102.

Rotello, C. M., Macmillan, N. A., & Reder, J. A. (2004). Sum-difference theory of remember and knowing: A two-dimensional signal-detection model. *Psychological Review, 111*, 588–616.

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

Schacter, D. L., & Wiseman, A. L. (2006). Reducing memory errors: The distinctiveness heuristic. In R. R. Hunt & J. Worthen (Eds.), *Distinctiveness and memory* (pp. 89–107). New York, NY: Oxford University Press.

Singer, M. (2009). Strength-based criterion shifts in recognition memory. *Memory & Cognition, 37*, 976–984.

Starns, J. J., Hicks, J. L., & Marsh, R. L. (2006). Repetition effects in associative false recognition: Theme-based criterion shifts are the exception, not the rule. *Memory, 14*, 742–761.

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.

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

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology, 26*, 1–12.

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

Wixted, J. T. (2007). Dual-process theory and signal detection theory of recognition memory. *Psychological Review, 142*, 152–176.

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review, 117*, 1025–1054.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441–517.